


The IMA Volumes in Mathematics and its Applications

George Yin
Qing Zhang *Editors*

Modeling, Stochastic Control, Optimization, and Applications



 Springer

The IMA Volumes in Mathematics and its Applications

Volume 164

Series editor

Daniel Spirn, *Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN, USA*

Institute for Mathematics and its Applications (IMA)

The Institute for Mathematics and its Applications (IMA) was established in 1982 as a result of a National Science Foundation competition. The mission of the IMA is to connect scientists, engineers, and mathematicians in order to address scientific and technological challenges in a collaborative, engaging environment, developing transformative, new mathematics and exploring its applications, while training the next generation of researchers and educators. To this end the IMA organizes a wide variety of programs, ranging from short intense workshops in areas of exceptional interest and opportunity to extensive thematic programs lasting nine months. The IMA Volumes are used to disseminate results of these programs to the broader scientific community.

The full list of IMA books can be found at the Web site of the Institute for Mathematics and its Applications:

<http://www.ima.umn.edu/springer/volumes.html>.

Presentation materials from the IMA talks are available at

<http://www.ima.umn.edu/talks/>.

Video library is at

<http://www.ima.umn.edu/videos/>.

Daniel Spim, Director of the IMA

More information about this series at <http://www.springer.com/series/811>

George Yin · Qing Zhang
Editors

Modeling, Stochastic Control, Optimization, and Applications

 Springer

Editors

George Yin
Department of Mathematics
Wayne State University
Detroit, MI, USA

Qing Zhang
Department of Mathematics
University of Georgia
Athens, GA, USA

ISSN 0940-6573

ISSN 2198-3224 (electronic)

The IMA Volumes in Mathematics and its Applications

ISBN 978-3-030-25497-1

ISBN 978-3-030-25498-8 (eBook)

<https://doi.org/10.1007/978-3-030-25498-8>

Mathematics Subject Classification (2010): 60H10, 60H15, 60J60, 60J75, 93E20, 93E03

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

This volume contains a selection of papers based on a thematic summer program “Modeling, Stochastic Control, Optimization, and Related Applications” held at the Institute for Mathematics and its Applications from May 1-June 30, 2018 and organized by George Yin and Qing Zhang. This summer program included four week-long workshops over the eight week focus period:

- Stochastic Control, Computational methods, and Applications,
- Queuing Theory and Networked Systems,
- Ecological and Biological Applications, and
- Finance and Economics Applications.

The papers found in this volume cover a broad array of application areas and utilize a wide variety of techniques, from applied probability and network science to stochastic differential equations and numerical methods. We would like to thank volume editors, George Yin and Qing Zhang. Finally, we acknowledge the National Science Foundation for its support of this program.

The Institute for Mathematics and its Applications was established by a grant from the National Science Foundation to the University of Minnesota in 1982. The primary mission of the IMA is to foster research of a truly interdisciplinary nature, establishing links between mathematics of the highest caliber and important scientific and technological problems from other disciplines and industries. IMA Volumes are used to communicate results of these programs that we believe are of particular value to the broader scientific community. The full list of IMA books can be found at the web site of the Institute for Mathematics and its Applications:

<http://www.ima.umn.edu/springer/volumes.html>.

Presentation materials from the IMA talks are available at

<http://www.ima.umn.edu/talks/>.

A video library is at

<http://www.ima.umn.edu/videos/>.

Minneapolis, MN, USA, May 2019

Daniel Spirn

Preface

This volume collects papers, based on invited talks given at the IMA workshop in Modeling, Stochastic Control, Optimization, and Related Applications, held at the Institute for Mathematics and Its Applications, University of Minnesota, during May and June, 2018.

There were four week-long workshops during the conference. They are (1) stochastic control, computation methods, and applications, (2) queueing theory and networked systems, (3) ecological and biological applications, and (4) finance and economics applications. For broader impacts, researchers from different fields covering both theoretically oriented and application intensive areas were invited to participate in the conference. It brought together researchers from multi-disciplinary communities in applied mathematics, applied probability, engineering, biology, ecology, and networked science, to review, and substantially update most recent progress. As an archive, this volume presents some of the highlights of the workshops, and collect papers covering a broad range of topics. After the conference, 23 papers were submitted. All papers were reviewed.

Without the help and assistance of the IMA directors and staffs, the conference could not come into being. We thank Fadil Santosa for the initial suggestion and encouragement as well as subsequent help. Our thanks go to Daniel Spirn, Benjamin Brubaker, and Katherine Dowd for their support throughout the workshops. It had been a pleasure working with Rebecca Malkovich and Georgia Kroll and all the IMA professionals. Their supports are greatly acknowledged. Finally, we thank the contributors of this volume, invited speakers of the workshops, and the attendees for making the conference a successful and memorable event.

Detroit and Athens,
Feb., 2019,

George Yin
Qing Zhang

Table of Contents

Foreword	v
Preface	vii
Uniform Polynomial Rates of Convergence for A Class of Lévy-Driven Controlled SDEs Arising in Multiclass Many-Server Queues	1
A. Arapostathis, H. Hmedi, G. Pang, and N. Sandrić	
Nudged Particle Filters in Multiscale Chaotic Systems with Correlated Sensor Noise	21
R. Beeson and N.S. Namachchivaya	
Postponing Collapse: Ergodic Control with a Probabilistic Constraint ...	57
V.S. Borkar and J.A. Filar	
Resource Sharing Networks and Brownian Control Problems	67
A. Budhiraja and M. Conroy	
American Option Model and Negative Fichera Function on Degenerate Boundary	95
X. Chen, Z. Jin, and Q. Song	
Continuous-Time Markov Chain and Regime Switching Approximations with Applications to Options Pricing	115
Z. Cui, J.L. Kirkby, and D. Nguyen	
Numerical Approximations for Discounted Continuous Time Markov Decision Processes	147
F. Dufour and T. Prieto-Rumeau	
Some Linear-Quadratic Stochastic Differential Games Driven by State Dependent Gauss-Volterra Processes	173
T.E. Duncan and B. Pasik-Duncan	
Correlated Equilibria for Infinite Horizon Nonzero-Sum Stochastic Differential Games	181
B.A. Escobedo-Trujillo and H. Jasso-Fuentes	

Lattice Dynamical Systems in the Biological Sciences	201
X. Han and P.E. Kloeden	
Balancing Prevention and Suppression of Forest Fires with Fuel Management as a Stock	235
B. Heines, S. Lenhart, and C. Sims	
A Free-Model Characterization of the Asymptotic Certainty Equivalent by the Arrow-Pratt Index	261
D. Hernández-Hernández and E. Treviño-Aguilar	
Binary Mean Field Stochastic Games: Stationary Equilibria and Comparative Statics Queues	283
M. Huang and Y. Ma	
Equivalence of Fluid Models for $G_t/GI/N + GI$ Queues	315
W. Kang and G. Pang	
Stochastic HJB Equations and Regular Singular Points	351
A.J. Krener	
Information Diffusion in Social Networks: Friendship Paradox based Models and Statistical Inference	369
V. Krishnamurthy and B. Nettasinghe	
Portfolio Optimization Using Regime-Switching Stochastic Interest Rate and Stochastic Volatility Models	407
R.H. Liu and D. Ren	
On Optimal Stopping and Impulse Control with Constraint	427
J.L. Menaldi and M. Robin	
Linear-Quadratic McKean-Vlasov Stochastic Differential Games	451
E. Miller and H. Pham	
Stochastic Multigroup Epidemic Models: Duration and Final Size	483
A. Nandi and L.J.S. Allen	
H_2 Dynamic Output Feedback Control for Hidden Markov Jump Linear Systems	509
A.M. de Oliveira, O.L.V. Costa, and J. Daafouz	
Time-Inconsistent Optimal Control Problems and Related Issues	533
W. Yan and J. Yong	
Regime-Switching Jump Diffusions with Non-Lipschitz Coefficients and Countably Many Switching States: Existence and Uniqueness, Feller, and Strong Feller Properties	571
F. Xi, G. Yin, and C. Zhu	



Uniform Polynomial Rates of Convergence for A Class of Lévy-Driven Controlled SDEs Arising in Multiclass Many-Server Queues

Ari Arapostathis, Hassan Hmedi, Guodong Pang, and Nikola Sandrić

Abstract We study the ergodic properties of a class of controlled stochastic differential equations (SDEs) driven by α -stable processes which arise as the limiting equations of multiclass queueing models in the Halfin–Whitt regime that have heavy-tailed arrival processes. When the safety staffing parameter is positive, we show that the SDEs are uniformly ergodic and enjoy a polynomial rate of convergence to the invariant probability measure in total variation, which is uniform over all stationary Markov controls resulting in a locally Lipschitz continuous drift. We also derive a matching lower bound on the rate of convergence (under no abandonment). On the other hand, when all abandonment rates are positive, we show that the SDEs are exponentially ergodic uniformly over the above-mentioned class of controls. Analogous results are obtained for Lévy-driven SDEs arising from multiclass many-server queues under asymptotically negligible service interruptions. For these equations, we show that the aforementioned ergodic properties are uniform over all stationary Markov controls. We also extend a key functional central limit theorem concerning diffusion approximations so as to make it applicable to the models studied here.

Key words: subexponential ergodicity, multiclass queues, Lévy processes

Ari Arapostathis

Department of ECE, The University of Texas at Austin, EER 7.824, Austin, TX 78712, e-mail: ari@ece.utexas.edu

Hassan Hmedi

Department of ECE, The University of Texas at Austin, EER 7.834, Austin, TX 78712, e-mail: hmedi@utexas.edu

Guodong Pang

The Harold and Inge Marcus Dept. of Industrial and Manufacturing Eng., College of Engineering, Pennsylvania State University, University Park, PA 16802, e-mail: gup3@psu.edu

Nikola Sandrić

Department of Mathematics, University of Zagreb, Bijenička cesta 30, 10000 Zagreb, Croatia, e-mail: nsandric@math.hr

1 Introduction

Lévy-driven controlled stochastic differential equations (SDEs) arise as scaling limits for multiclass many-server queues with heavy-tailed arrival processes and/or with asymptotically negligible service interruptions; see [4, 12, 13]. In these equations, the control appears only in the drift and corresponds to a work-conserving scheduling policy in multiclass many-server queues, that is, the allocation of the available service capacity to each class under a non-idling condition (no server idles whenever there are jobs in queue). For the limiting process, we focus on stationary Markov controls, namely time-homogeneous functions of the process. When the arrival process of each class is heavy-tailed (for example, with regularly varying interarrival times), the Lévy process driving the SDE is a multidimensional anisotropic α -stable process, $\alpha \in (1, 2)$. When the system is subject to service interruptions (in an alternating renewal environment affecting the service processes only), the Lévy process is a combination of either a Brownian motion, or an anisotropic α -stable process, $\alpha \in (1, 2)$, and an independent compound Poisson process.

Ergodic properties of these controlled SDEs are of great interest since they help to understand the performance of the queueing systems. In [4], the ergodic properties of the SDEs under constant controls are thoroughly studied. It is shown that when the safety staffing is positive, the SDEs have a polynomial rate of convergence to stationarity in total variation, while when the abandonment rates are positive, the rate of convergence is exponential. However, the technique developed in [4] does not equip us to investigate the ergodic properties of these SDEs beyond the constant controls, since the Lyapunov functions employed are modifications of the common quadratic functions that have been developed for piecewise linear diffusions [5].

It was recently shown in [7] that the Markovian multiclass many-server queues with positive safety staffing in the Halfin–Whitt regime are stable under any work-conserving scheduling policies. Motivated by this significant result, Arapostathis et al. (2018) [3] have developed a unified approach via a Lyapunov function method which establishes Foster-Lyapunov equations which are uniform under stationary Markov controls for the limiting diffusion and the prelimit diffusion-scaled queueing processes simultaneously. It is shown that the limiting diffusion is uniformly exponentially ergodic under any stationary Markov control.

In this paper we adopt and extend the approach in [3] to establish uniform ergodic properties for Lévy-driven SDEs. As done in [4], we distinguish two cases: (i) positive safety staffing, and (ii) positive abandonment rates. We focus primarily on the first case, which exhibits ergodicity at a polynomial rate, a result which is somewhat surprising. The second case always results in uniform exponential ergodicity. By employing a polynomial Lyapunov function instead of the exponential function used in [3], we first establish an upper bound on the rate of convergence which is polynomial. The drift inequalities carry over with slight modifications from [3], while the needed properties of the non-local part of the generator are borrowed from [2]. As in [4], we use the technique in [9] to establish a lower bound on the rate of convergence, which actually matches the upper bound. As a result, we establish that with positive safety staffing, the rate of convergence to stationarity in

total variation is polynomial with a rate that is uniform over the family of Markov controls which result in a locally Lipschitz continuous drift.

When the SDE is driven by an α -stable process (isotropic or anisotropic), in order for the process to be open-set irreducible and aperiodic, it suffices to require that the controls are stationary Markov and the drift is locally Lipschitz continuous. However, the existing proof of the convergence of the scaled queueing processes of the multiclass many-server queues with heavy-tailed arrivals to this limit process, assumes that the drift is Lipschitz continuous [13]. In this paper, we extend this result on the continuity of the integral mapping (Theorem 1.1 in [13]) to drifts that are locally Lipschitz continuous with at most linear growth (see Lemma 4). Applying this, we also present an extended functional central limit theorem (FCLT) for multiclass many-server queues with heavy-tailed arrival processes (see Theorem 6).

On the other hand, when the Lévy process consists of a Brownian motion and a compound Poisson process, which arises in the multiclass many-server queues with asymptotically negligible interruptions under the \sqrt{n} scaling, the SDE has a unique strong solution that is open-set irreducible and aperiodic under any stationary Markov control. To study uniform ergodic properties, we also need to account for the second order derivatives in the infinitesimal generator. For this reason we modify the Lyapunov function with suitable titling on the positive and negative half state spaces. We also discuss the model with a Lévy process consisting of a α -stable process and a compound Poisson process.

1.1 Organization of the paper

In Section 2, we present a class of SDEs driven by an α -stable process, whose ergodic properties are studied in Section 3. In Section 4, we study the ergodic properties of Lévy-driven SDEs arising from the multiclass queueing models with service interruptions. In Section 5, we provide a description of the multiclass many-server queues with heavy-tailed arrival processes, and establish the continuity of the integral mapping with a locally Lipschitz continuous function that has at most linear growth, as well as the associated FCLT.

1.2 Notation

We summarize some notation used throughout the paper. We use \mathbb{R}^m (and \mathbb{R}_+^m), $m \geq 1$, to denote real-valued m -dimensional (nonnegative) vectors, and write \mathbb{R} for $m = 1$. For $x, y \in \mathbb{R}$, we write $x \vee y = \max\{x, y\}$, $x \wedge y = \min\{x, y\}$, $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$. For a set $A \subseteq \mathbb{R}^m$, we use A^c , ∂A , and $\mathbb{1}_A$ to denote the complement, the boundary, and the indicator function of A , respectively. A ball of radius $r > 0$ in \mathbb{R}^m around a point x is denoted by $\mathcal{B}_r(x)$, or simply as \mathcal{B}_r if $x = 0$. We also let $\mathcal{B} \equiv \mathcal{B}_1$. The Euclidean norm on \mathbb{R}^m is denoted by $|\cdot|$, and $\langle \cdot, \cdot \rangle$ stands

for the inner product. For $x \in \mathbb{R}^m$, we let $\|x\|_1 := \sum_i |x_i|$, and we use x' to denote the transpose of x . We use the symbol e to denote the vector whose elements are all equal to 1, and e_i for the vector whose i^{th} element is equal to 1 and the rest are equal to 0.

We let $\mathcal{B}(\mathbb{R}^m)$, $\mathcal{B}_b(\mathbb{R}^m)$, and $\mathcal{P}(\mathbb{R}^m)$ denote the classes of Borel measurable functions, bounded Borel measurable functions, and Borel probability measures on \mathbb{R}^m , respectively. By $\mathcal{P}_p(\mathbb{R}^m)$, $p > 0$, we denote the subset of $\mathcal{P}(\mathbb{R}^m)$ containing all probability measures $\pi(dx)$ with the property that $\int_{\mathbb{R}^m} |x|^p \pi(dx) < \infty$. For a finite signed measure ν on \mathbb{R}^m , and a Borel measurable $f: \mathbb{R}^m \rightarrow [1, \infty)$, $\|\nu\|_f := \sup_{|g| \leq f} \int_{\mathbb{R}^m} |g(x)| \nu(dx)$, where the supremum is over all Borel measurable functions g satisfying this inequality.

2 The model

We consider an m -dimensional stochastic differential equation (SDE) of the form

$$dX_t = b(X_t, U_t) dt + d\widehat{A}_t, \quad X_0 = x \in \mathbb{R}^m. \quad (1)$$

All random processes in (1) live in a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We have the following structural hypotheses.

(A1) The control process $\{U_t\}_{t \geq 0}$ lives in the $(m-1)$ -simplex

$$\Delta := \{u \in \mathbb{R}^m : u \geq 0, \langle e, u \rangle = 1\},$$

and the drift $b: \mathbb{R}^m \times \Delta \rightarrow \mathbb{R}^m$ is given by

$$\begin{aligned} b(x, u) &= \ell - M(x - \langle e, x \rangle^+ u) - \langle e, x \rangle^+ \Gamma u \\ &= \begin{cases} \ell - (M + (\Gamma - M)ue')x, & \langle e, x \rangle > 0, \\ \ell - Mx, & \langle e, x \rangle \leq 0, \end{cases} \end{aligned} \quad (2)$$

where $\ell \in \mathbb{R}^m$, $M = \text{diag}(\mu_1, \dots, \mu_m)$ with $\mu_i > 0$, and $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_m)$ with $\gamma_i \in \mathbb{R}_+$, $i = 1, \dots, m$.

(A2) The process $\{\widehat{A}_t\}_{t \geq 0}$ is an anisotropic Lévy process with independent symmetric one-dimensional α -stable components for $\alpha \in (1, 2)$.

Define

$$\mathcal{K}_+ := \{x \in \mathbb{R}^m : \langle e, x \rangle > 0\}, \quad \text{and} \quad \mathcal{K}_- := \{x \in \mathbb{R}^m : \langle e, x \rangle \leq 0\}.$$

A control U_t is called stationary Markov, if it takes the form $U_t = v(X_t)$ for a Borel measurable function $v: \mathcal{K}_+ \rightarrow \Delta$. We let \mathcal{U}_{sm} denote the class of stationary Markov controls, and \mathcal{U}'_{sm} its subset consisting of those controls under which

$$b_v(x) := b(x, v(x))$$

is locally Lipschitz continuous. These controls can be identified with the function v . Note that if $v: \mathcal{K}_+ \rightarrow \Delta$ is Lipschitz continuous when restricted to any set $\mathcal{K}_+ \cap \mathcal{B}_R$, $R > 0$, then $v \in \mathfrak{U}_{\text{sm}}$, but this property is not necessary for membership in $\widetilde{\mathfrak{U}}_{\text{sm}}$.

Clearly, for any $v \in \mathfrak{U}_{\text{sm}}$, the drift $b_v(x)$ has at most linear growth. Therefore, if $v \in \widetilde{\mathfrak{U}}_{\text{sm}}$, then using [1, Theorem 3.1, and Propositions 4.2 and 4.3], one can conclude that the SDE (1) admits a unique nonexplosive strong solution $\{X_t\}_{t \geq 0}$ which is a strong Markov process and it satisfies the C_b -Feller property. In addition, in the same reference, it is shown that the infinitesimal generator $(\mathcal{A}^v, \mathcal{D}_{\mathcal{A}^v})$ of $\{X_t\}_{t \geq 0}$ (with respect to the Banach space $(\mathcal{B}_b(\mathbb{R}^m), \|\cdot\|_\infty)$) satisfies $C_c^2(\mathbb{R}^m) \subseteq \mathcal{D}_{\mathcal{A}^v}$ and

$$\mathcal{A}^v \Big|_{C_c^2(\mathbb{R}^m)} f(x) := \langle b_v(x), \nabla f(x) \rangle + \mathfrak{I}_\alpha f(x), \quad (3)$$

where

$$\mathfrak{I}_\alpha f(x) := \sum_{i=1}^d \int_{\mathbb{R}_*} \partial f(x; y_i e_i) \frac{\xi_i dy_i}{|y_i|^{1+\alpha}},$$

for some positive constants ξ_1, \dots, ξ_m , and

$$\partial f(x; y) := f(x+y) - f(x) - \langle y, \nabla f(x) \rangle, \quad f \in C^1(\mathbb{R}^m). \quad (4)$$

Here, $\mathcal{D}_{\mathcal{A}^v}$ and $C_c^2(\mathbb{R}^m)$ denote the domain of \mathcal{A}^v and the space of twice continuously differentiable functions with compact support, respectively.

We let \mathbb{P}_x^v and \mathbb{E}_x^v denote the probability measure and expectation operator on the canonical space of the solution of (1) under $v \in \mathfrak{U}_{\text{sm}}$ and starting at x . Also, $P_t^v(x, dy)$ denotes its transition probability. From the proof of Theorem 3.1 (iv) in [4] we have the following result.

Theorem 1. *Under any $v \in \widetilde{\mathfrak{U}}_{\text{sm}}$, $P_t^v(x, B) > 0$ for all $t > 0$, $x \in \mathbb{R}^m$ and $B \in \mathcal{B}(\mathbb{R}^m)$ with positive Lebesgue measure. In particular, under any $v \in \widetilde{\mathfrak{U}}_{\text{sm}}$, the process $\{X_t\}_{t \geq 0}$ is open-set irreducible and aperiodic in the sense of [11].*

Remark 1. As far as the results in this paper are concerned we can replace the anisotropic non-local operator \mathfrak{I}_α with the isotropic operator

$$\int_{\mathbb{R}_*} \partial f(x; y) \frac{dy}{|y|^{m+\alpha}},$$

as done in [4].

We also define

$$\mathcal{A}^u f(x) := \langle b(x, u), \nabla f(x) \rangle + \mathfrak{I}_\alpha f(x), \quad u \in \Delta.$$

In the next section we study the ergodic properties of $\{X_t\}_{t \geq 0}$. To facilitate the analysis, we define the *spare capacity*, or *safety staffing*, β as

$$\beta := -\langle e, M^{-1}\ell \rangle. \quad (5)$$

Note that if we let $\zeta = \frac{\beta}{m}e + M^{-1}\ell$, with β as in (5), then a mere translation of the origin of the form $\tilde{X}_t = X_t - \zeta$ results in an SDE of the same form, with the only difference that the constant term ℓ in the drift equals $-\frac{\beta}{m}Me$. Since translating the origin does not alter the ergodic properties of the process, without loss of generality, we assume throughout the paper that the drift in (2) has the form

$$b(x, u) = -\frac{\beta}{m}Me - M(x - \langle e, x \rangle^+ u) - \langle e, x \rangle^+ \Gamma u. \quad (6)$$

3 Uniform ergodic properties

We recall some important definitions used in [3, Section 2.3].

Definition 1. We fix some convex function $\psi \in C^2(\mathbb{R})$ with the property that $\psi(t)$ is constant for $t \leq -1$, and $\psi(t) = t$ for $t \geq 0$. The particular form of this function is not important. But to aid some calculations we fix this function as

$$\psi(t) := \begin{cases} -\frac{1}{2}, & t \leq -1, \\ (t+1)^3 - \frac{1}{2}(t+1)^4 - \frac{1}{2} & t \in [-1, 0], \\ t & t \geq 0. \end{cases}$$

Let $\mathcal{J} = \{1, \dots, m\}$. With δ and p positive constants, we define

$$\Psi(x) := \sum_{i \in \mathcal{J}} \frac{\Psi(x_i)}{\mu_i}, \quad \text{and} \quad V_p(x) := \left(\delta \Psi(-x) + \Psi(x) + \frac{m}{\min_{i \in \mathcal{J}} \mu_i} \right)^p.$$

Note that the term inside the parenthesis in the definition of V_p , or in other words V_1 , is bounded away from 0 uniformly in $\delta \in (0, 1]$. The function V_p also depends on the parameter δ which is suppressed in the notation.

For $x \in \mathbb{R}^m$ we let $x^\pm := (x_1^\pm, \dots, x_m^\pm)$. The results which follows is a corollary of Lemma 2.1 in [3], but we sketch the proof for completeness.

Lemma 1. Assume $\beta > 0$, and let $\delta \in (0, 1]$ satisfy

$$\left(\max_{i \in \mathcal{J}} \frac{\mu_i}{\mu_i} - 1 \right)^+ \delta \leq 1. \quad (7)$$

Then, the function V_p in Definition 1 satisfies, for any $p > 1$ and for all $u \in \Delta$,

$$\langle b(x, u), \nabla V_p(x) \rangle \leq p \left(\delta \beta + \frac{m}{2}(1 + \delta) - \delta \|x\|_1 \right) V_{p-1}(x) \quad \forall x \in \mathcal{K}_-, \quad (8)$$

$$\langle b(x, u), \nabla V_p(x) \rangle \leq -p \left(\frac{\beta}{m} - \delta \beta - \delta \frac{m}{2} + \delta \|x^-\|_1 \right) V_{p-1}(x) \quad \forall x \in \mathcal{K}_+. \quad (9)$$

Proof. We have

$$\begin{aligned} \langle b(x, u), \nabla \Psi(x) \rangle &= -\frac{\beta}{m} \sum_{i \in \mathcal{J}} \psi'(x_i) - \sum_{i \in \mathcal{J}} \psi'(x_i) (x_i - \langle e, x \rangle^+ u_i) \\ &\quad - \langle e, x \rangle^+ \sum_{i \in \mathcal{J}} \psi'(x_i) \frac{\gamma_i}{\mu_i} u_i, \end{aligned} \quad (10)$$

and

$$\begin{aligned} \langle b(x, u), \nabla \Psi(-x) \rangle &= \frac{\beta}{m} \sum_{i \in \mathcal{J}} \psi'(-x_i) + \sum_{i \in \mathcal{J}} \psi'(-x_i) x_i \\ &\quad - \langle e, x \rangle^+ \sum_{i \in \mathcal{J}} \psi'(-x_i) \left(1 - \frac{\gamma_i}{\mu_i}\right)^+ u_i \\ &\quad + \langle e, x \rangle^+ \sum_{i \in \mathcal{J}} \psi'(-x_i) \left(\frac{\gamma_i}{\mu_i} - 1\right)^+ u_i. \end{aligned} \quad (11)$$

It is easy to verify that $\psi'(-1/2) = 1/2$, from which we obtain

$$\sum_{i \in \mathcal{J}} \psi'(x_i) x_i \geq \|x^+\|_1 - \frac{m}{2}, \quad \text{and} \quad -\sum_{i \in \mathcal{J}} \psi'(-x_i) x_i \geq \|x^-\|_1 - \frac{m}{2}. \quad (12)$$

Therefore, (8) follows by using (12) in (10)–(11).

We next turn to the proof of (9). If $\gamma_i \leq \mu_i$ for all $i \in \mathcal{J}$, then the proof is simple. This is because the inequality $\sum_{i \in \mathcal{J}} \psi'(x_i) x_i \geq \langle e, x \rangle$ and the fact that $\|\psi'\|_\infty \leq 1$ implies that

$$\sum_{i \in \mathcal{J}} \psi'(x_i) (x_i - \langle e, x \rangle^+ u_i) \geq 0 \quad \text{for } x \in \mathcal{K}_+,$$

which together with (10) shows that

$$\langle b(x, u), \nabla \Psi(x) \rangle \leq -\frac{\beta}{m} \sum_{i \in \mathcal{J}} \psi'(x_i) \leq -\frac{\beta}{m} \quad \text{on } \mathcal{K}_+. \quad (13)$$

On the other hand, by (11) and (12) we obtain

$$\begin{aligned} \delta \langle b(x, u), \nabla \Psi(-x) \rangle &\leq \delta \frac{\beta}{m} \sum_{i \in \mathcal{J}} \psi'(-x_i) + \delta \sum_{i \in \mathcal{J}} \psi'(-x_i) x_i \\ &\leq \delta \beta + \delta \frac{m}{2} - \delta \|x^-\|_1 \quad \text{on } \mathbb{R}^m. \end{aligned} \quad (14)$$

Therefore, when $\gamma_i \leq \mu_i$ for all $i \in \mathcal{J}$, (9) follows by adding (13) and (14).

Without assuming that $\gamma_i \leq \mu_i$, a careful comparison of the terms in (10)–(11), shows that (see [3, Lemma 2.1])

$$\begin{aligned} \delta \langle e, x \rangle^+ \sum_{i \in \mathcal{J}} \psi'(-x_i) \left(\frac{\gamma_i}{\mu_i} - 1\right)^+ u_i - \sum_{i \in \mathcal{J}} \psi'(x_i) (x_i - \langle e, x \rangle^+ u_i) \\ - \langle e, x \rangle^+ \sum_{i \in \mathcal{J}} \psi'(x_i) \frac{\gamma_i}{\mu_i} u_i \leq 0 \quad \forall (x, u) \in \mathcal{K}_+ \times \Delta. \end{aligned} \quad (15)$$

Thus (9) follows by using (13)–(15) in (10)–(11). This completes the proof. \square

On the other hand, when $\Gamma > 0$, the proof of [3, Theorem 2.2] implies the following.

Lemma 2. *Assume that $\Gamma > 0$. Then there exists a positive constant δ such that for any $p > 1$,*

$$\langle b(x, u), \nabla V_p(x) \rangle \leq c_0 - c_1 V_p(x) \quad \forall (x, u) \in \mathbb{R}^m \times \Delta,$$

for some positive constants c_0 and c_1 depending only on δ .

Another result that we borrow is Proposition 5.1 in [2], whose proof implies the following.

Lemma 3. *The map $x \mapsto |x|^{\alpha-p} \mathfrak{J}_\alpha V_p(x)$ is bounded on \mathbb{R}^m for any $p \in (0, \alpha)$.*

Theorems 2 and 3 that follow establish ergodic properties which are uniform over controls in $\tilde{\mathfrak{U}}_{\text{sm}}$ in the case of positive safety staffing and positive abandonment rates, respectively.

Theorem 2. *Assume $\beta > 0$. In addition to (7), let*

$$\delta < \frac{\beta}{2m(2\beta + m)}. \quad (16)$$

We have the following.

(a) *For any $p \in (1, \alpha)$, the function $V_p(x)$ in Definition 1 satisfies the Foster–Lyapunov equation*

$$\mathcal{A}^u V_p(x) \leq C_0(p) - p \left(\frac{\beta}{2m} + \delta \|x^-\|_1 \right) V_{p-1}(x) \quad \forall (x, u) \in \mathbb{R}^m \times \Delta, \quad (17)$$

for some positive constant $C_0(p)$ depending only on p .

(b) *Under any $v \in \tilde{\mathfrak{U}}_{\text{sm}}$, the process $\{X_t\}_{t \geq 0}$ in (1) admits a unique invariant probability measure $\bar{\pi}_v \in \mathcal{P}(\mathbb{R}^m)$.*

(c) *There exists a constant $C_1(\varepsilon)$ depending only on $\varepsilon \in (0, \alpha)$, such that, under any $v \in \tilde{\mathfrak{U}}_{\text{sm}}$, the process $\{X_t\}_{t \geq 0}$ in (1) satisfies*

$$\|P_t^v(x, \cdot) - \bar{\pi}_v(\cdot)\|_{\text{TV}} \leq C_1(\varepsilon) (t \vee 1)^{1+\varepsilon-\alpha} |x|^{\alpha-\varepsilon} \quad \forall x \in \mathbb{R}^m. \quad (18)$$

Proof. Note that, since $\alpha > 1$, Lemma 3 implies that $\frac{\mathfrak{J}_\alpha V_p(x)}{1+|V_{p-1}(x)|}$ vanishes at infinity.

Using δ as in (16), it is clear that $\delta\beta + \delta\frac{m}{2} \leq \frac{\beta}{2m}$. Thus, (17) is a direct consequence of Lemmas 1 and 3 together with the definition in (3).

Clearly, (17) implies that

$$\mathcal{A}^v V_p(x) \leq C_0(p) - p \frac{\beta}{2m} V_{p-1}(x) \quad \forall x \in \mathbb{R}^m, \quad (19)$$

and for any $v \in \widetilde{\mathfrak{U}}_{\text{sm}}$. It is well known that the existence of an invariant probability measure $\bar{\pi}_v$ follows from the C_b -Feller property and (19), while the open-set irreducibility asserted in Theorem 1 implies its uniqueness.

Equation (18) is a direct result of (19), Theorem 1 and [6, Theorem 3.2]. This completes the proof. \square

Theorem 3. *Assume that $\Gamma > 0$ and $p \in [1, \alpha)$. Then, there exists a positive constant δ such that*

$$A^u V_p(x) \leq \tilde{\kappa}_0 - \tilde{\kappa}_1 V_p(x) \quad \forall (x, u) \in \mathbb{R}^m \times \Delta.$$

for some positive constants $\tilde{\kappa}_0$ and $\tilde{\kappa}_1$. Moreover, under any $v \in \widetilde{\mathfrak{U}}_{\text{sm}}$, the process $\{X_t\}_{t \geq 0}$ admits a unique invariant probability measure $\bar{\pi}_v \in \mathcal{P}(\mathbb{R}^m)$, and for any $\gamma \in (0, \tilde{\kappa}_1)$ there exists a positive constant C_γ such that

$$\|P_t^v(x, \cdot) - \bar{\pi}_v(\cdot)\|_{V_p} \leq C_\gamma V_p(x) e^{-\gamma t}, \quad x \in \mathbb{R}^m, t \geq 0.$$

Remark 2. We limited our attention to controls in $\widetilde{\mathfrak{U}}_{\text{sm}}$ only to take advantage of Theorem 1. However, if under some $v \in \mathfrak{U}_{\text{sm}}$ the SDE in (1) has a unique weak solution which is an open-set irreducible and aperiodic C_b -Feller process, then it has a unique invariant probability measure $\bar{\pi}_v$, and the conclusions of Theorems 2 and 3 follow.

Concerning the lower bound on the rate of convergence, we need not restrict the controls in \mathfrak{U}_{sm} . The lack of integrability of functions that have strict polynomial growth of order α (or higher) under the Lévy measure of \mathfrak{J}_α , plays a crucial role in determining this lower bound. Consider a $v \in \mathfrak{U}_{\text{sm}}$ as in Remark 2, and suppose that $\beta > 0$.

Then it is shown in Lemma 5.7 (b) of [4] that

$$\int_{\mathbb{R}^m} (\langle e, M^{-1}x \rangle^+)^p \bar{\pi}_v(dx) < \infty \quad \text{for some } p > 0 \quad \implies \quad p < \alpha - 1. \quad (20)$$

We use this property in the proof of Theorem 4 which follows. To simplify the notation, for a function f which is integrable under $\bar{\pi}_v$, we let $\bar{\pi}_v(f) := \int_{\mathbb{R}^m} f(x) \bar{\pi}_v(dx)$.

Theorem 4. *We assume $\beta > 0$. Suppose that under some $v \in \mathfrak{U}_{\text{sm}}$ such that $\Gamma v = 0$ a.e. the SDE in (1) has a unique weak solution which is an open-set irreducible and aperiodic C_b -Feller process. Then the process $\{X_t\}_{t \geq 0}$ is polynomially ergodic. In particular, there exists a positive constant C_2 not depending on v , such that for all $\varepsilon > 0$ we have*

$$\|P_t^v(x, \cdot) - \bar{\pi}_v(\cdot)\|_{\text{TV}} \geq C_2 \left(\frac{t \vee 1}{\varepsilon} + |x|^{\alpha - \varepsilon} \right)^{\frac{1 - \alpha}{1 - \varepsilon}} \quad \forall (t, x) \in \mathbb{R}_+ \times \mathbb{R}^m.$$

Proof. The proof uses [9, Theorem 5.1] and some results from [4]. Recall the function ψ , and define

$$\check{\chi}(t) := 1 + \psi(t), \quad \text{and} \quad \chi(t) := -\check{\chi}(-t).$$

Also, we scale $\chi(t)$ using $\chi_R(t) := R + \chi(t - R)$, $R \in \mathbb{R}$. Thus, $\chi_R(t) = t$ for $t \leq R - 1$ and $\chi_R(t) = R - \frac{1}{2}$ for $t \geq R$.

Let

$$F(x) := \check{\chi}(\langle e, M^{-1}x \rangle), \quad \text{and} \quad F_{\kappa, R}(x) := \chi_R \circ F^\kappa(x), \quad x \in \mathbb{R}^m, \quad R > 0,$$

where $F^\kappa(x)$ denotes the κ^{th} power of $F(x)$, with $\kappa > 0$.

Using the same notation as in [9, Theorem 5.1] whenever possible, we define $G(x) := F^{\alpha-\varepsilon}(x)$, for $\varepsilon \in (0, \alpha - 1)$. Then $\bar{\pi}_v(F^{\alpha-\varepsilon}) = \infty$ by (20). Applying the Itô formula to (19) we obtain

$$\mathbb{E}_x^v[V_{\alpha-\varepsilon}(X_t)] - V_{\alpha-\varepsilon}(x) \leq C_0(\alpha - \varepsilon)t, \quad x \in \mathbb{R}^m.$$

Since $F^{\alpha-\varepsilon} \leq \bar{C}_0 V_{\alpha-\varepsilon}$ for some constant $\bar{C}_0 \geq 1$, the preceding inequality implies that

$$\mathbb{E}_x^v[F^{\alpha-\varepsilon}(X_t)] \leq \bar{C}_0(C_0(\alpha - \varepsilon)t + V_{\alpha-\varepsilon}(x)) =: g(x, t).$$

Next, we compute a suitable lower bound $f(t)$ for $\bar{\pi}_v(\{x: G(x) \geq t\})$. We have

$$\begin{aligned} \mathcal{A}^v F_{1,R}(x) &= \mathcal{J}_\alpha F_{1,R}(x) + \chi'_R(F(x)) \langle b_v(x), \nabla F(x) \rangle \\ &= \mathcal{J}_\alpha F_{1,R}(x) + \chi'_R(F(x)) \check{\chi}'(\langle e, M^{-1}x \rangle) (-\beta + \langle e, x \rangle^-). \end{aligned} \quad (21)$$

Integrating (21) with respect to $\bar{\pi}_v$, and replacing the variable R with t , we obtain

$$\beta \bar{\pi}_v(\chi'_t(F)h) = \bar{\pi}_v(\mathcal{J}_\alpha F_{1,t}) + \bar{\pi}_v(\chi'_t(F)\tilde{h}), \quad (22)$$

where

$$h(x) := \check{\chi}'(\langle e, M^{-1}x \rangle), \quad \text{and} \quad \tilde{h}(x) := h(x) \langle e, x \rangle^-.$$

Taking limits as $t \rightarrow \infty$ in (22), we obtain

$$\beta \bar{\pi}_v(h) = \bar{\pi}_v(\mathcal{J}_\alpha F) + \bar{\pi}_v(\tilde{h}). \quad (23)$$

Subtracting (22) from (23), gives

$$\beta \bar{\pi}_v(h - \chi'_t(F)h) = \bar{\pi}_v(\mathcal{J}_\alpha(F - F_{1,t})) + \bar{\pi}_v(\tilde{h} - \chi'_t(F)\tilde{h}). \quad (24)$$

Note that all the terms in this equation are nonnegative. Moreover, $\mathcal{J}_\alpha(F - F_{1,t})(x)$ is nonnegative by convexity, and thus

$$\begin{aligned} \bar{\pi}_v(\mathcal{J}_\alpha(F - F_{1,t})) &\geq \inf_{x \in \mathcal{B}} (\mathcal{J}_\alpha(F - F_{1,t})(x)) \bar{\pi}_v(\mathcal{B}) \\ &\geq \mathcal{J}_\alpha(F - F_{1,t})(0) \bar{\pi}_v(\mathcal{B}). \end{aligned} \quad (25)$$

It is straightforward to show that $\mathcal{J}_\alpha(F - F_{1,t})(0) \geq \hat{\kappa} t^{1-\alpha}$ for some positive constant $\hat{\kappa}$. Therefore, by (24)–(25) and the definition of the functions F , $F_{1,R}$ and h , we obtain

$$\begin{aligned}
\bar{\pi}_v(\{x: \langle e, M^{-1}x \rangle > t\}) &\geq \bar{\pi}_v(h - \chi'_t(F)h) \\
&\geq \beta^{-1} \bar{\pi}_v(\mathcal{B}) \mathfrak{J}_\alpha(F - F_{1,t})(0) \\
&\geq \hat{\kappa} t^{1-\alpha}.
\end{aligned} \tag{26}$$

Therefore, by (26), we have

$$\begin{aligned}
\bar{\pi}_v(\{x: G(x) \geq t\}) &= \bar{\pi}_v(\{x: (\langle e, M^{-1}x \rangle)^{\alpha-\varepsilon} > t\}) \\
&= \bar{\pi}_v(\{x: \langle e, M^{-1}x \rangle > t^{\frac{1}{\alpha-\varepsilon}}\}) \\
&\geq \hat{\kappa} t^{\frac{1-\alpha}{\alpha-\varepsilon}} =: f(t).
\end{aligned}$$

Next we solve $yf(y) = 2g(x, t)$ for $y = y(t)$, and this gives us $y = (\hat{\kappa}^{-1}2g(x, t))^{\frac{\alpha-\varepsilon}{1-\varepsilon}}$, and

$$f(y) = \hat{\kappa}(\hat{\kappa}^{-1}2g(x, t))^{\frac{1-\alpha}{1-\varepsilon}} = \bar{C}_1(C_0(\alpha - \varepsilon)t + V_{\alpha-\varepsilon}(x))^{\frac{1-\alpha}{1-\varepsilon}},$$

with

$$\bar{C}_1 := (2\bar{C}_0)^{\frac{1-\alpha}{1-\varepsilon}} \hat{\kappa}^{\frac{\alpha-\varepsilon}{1-\varepsilon}}.$$

Therefore, by [9, Theorem 5.1], and since ε is arbitrary, we have

$$\begin{aligned}
\|P_t^v(x, \cdot) - \bar{\pi}_v(\cdot)\|_{\text{TV}} &\geq f(y) - \frac{g(x, t)}{y} \\
&= \frac{\bar{C}_1}{2} (C_0(\alpha - \varepsilon)t + V_{\alpha-\varepsilon}(x))^{\frac{1-\alpha}{1-\varepsilon}}
\end{aligned} \tag{27}$$

for all $t \geq 0$ and $\varepsilon \in (0, \alpha - 1)$.

As shown in the proof of [4, Theorem 3.4], there exists a positive constant κ'_0 , not depending on ε , such that

$$C_0(\alpha - \varepsilon) \geq \kappa'_0(1 + \varepsilon^{-1}). \tag{28}$$

Thus the result follows by (27)–(28). \square

4 Ergodic properties of the limiting SDEs arising from queueing models with service interruptions

The limiting equations of multiclass $G/M/n + M$ queues with asymptotically negligible service interruptions under the \sqrt{n} -scaling in the Halfin–Whitt regime are Lévy-driven SDEs of the form

$$dX_t = b(X_t, U_t) dt + \sigma dW_t + dL_t, \quad X_0 = x \in \mathbb{R}^m. \tag{29}$$

Here, the drift b is as in Section 2, σ is a nonsingular diagonal matrix, and $\{L_t\}_{t \geq 0}$ is a compound Poisson process, with a drift ϑ , and a finite Lévy measure $\eta(dy)$ which

is supported on a half-line of the form $\{tw : t \in [0, \infty)\}$, with $\langle e, M^{-1}w \rangle > 0$. This can be established as in Theorem 6 in Section 5, assuming that the control is of the form $U_t = v(X_t)$ for a map $v : \mathcal{K}_+ \rightarrow \Delta$, such that $b_v(x)$ is locally Lipschitz, when the scaling is of order \sqrt{n} (see also Section 4.2 of [4]).

As we explain later, under any stationary Markov control, the SDE in (29) has a unique strong solution which is an open-set irreducible and aperiodic strong Feller process. Therefore, as far as the study of the process $\{X_t\}_{t \geq 0}$ is concerned, we do not need to impose a local Lipschitz continuity condition on the drift, but can allow the control to be any element of \mathcal{U}_{sm} .

There are two important parameters to consider. The first is the parameter θ_c , which is defined by

$$\theta_c := \sup \{ \theta \in \Theta_c \}, \quad \text{with} \quad \Theta_c := \left\{ \theta > 0 : \int_{\mathcal{B}^c} |y|^\theta \eta(dy) < \infty \right\}.$$

The second is the *effective spare capacity*, defined as

$$\tilde{\beta} := -\langle e, M^{-1}\tilde{\ell} \rangle,$$

where

$$\tilde{\ell} := \begin{cases} \ell + \vartheta + \int_{\mathcal{B}^c} y \eta(dy), & \text{if } \int_{\mathcal{B}^c} |y| \eta(dy) < \infty \\ \ell + \vartheta, & \text{otherwise.} \end{cases}$$

Suppose that $v \in \mathcal{U}_{\text{sm}}$ is such that $\Gamma v(x) = 0$ a.e. x in \mathbb{R}^m . Then as shown in Lemma 5.7 of [4], the process $\{X_t\}_{t \geq 0}$ controlled by v cannot have an invariant probability measure $\bar{\pi}_v$ unless $1 \in \Theta_c$ and $\tilde{\beta} > 0$, and moreover,

$$\int_{\mathbb{R}^m} ((e, M^{-1}x)^+)^p \bar{\pi}_v(dx) < \infty \quad \text{for some } p > 0 \quad \implies \quad p+1 \in \Theta_c.$$

In addition, $\tilde{\beta} = \int_{\mathbb{R}^m} \langle e, x \rangle^- \bar{\pi}_v(dx)$ [4, Theorem 3.4 (b)]. Conversely, $1 \in \Theta_c$ and $\tilde{\beta} > 0$ are sufficient for $\{X_t\}_{t \geq 0}$ to have an invariant probability measure $\bar{\pi}_v$ under any constant control v , and $\bar{\pi}_v \in \mathcal{P}_p(\mathbb{R}^m)$ if $p+1 \in \Theta_c$ (see Theorems 3.2 and 3.4 (b) in [4]).

On the other hand, if $\Gamma > 0$, that is, it has positive diagonal elements, then $\{X_t\}_{t \geq 0}$ is geometrically ergodic under any constant Markov control, and $\bar{\pi}_v \in \mathcal{P}_\theta(\mathbb{R}^m)$ for any $\theta \in \Theta_c$ [4, Theorem 3.5]. This bound is tight since, in general, if under some Markov control v the process $\{X_t\}_{t \geq 0}$ has an invariant probability measure $\bar{\pi}_v \in \mathcal{P}_p(\mathbb{R}^m)$, then necessarily $p \in \Theta_c$.

We extend the results derived for constant Markov controls in [4] to all controls in \mathcal{U}_{sm} . Recall the definition in (4). Let

$$\tilde{b}(x, u) := b(x, u) + \tilde{\ell} - \ell,$$

and $\tilde{b}_v(x) = \tilde{b}(x, v(x))$ for $v \in \mathcal{U}_{\text{sm}}$. As explained in Section 2, we assume, without loss of generality, that the constant term in \tilde{b} is as in (6) with β replaced by $\tilde{\beta}$.

We define the operator \mathcal{A}^u on C^2 functions by

$$\mathcal{A}^u f(x) := \mathcal{L}^u f(x) + \mathfrak{J}_\eta f(x), \quad (x, u) \in \mathbb{R}^m \times \Delta,$$

where

$$\mathcal{L}^u f(x) = \frac{1}{2} \text{trace}(\sigma \sigma' \nabla^2 f(x)) + \langle \tilde{b}(x, u), \nabla f(x) \rangle, \quad (x, u) \in \mathbb{R}^m \times \Delta, \quad (30)$$

and

$$\mathfrak{J}_\eta f(x) := \int_{\mathbb{R}^m} \mathfrak{d}f(x; y) \eta(dy), \quad x \in \mathbb{R}^m.$$

Also, \mathcal{L}^v is defined as in (30) by replacing u with $v(x)$ for a control $v \in \mathfrak{U}_{\text{sm}}$, and analogously for \mathcal{A}^v .

It follows from the results in [8] that, for any $v \in \mathfrak{U}_{\text{sm}}$, the diffusion

$$d\tilde{X}_t = \tilde{b}(\tilde{X}_t, v(\tilde{X}_t)) dt + \sigma(\tilde{X}_t) dW_t, \quad \tilde{X}_0 = x \in \mathbb{R}^d \quad (31)$$

has a unique strong solution. Also, as shown in [14], since the Lévy measure is finite, the solution of (29) can be constructed in a piecewise fashion using the solution of (31) (see also [10]). It thus follows that, under any stationary Markov control, (29) has a unique strong solution which is a strong Markov process. In addition, its transition probability $P_t^v(x, dy)$ satisfies $P_t^v(x, B) > 0$ for all $t > 0, x \in \mathbb{R}^m$ and $B \in \mathcal{B}(\mathbb{R}^m)$ with positive Lebesgue measure. Thus, under any $v \in \mathfrak{U}_{\text{sm}}$, the process $\{X_t\}_{t \geq 0}$ is open-set irreducible and aperiodic.

Recall Definition 1. In order to handle the second order derivatives in \mathcal{A}^u we need to scale the Lyapunov function V_p . This is done as follows. With ψ as in Definition 1, we define

$$\psi_\delta(t) := \psi(\delta t), \quad \text{and} \quad \Psi_\delta(x) := \sum_{i \in \mathcal{J}} \frac{\psi_\delta(x_i)}{\mu_i}, \quad \delta \in (0, 1],$$

and let

$$\mathcal{V}_{p, \delta}(x) := \left(\delta^2 \Psi(-x) + \Psi_\delta(x) + \frac{m}{\min_{i \in \mathcal{J}} \mu_i} \right)^p.$$

Note that $\mathcal{V}_{1, \delta}$ is bounded away from 0 uniformly in $\delta \in (0, 1]$. Here we use the inequality $\sum_{i \in \mathcal{J}} \psi'_\delta(x_i) x_i \geq \delta \|x^+\|_1 - \frac{m}{2}$. Then, under the assumption that $\tilde{\beta} > 0$, the drift inequalities take the form

$$\begin{aligned} & \langle \tilde{b}(x, u), \nabla \mathcal{V}_{p, \delta}(x) \rangle \\ & \leq \begin{cases} p\delta \left(\delta \tilde{\beta} + \frac{m}{2\delta} (1 + \delta^2) - \delta \|x\|_1 \right) \mathcal{V}_{p-1, \delta}(x) & \forall x \in \mathcal{K}_-, \\ -p\delta \left(\frac{\tilde{\beta}}{m} - \delta \tilde{\beta} - \delta \frac{m}{2} + \delta \|x^-\|_1 \right) \mathcal{V}_{p-1, \delta}(x) & \forall (x, u) \in \mathcal{K}_+ \times \Delta. \end{cases} \end{aligned} \quad (32)$$

The following result is analogous to Theorem 2.

Theorem 5. Assume $\tilde{\beta} > 0$, and $1 \in \Theta_c$. Let $p \in \Theta_c$ with $p > 1$. Then the following hold.

(a) There exists $\delta > 0$, a positive constant \tilde{C}_0 , and a compact set K such that

$$\mathcal{A}^u \mathcal{V}_{p,\delta}(x) \leq \tilde{C}_0 \mathbb{1}_K(x) - p\delta \frac{\tilde{\beta}}{2m} \mathcal{V}_{p-1,\delta}(x) \quad \forall (x, u) \in \mathbb{R}^m \times \Delta. \quad (33)$$

(b) Under any $v \in \mathfrak{U}_{\text{sm}}$, the process $\{X_t\}_{t \geq 0}$ in (1) admits a unique invariant probability measure $\tilde{\pi}_v \in \mathcal{P}(\mathbb{R}^m)$.

(c) For any $\theta \in \Theta_c$ there exists a constant $\tilde{C}_1(\theta)$ depending only on θ , such that, under any $v \in \mathfrak{U}_{\text{sm}}$, the process $\{X_t\}_{t \geq 0}$ in (1) satisfies

$$\|P_t^v(x, \cdot) - \tilde{\pi}_v(\cdot)\|_{\text{TV}} \leq \tilde{C}_1(\theta)(t \vee 1)^{1-\theta} |x|^\theta \quad \forall x \in \mathbb{R}^m.$$

Proof. It is straightforward to show that $\psi''_\delta(t) \leq 2\delta^2$ and $\psi'_\delta(t) \leq \delta$ for all $t \in \mathbb{R}$. An easy calculation then shows that there exists a positive constant C such that

$$\text{trace}(\sigma \sigma' \nabla^2 \mathcal{V}_{p,\delta}(x)) \leq Cp^2 \delta^2 (\mathcal{V}_{p-1,\delta}(x) + \mathcal{V}_{p-2,\delta}(x)) \quad (34)$$

for all $p \geq 1$ and $x \in \mathbb{R}^m$. Recall that $\mathcal{V}_{1,\delta}$ is bounded away from 0 uniformly in $\delta \in (0, 1]$. This of course implies that $\mathcal{V}_{p-2,\delta}$ is bounded by some fixed multiple of $\mathcal{V}_{p-1,\delta}$ for all $p \geq 1$. Therefore, (32) and (34) imply that for some small enough positive δ we can chose a positive constant \tilde{C}'_0 , and a compact set K' such that

$$\mathcal{L}^u \mathcal{V}_{p,\delta}(x) \leq \tilde{C}'_0 \mathbb{1}_{K'}(x) - p\delta \frac{3\tilde{\beta}}{4m} \mathcal{V}_{p-1,\delta}(x) \quad \forall (x, u) \in \mathbb{R}^m \times \Delta. \quad (35)$$

If $p \in \Theta_c$, then [4, Lemma 5.1] asserts that $\mathfrak{J}_\eta \mathcal{V}_{p,\delta}$ vanishes at infinity for $p < 2$, and $\mathfrak{J}_\eta \mathcal{V}_{p,\delta}$ is of order $|x|^{p-2}$ for $p \geq 2$. This together with (35) implies (33). The rest are as in the proof of Theorem 2. \square

If $\Gamma > 0$, then the arguments in the proof of Theorem 5 together with Lemma 2 show that the process $\{X_t\}_{t \geq 0}$ is geometrically ergodic uniformly over $v \in \mathfrak{U}_{\text{sm}}$. Thus we obtain the analogous results to Theorem 3. We omit the details which are routine.

Note that the assumption that the Lévy measure $\eta(dy)$ is supported on a half-line of the form $\{tw : t \in [0, \infty)\}$, with $\langle e, M^{-1}w \rangle > 0$ has not been used, and is not needed in Theorem 5. Under this assumption we can obtain a lower bound of the rate of convergence analogous to equation (3.9) in [4], by mimicking the arguments in that paper. We leave the details to the reader.

Remark 3. With heavy-tailed arrivals and asymptotically negligible service interruptions under the common $n^{1/\alpha}$ -scaling for $\alpha \in (1, 2)$, in the modified Halfin–Whitt regime, the limit process is an SDE driven by an anisotropic α -stable process (with independent α -stable components) as in (1), and a compound Poisson process with a finite Lévy measure as in (29). This can be established as in Theorem 6, under the same scaling assumptions in Section 4.2 of [4]. Thus the generator is given by

$$\hat{A}^u f(x) := \langle \tilde{b}(x, u), \nabla f(x) \rangle + \mathfrak{J}_\eta f(x) + \mathfrak{J}_\alpha f(x),$$

and \hat{A}^v is defined analogously by replacing u with $v(x)$ for $v \in \tilde{\mathcal{U}}_{\text{sm}}$.

To study this equation, we use the Lyapunov function V_p in Definition 1, with $p \in [1, \alpha) \cap \Theta_c$. Following the proof of Theorem 5, and also using Lemma 3, it follows that there exists $\delta > 0$ sufficiently small, a constant \hat{C}_0 and a compact set \hat{K} such that

$$\hat{A}^u V_p(x) \leq \hat{C}_0 \mathbb{1}_{\hat{K}}(x) - p \frac{\tilde{\beta}}{2m} V_{p-1}(x) \quad \forall (x, u) \in \mathbb{R}^m \times \Delta.$$

Thus, (18) holds for any ε such that $\alpha - \varepsilon \in \Theta_c$. The results of Theorem 3 also follow provided we select $p \in [1, \alpha) \cap \Theta_c$. However the lower bound is not necessarily the one in Theorem 4. Instead we can obtain a lower bound in the form of equation (3.9) in [4].

5 Multiclass $G/M/n + M$ queues with heavy-tailed arrivals

As in [4, Subsection 4.1], consider $G/M/n + M$ queues with m classes of customers and one server pool of n parallel servers. Customers of each class form their own queue and are served in the first-come first-served (FCFS) service discipline. Customers of different classes are scheduled to receive service under the work conserving constraint, that is, non-idling whenever customers are in queue. We assume that the arrival process of each class is renewal with heavy-tailed interarrival times. The service and patience times are exponentially distributed with class-dependent rates. The arrival, service and abandonment processes of each class are mutually independent.

We consider a sequence of such queueing models indexed by n and let $n \rightarrow \infty$. Let A_i^n , $i = 1, \dots, m$, be the arrival process of class- i customers with arrival rate λ_i^n . Assume that A_i^n 's are mutually independent. Define the FCLT-scaled arrival processes $\hat{A}^n = (\hat{A}_1^n, \dots, \hat{A}_m^n)'$ by $\hat{A}_i^n := n^{-1/\alpha}(A_i^n - \lambda_i^n \varpi)$, $i = 1, \dots, m$, where $\varpi(t) \equiv t$ for each $t \geq 0$, and $\alpha \in (1, 2)$. We assume that

$$\lambda_i^n/n \rightarrow \lambda_i > 0, \quad \text{and} \quad \ell_i^n := n^{-1/\alpha}(\lambda_i^n - n\lambda_i) \rightarrow \ell_i \in \mathbb{R}, \quad (36)$$

for each $i = 1, \dots, m$, as $n \rightarrow \infty$, and that the arrival processes satisfy an FCLT

$$\hat{A}^n \Rightarrow \hat{A} = (\hat{A}_1, \dots, \hat{A}_m)' \quad \text{in } (D_m, M_1), \text{ as } n \rightarrow \infty,$$

where the limit processes \hat{A}_i , $i = 1, \dots, m$, are mutually independent symmetric α -stable processes with $\hat{A}_i(0) \equiv 0$, and \Rightarrow denotes weak convergence and (D_m, M_1) is the space of \mathbb{R}^m -valued càdlàg functions endowed with the product M_1 topology [15]. The processes \hat{A}_i have the same stability parameter α , with possibly different “scale” parameters ξ_i . Note that if the arrival process of each class is renewal with

regularly varying interarrival times of parameter α , then we obtain the above limit process. Let μ_i and γ_i be the service and abandonment rates for class- i customers, respectively.

The modified Halfin-Whitt regime. The parameters satisfy

$$n^{1-1/\alpha}(1-\rho^n) \xrightarrow{n \rightarrow \infty} \rho = -\sum_{i=1}^m \frac{\ell_i}{\mu_i},$$

where $\rho^n := \sum_{i=1}^m \frac{\lambda_i^n}{n\mu_i}$ is the aggregate traffic intensity. This follows from (36). Let $\rho_i := \lambda_i/\mu_i$ for $i \in J$.

Let $X^n = (X_1^n, \dots, X_d^n)'$, $Q^n = (Q_1^n, \dots, Q_d^n)'$, and $Z^n = (Z_1^n, \dots, Z_d^n)'$ be the processes counting the number of customers of each class in the system, in queue, and in service, respectively. We consider work-conserving scheduling policies that are non-anticipative and allow preemption (namely, service of a customer can be interrupted at any time to serve some other class of customers and will be resumed at a later time). Scheduling policies determine the allocation of service capacity, i.e., the Z^n process, which must satisfy the condition that $\langle e, Z^n \rangle = \langle e, X^n \rangle \wedge n$ at each time, as well as the balance equations $X_i^n = Q_i^n + Z_i^n$ for each i .

Define the FCLT-scaled processes $\hat{X}^n = (\hat{X}_1^n, \dots, \hat{X}_d^n)'$, $\hat{Q}^n = (\hat{Q}_1^n, \dots, \hat{Q}_d^n)'$, and $\hat{Z}^n = (\hat{Z}_1^n, \dots, \hat{Z}_d^n)'$ by

$$\hat{X}_i^n := n^{-1/\alpha}(X_i^n - \rho_i n), \quad \hat{Q}_i^n := n^{-1/\alpha}Q_i^n, \quad \hat{Z}_i^n := n^{-1/\alpha}(Z_i^n - \rho_i n).$$

We need the following extension of Theorem 1.1 in [13]. Let $\phi: D([0, T], \mathbb{R}^m) \rightarrow D([0, T], \mathbb{R}^m)$ denote the mapping $x \mapsto y$ defined by the integral representation

$$y(t) = x(t) + \int_0^t h(y(s)) ds, \quad t \geq 0.$$

It is shown in [13, Theorem 1.1] that the mapping ϕ is continuous in the Skorohod M_1 topology when $m = 1$ and the function h is Lipschitz continuous. The lemma which follows extends this result to functions $h: \mathbb{R}^m \rightarrow \mathbb{R}^m$ which are locally Lipschitz continuous and have at most linear growth.

Lemma 4. *Assume that h is locally Lipschitz and has at most linear growth. Then the mapping ϕ defined above is continuous in (D_m, M_1) , the space $D([0, T], \mathbb{R}^m)$ endowed with the product M_1 topology.*

Proof. Assume that $x_n \rightarrow x$ in D_m with the product M_1 topology as $n \rightarrow \infty$. Let x^i be the i^{th} component of x , and similarly for x_n^i . Let

$$G_x := \{(z, t) \in \mathbb{R}^m \times [0, T]: z^i \in [x^i(t-), x^i(t)] \text{ for each } i = 1, \dots, m\},$$

be the (weak) graph of x , and similarly, G_{x_n} for x_n ; see Chapter 12.3.1 in [15]. Then following the proof of Theorem 1.2 in [13], it can be shown that there exist parametric representations (u, r) and (u_n, r_n) of x and x_n , that map $[0, 1]$ onto the graphs G_x and G_{x_n} of x and x_n , respectively, and satisfy the properties below. In

the construction of the time component as in Lemma 4.3 of [13], the discontinuity points of all the x^j components need to be included, and then the spatial component can be done similarly as in the proof of that lemma.

- (a) The time (domain) components $r, r_n \in C([0, 1], [0, T])$ are nondecreasing functions satisfying $r(0) = r_n(0) = 0$ and $r_n(1) = r_n(1) = T$, and such that r and r_n are absolutely continuous with respect to Lebesgue measure on $[0, 1]$.
- (b) The derivatives r' and r'_n exist for all n and satisfy $\|r'\|_\infty < \infty$, $\sup_n \|r'_n\|_\infty < \infty$, and $\|r'_n - r'\|_{L^1} \rightarrow 0$, where $\|r\|_\infty := \sup_{s \in [0, 1]} |r(s)|$, and $\|\cdot\|_{L^1}$ denotes the L^1 norm.
- (c) The spatial components $u = (u^1, \dots, u^m)$ and $u_n = (u_n^1, \dots, u_n^m)$, $n \in \mathbb{N}$, lie in $C([0, 1], \mathbb{R}^m)$, and satisfy $u(0) = x(0)$, $u(1) = x(T)$, $u_n(0) = x_n(0)$, $u_n(1) = x_n(T)$, and $\|u_n - u\|_\infty \rightarrow 0$ as $n \rightarrow \infty$.

As shown in the proof of Theorem 1.1 in [13], there exist parametric representations (u_y, r_y) and (u_{y_n}, r_{y_n}) of y and y_n , respectively, with $r_y = r$ and $r_{y_n} = r_n$, satisfying

$$u_{y_n}(s) = u_n(s) + \int_0^s h(u_{y_n}(w)) r'_n(w) dw, \quad s \in [0, 1], \quad (37)$$

and similarly for $u_y(s)$. Here, (u, r) and (u_n, r_n) are the parametric representations of x and x_n , respectively, whose properties are summarized above.

Since $x_n \rightarrow x$ in (D_m, M_1) as $n \rightarrow \infty$, we have $\sup_n \|u_n\|_\infty < \infty$. Taking norms in (37), and using also the property $\sup_n \|r'_n\|_\infty < \infty$, and the linear growth of h , an application of Gronwall's lemma shows that $\sup_n \|u_{y_n}\|_\infty \leq R$ for some constant R . Enlarging this constant if necessary, we may also assume that $\|u_y\|_\infty \leq R$. By the representation in (37), we have

$$\begin{aligned} |u_{y_n}(s) - u_y(s)| &\leq |u_n(s) - u(s)| + \left| \int_0^s (h(u_{y_n}(w)) - h(u_y(w))) r'_n(w) dw \right| \\ &\quad + \left| \int_0^s h(u_y(w)) r'_n(w) dw - \int_0^s h(u_y(w)) r'(w) dw \right|. \end{aligned}$$

Let κ_R be a Lipschitz constant of h on the ball \mathcal{B}_R . Then, applying Gronwall's lemma once more, we obtain

$$\|u_{y_n} - u_y\|_\infty \leq \left(\|u_n - u\|_\infty + \|r'_n - r'\|_{L^1} \sup_{\mathcal{B}_R} h \right) e^{\kappa_R \|r'_n\|_\infty} \xrightarrow{n \rightarrow \infty} 0.$$

This completes the proof. \square

Remark 4. Suppose h, x, x_n , and y are as in Lemma 4, but y_n satisfies

$$y_n(t) = x_n(t) + \int_0^t h_n(y_n(s)) ds, \quad t \geq 0,$$

for some sequence h_n which converges to h uniformly on compacta. Then a slight variation of the proof of Lemma 4, shows that $y_n \rightarrow y$ in D_m .

Control approximation. Given a continuous map $v: \mathcal{X}_+ \rightarrow \Delta$, we construct a stationary Markov control for the n -system which approximates it in a suitable manner.

Recall that $\langle e, \rho \rangle = 1$. Let

$$\mathcal{X}_n := \{n^{-1/\alpha}(y - \rho n) : y \in \mathbb{Z}_+^m, \langle e, y \rangle > n\},$$

and $\mathfrak{Z}_n = \mathfrak{Z}_n(\hat{x})$ denote the set of work-conserving actions at $\hat{x} \in \mathcal{X}_n$. It is clear that a work-conserving action $\hat{z}^n \in \mathfrak{Z}_n(\hat{x})$ can be parameterized via a map $\widehat{U}^n: \mathcal{X}_+ \rightarrow \Delta$, satisfying

$$\hat{z}_i^n(\hat{x}) = \hat{x}_i - \langle e, \hat{x} \rangle^+ \widehat{U}_i^n(\hat{x}). \quad (38)$$

Consider the mapping defined in (38) from $\hat{z}^n \in \mathfrak{Z}_n(\hat{x})$ to \widehat{U}^n , and denote its image as $\widehat{\mathcal{U}}_n(\hat{x})$. Let

$$\widehat{U}^n[v](\hat{x}) \in \underset{u \in \widehat{\mathcal{U}}_n(\hat{x})}{\text{Arg min}} \left| \langle e, \hat{x} \rangle u - \langle e, \hat{x} \rangle v(\hat{x}) \right|, \quad \hat{x} \in \mathcal{X}_n. \quad (39)$$

The function $\widehat{U}^n[v]$ has the following property. There exists a constant \check{c} such that with $\widehat{\mathcal{B}}_n$ denoting the ball of radius $\check{c}n^{\check{\alpha}}$ in \mathbb{R}^m , with $\check{\alpha} := 1 - 1/\alpha$, then

$$\sup_{\hat{x} \in \widehat{\mathcal{B}}_n \cap \mathcal{X}_n} \left| \langle e, \hat{x} \rangle \widehat{U}^n[v](\hat{x}) - \langle e, \hat{x} \rangle v(\hat{x}) \right| \leq n^{-1/\alpha}. \quad (40)$$

We have the following functional limit theorem.

Theorem 6. *Let $v \in \widetilde{\mathfrak{U}}_{\text{sm}}$. Under any stationary Markov control $\widehat{U}^n[v]$ defined in (39), and provided there exists $X(0)$ such that $\widehat{X}^n(0) \Rightarrow X(0)$ as $n \rightarrow \infty$, we have*

$$\widehat{X}^n \Rightarrow X \quad \text{in } (D_m, M_1) \quad \text{as } n \rightarrow \infty,$$

where the limit process X is the unique strong solution to the SDE in (1). The parameters in the drift are given by ℓ_i in (36), μ_i , and γ_i , for $i = 1, \dots, m$.

Proof. The FCLT-scaled processes \widehat{X}_i^n , $i = 1, \dots, m$, can be represented as

$$\widehat{X}_i^n(t) = \widehat{X}_i^n(0) + \ell_i^n t - \mu_i \int_0^t \widehat{Z}_i^n(s) ds - \gamma_i \int_0^t \widehat{Q}_i^n(s) ds + \widehat{A}_i^n(t) - \widehat{M}_{S,i}^n(t) - \widehat{M}_{R,i}^n(t)$$

where ℓ_i^n is defined in (36), with

$$\widehat{M}_{S,i}^n(t) = n^{-1/\alpha} \left(S_i^n \left(\mu_i \int_0^t \widehat{Z}_i^n(s) ds \right) - \mu_i \int_0^t Z_i^n(s) ds \right),$$

$$\widehat{M}_{R,i}^n(t) = n^{-1/\alpha} \left(R_i^n \left(\gamma_i \int_0^t \widehat{Q}_i^n(s) ds \right) - \gamma_i \int_0^t Q_i^n(s) ds \right),$$

and S_i^n, R_i^n , $i = 1, \dots, m$, are mutually independent rate-one Poisson processes, representing the service and reneging (abandonment), respectively.

The result can be established by mimicking the arguments in the proof of in [4, Theorem 4.1], and applying Lemma 4 and Remark 4, using the function

$$h_n(x) := \ell^n + M(x - \langle e, x \rangle^+ \hat{U}^n[v](x)) - \langle e, x \rangle^+ \Gamma \hat{U}^n[v](x),$$

and the bound in (40). \square

6 Concluding remarks

We have extended some of the results in [4] stated for constant controls, to stationary Markov controls resulting in a locally Lipschitz drift in the case of SDEs driven by α -stable processes, and to all stationary Markov controls in the case of SDEs driven by a Wiener process and a compound Poisson process. The results in this paper can also be viewed as an extension of some results in [3]. However, the work in [3] also studies the prelimit process and establishes tightness of the stationary distributions. To the best of our knowledge, this is an open problem for systems with arrival processes which are renewal with heavy-tailed interarrival times (no second moments). This problem is very important and worth pursuing.

Acknowledgements This research was supported in part by the Army Research Office through grant W911NF-17-1-001, in part by the National Science Foundation through grants DMS-1715210, CMMI-1538149 and DMS-1715875, and in part by Office of Naval Research through grant N00014-16-1-2956. Financial support through Croatian Science Foundation under the project 8958 (for N. Sandrić) is gratefully acknowledged.

References

1. Albeverio, S., Brzeźniak, Z., Wu, J.L.: Existence of global solutions and invariant measures for stochastic differential equations driven by Poisson type noise with non-Lipschitz coefficients. *J. Math. Anal. Appl.* **371**(1), 309–322 (2010). DOI 10.1016/j.jmaa.2010.05.039
2. Arapostathis, A., Biswas, A., Caffarelli, L.: The Dirichlet problem for stable-like operators and related probabilistic representations. *Comm. Partial Differential Equations* **41**(9), 1472–1511 (2016). DOI 10.1080/03605302.2016.1207084
3. Arapostathis, A., Hmedi, H., Pang, G.: On uniform exponential ergodicity of Markovian multiclass many-server queues in the Halfin–Whitt regime. *ArXiv e-prints* **1812.03528** (2018)
4. Arapostathis, A., Pang, G., Sandrić, N.: Ergodicity of a Lévy-driven SDE arising from multiclass many-server queues. *Ann. Appl. Probab.* **29**(2), 1070–1126 (2019). DOI 10.1214/18-AAP1430
5. Dieker, A.B., Gao, X.: Positive recurrence of piecewise Ornstein–Uhlenbeck processes and common quadratic Lyapunov functions. *Ann. Appl. Probab.* **23**(4), 1291–1317 (2013). DOI 10.1214/12-aap870
6. Douc, R., Fort, G., Guillin, A.: Subgeometric rates of convergence of f -ergodic strong Markov processes. *Stochastic Process. Appl.* **119**(3), 897–923 (2009). DOI 10.1016/j.spa.2008.03.007

7. Gamarnik, D., Stolyar, A.L.: Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime: asymptotics of the stationary distribution. *Queueing Syst.* **71**(1-2), 25–51 (2012). DOI 10.1007/s11134-012-9294-x
8. Gyöngy, I., Krylov, N.: Existence of strong solutions for Itô's stochastic equations via approximations. *Probab. Theory Related Fields* **105**(2), 143–158 (1996). DOI 10.1007/BF01203833
9. Hairer, M.: Convergence of Markov Processes. Lecture Notes, University of Warwick (2016). Available at <http://www.hairer.org/notes/Convergence.pdf>
10. Li, C.W.: Lyapunov exponents of nonlinear stochastic differential equations with jumps. In: Stochastic inequalities and applications, *Progr. Probab.*, vol. 56, pp. 339–351. Birkhäuser, Basel (2003)
11. Meyn, S.P., Tweedie, R.L.: Stability of Markovian processes. II. Continuous-time processes and sampled chains. *Adv. in Appl. Probab.* **25**(3), 487–517 (1993). DOI 10.2307/1427521
12. Pang, G., Whitt, W.: Heavy-traffic limits for many-server queues with service interruptions. *Queueing Syst.* **61**(2-3), 167–202 (2009). DOI 10.1007/s11134-009-9104-2
13. Pang, G., Whitt, W.: Continuity of a queueing integral representation in the M_1 topology. *Ann. Appl. Probab.* **20**(1), 214–237 (2010). DOI 10.1214/09-AAP611
14. Skorokhod, A.V.: Asymptotic methods in the theory of stochastic differential equations, *Translations of Mathematical Monographs*, vol. 78. American Mathematical Society, Providence, RI (1989). Translated from the Russian by H. H. McFaden
15. Whitt, W.: Stochastic-process limits. An introduction to stochastic-process limits and their application to queues. Springer Series in Operations Research. Springer-Verlag, New York (2002)



Nudged Particle Filters in Multiscale Chaotic Systems with Correlated Sensor Noise

Ryne Beeson and N. Sri Namachchivaya

Abstract In this work we present recent and new results for the theory and algorithms of efficient estimation of the coarse-grain dynamics of a multiscale chaotic dynamical system, where observations may be limited both spatial and in time, and the observations are correlated with the slow states. The rigorous mathematical statement and convergence result with a rate of convergence to the reduced order filter problem is given for the case of correlated sensor noise. Based on this result, algorithms for efficient numerical solution of the filtering problem for the coarse-grain dynamics are provided. We then address a second issue, which presents itself in the case of chaotic systems and degrades particle filtering performance; the growth of small errors at an exponential rate. We solve this problem by introducing an optimal control problem for the solution of the proposal distribution and develop a numerical algorithm for its solution. The algorithms developed in this work are demonstrated on the widely used multiscale chaotic Lorenz 1996 model, that mimics mid-latitude convection.

1 Introduction

With the continual growth in computational power, higher dimensional, more complex physical models are more readily used in engineering and science applications. Simultaneously, there is more and more interest in applying controls to and estimating or forecasting these complex models. Often the physical models may have

Ryne Beeson

Department of Aerospace Engineering, University of Illinois at Urbana-Champaign,
Urbana, Illinois, USA

e-mail: rbeeson2@illinois.edu

N. Sri Namachchivaya

Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada

e-mail: nsnamachchivaya@uwaterloo.ca

multiple time and spatial scales. For instance, in climate and weather modeling, it is common practice to model the dynamics of the atmosphere or ocean on varying spatial grids with distinct time scale separations, and to use stochastic consistency, the additional of a stochastic process, to model the unresolved modes.

Whether one is interested in forecasting the states of their model, or applying control to the model, the need for estimating the states of the system at a time t given past observations of the system, is a critical problem. This is the problem of filtering theory, of which the main result is that the filter, a conditional measure, is characterized by the evolution of a certain stochastic differential equation (SDE) taking values in a probability measure space. General and flexible numerical algorithms have been devised for the solution of this SDE, but suffer from the curse of dimensionality. Hence the reduction of the model is crucial for improving the performance of these general filtering algorithms.

In the case where the model possesses large time-scale separations, one can leverage the theory of stochastic averaging to show that there exists a lower dimensional process X_t^0 that is close to the coarse-grain dynamics X_t , but is uncoupled from the fine-grain processes Z_t . In particular $X_t \Rightarrow X_t^0$ as a time-scale separation parameter tends to zero. Our interest lies in understanding whether we can show certain convergence results of the filter associated with X_t to a lower dimension filter; exploiting this property of $X_t \Rightarrow X_t^0$. Being able to show such results, allows potential insight into how more efficient lower dimensional filter algorithms can be created, and thus more accurate and tractable filtering methods for solving modern problems.

For instance, a motivating problem for the multiscale correlated filtering problem stems from atmospheric and climatology problems; for example, coupled atmosphere-ocean models, which immediately provide a multiscale model with fast atmospheric and slow ocean dynamics. In the case of climate prediction the ocean memory, due to its heat capacity, holds important information. Hence, the improved estimate of the ocean state, which is often the slow component, is of greater interest.

In particular, we are interested in working with the following problem. Consider the system of equations describing the multiple timescale correlated sensor noise problem, defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{Q})$, supporting an (unobserved) signal processes $(X_t^\varepsilon, Z_t^\varepsilon)$ and an observation process Y_t^ε , as follows

$$\begin{aligned} dX_t^\varepsilon &= b(X_t^\varepsilon, Z_t^\varepsilon)dt + \sigma(X_t^\varepsilon, Z_t^\varepsilon)dW_t, & X_0^\varepsilon &= x \in \mathbb{R}^m \\ dZ_t^\varepsilon &= \frac{1}{\varepsilon}f(X_t^\varepsilon, Z_t^\varepsilon)dt + \frac{1}{\sqrt{\varepsilon}}g(X_t^\varepsilon, Z_t^\varepsilon)dV_t, & Z_0^\varepsilon &= z \in \mathbb{R}^n \\ dY_t^\varepsilon &= h(X_t^\varepsilon, Z_t^\varepsilon)dt + \alpha dW_t + \gamma dU_t, & Y_0^\varepsilon &= 0 \in \mathbb{R}^d \end{aligned} \quad (1)$$

where $X_t^\varepsilon, Z_t^\varepsilon, Y_t^\varepsilon$ take values in $\mathbb{R}^m, \mathbb{R}^n$, and \mathbb{R}^d respectively for $m, n, d \in \mathbb{N}$. Here $0 < \varepsilon \ll 1$ is a timescale separation parameter; hence X_t^ε is a slow process, and Z_t^ε is a fast process. W_t, V_t, U_t are independent standard Brownian motions. The presence of $\alpha \neq 0$ indicates sensor-signal noise correlation in the observation process Y_t^ε . Assuming (1) possesses certain properties such that $X_t^\varepsilon \Rightarrow X_t^0$, to be properly defined in Section 2, and assuming the test function of interest is of the form $\varphi = \varphi(x)$, then

can we show that

$$\pi_t^{\varepsilon,x}(\varphi) \equiv \mathbb{E}_{\mathbb{Q}}[\varphi(X_t^\varepsilon, Z_t^\varepsilon) | \mathcal{Y}_t^\varepsilon] \Rightarrow \pi_t^0(\varphi) \equiv \mathbb{E}_{\mathbb{Q}}[\varphi(X_t^0) | \mathcal{Y}_t^\varepsilon],$$

where φ is an integrable test function, $\mathcal{Y}_t^\varepsilon \equiv \sigma(\{Y_s^\varepsilon | s \in [0, t]\})$ is the σ -algebra generated by the observation process, and $\pi_t^{\varepsilon,x}$ is the x -marginal of π_t^ε ? In particular, can a rate of convergence be shown for the case of correlated sensor-signal noise. And furthermore, what can be done to improve numerical methods for solving the filtering problem if (1) is a chaotic system or still high-dimensional after reduction.

The organization of this chapter is as follows. In Section 2 we properly introduce the filtering problem for the multiscale correlated sensor-signal noise problem. We present the main converge results and describe the mathematical techniques used to show the result, which includes the re-casting of the filter SDE into a backward stochastic partial differential equation (BSPDE) and finally into backward doubly stochastic differential equations (BDSDEs). In Section 3 we introduce the Lorenz 1996 model [23], which is an example of a chaotic system used to model a random multiscale natural system. It is a heuristic model that mimics mid-latitude atmospheric dynamics with microscopic convective processes. It is a useful tool for testing new data assimilation methods for use in numerical weather simulations owing to its transparency and low computational cost (c.f [1, 27, 15, 20]). We in fact present a stochastic version of the Lorenz 1996 model, where the stochastic forcing is justified by the need to account for unresolved modes [25]. The Lorenz '96 model will act as our testbed model for benchmarking various filtering methods in Section 6. Before leaving Section 3, we introduce the heterogeneous multiscale method (HMM) [9], which we use to numerical solve for the homogenized dynamics. We then compare the qualitative behavior of the full Lorenz '96 and homogenized models and verify that the relevant assumptions for applying HMM to the Lorenz '96 model are satisfied. In Section 4, we recall the derivation of the sequential importance sampling (SIS) particle filter (PF) for the solution of the filtering problem and then share an algorithm that combines the multiscale integration technique, HMM, with the SIS PF, which we simple refer to as PF from now on. This filtering algorithm is called the homogenized hybrid particle filter (HHPF). Lastly, in Section 4 we provide results regarding how to account for the case of correlated sensor-signal noise when filtering a continuous-time signal with sparse in-time discrete observations. In Section 5 we address a critical issue when applying a PF to a chaotic system with a sparse in-time observation process. To keep the number of particles low, yet attain desired accuracy, we introduce a control on the particles and solve an optimal control problem that results in an optimal proposal distribution for the PF. In Section 6 we demonstrate the aforementioned sequential monte carlo based methods on the Lorenz 1996 model with a correlated sensor-signal noise observation process. Section 7 provides concluding remarks.

2 Homogenized Correlated Nonlinear Filtering

The theoretical aim of filtering is to derive representations and convergence results of the filter π_t^ε ,

$$\pi_t^\varepsilon(\varphi) \equiv \mathbb{E}_{\mathbb{Q}}[\varphi(X_t^\varepsilon, Z_t^\varepsilon) \mid \mathcal{Y}_t^\varepsilon],$$

a conditional measure, where $\mathcal{Y}_t^\varepsilon \equiv \sigma(\{Y_s^\varepsilon \mid s \in [0, t]\})$ is the σ -algebra generated by the observation process, $\varphi(x, z)$ is an integrable test function of interest, and the dynamics of $(X_t^\varepsilon, Z_t^\varepsilon, Y_t^\varepsilon)$ is given by (1). In the case where for each fixed x , $Z_t^{\varepsilon, x}$ is ergodic and converges rapidly to its stationary distribution; that is,

$$dZ_t^{\varepsilon, x} = \frac{1}{\varepsilon} f(x, Z_t^{\varepsilon, x}) dt + \frac{1}{\sqrt{\varepsilon}} g(x, Z_t^{\varepsilon, x}) dV_t,$$

is ergodic, then the theory of stochastic averaging [28] tells us that $X_t^\varepsilon \Rightarrow X_t^0$ in distribution as $\varepsilon \rightarrow 0$, where X_t^0 satisfies the following averaged stochastic differential equation (SDE),

$$dX_t^0 = \bar{b}(X_t^0) dt + \sqrt{\bar{a}(X_t^0)} dW_t. \quad (2)$$

Equation (2) is also known as the effective dynamics. With the stationary distribution of $Z_t^{\varepsilon, x}$ denoted as $\mu_\infty(z; x)$, the averaged coefficients \bar{b} and \bar{a} are defined as,

$$\begin{aligned} \bar{b}(x) &\equiv \int b(x, z) \mu_\infty(dz; x), \\ \bar{a}(x) &\equiv \int \sigma \sigma^*(x, z) \mu_\infty(dz; x). \end{aligned} \quad (3)$$

Then $\sqrt{\bar{a}(x)}$ is the factor of the modified Cholesky decomposition of $\bar{a}(x)$.

2.1 Main Result

In the case $\varphi = \varphi(x)$, we consider the x -marginal of $\pi_t^\varepsilon(\varphi)$,

$$\pi_t^{\varepsilon, x}(\varphi) \equiv \int \varphi(x) \pi_t^\varepsilon(dx, dz)$$

If X_t^ε takes values in \mathbb{R}^m and Z_t^ε in \mathbb{R}^n with $m \leq n$, then it would be advantageous to consider the reduced (homogenized) filter equation

$$\pi_t^0(\varphi) \equiv \mathbb{E}_{\mathbb{Q}}[\varphi(X_t^0) \mid \mathcal{Y}_t^\varepsilon].$$

Yet the result $X_t^\varepsilon \Rightarrow X_t^0$ does not necessarily imply $\pi_t^{\varepsilon, x} \rightarrow \pi_t^0$. In [16], convergence of the x -marginal to the homogenized filter is shown for the non-correlated case, $\alpha = 0$. Specifically, it is proved that for any $T > 0$, the difference between the x -

marginal of the original filter and the filter for the coarse-grain dynamics goes to zero as $\varepsilon \rightarrow 0$ at the rate $\sqrt{\varepsilon}$,

$$\mathbb{E}_{\mathbb{Q}}[d(\pi_T^{\varepsilon,x}, \pi_T^0)] \leq \sqrt{\varepsilon}C, \quad (4)$$

where d denotes a suitable distance on the space of probability measures that generates the topology of weak convergence. Kushner [19] presents the next closest result to what we desire for two-timescale filtering problems, which is covered in great detail there, but does not obtain rates of convergence.

It is of interest to understand if a similar result holds in the correlated sensor noise case. For example, in our motivating problem of atmospheric or climate problems, sensors in those environments (e.g. floats, drifters, balloons) are coupled to their noisy environment. Further, as discussed in [14, 34], the correlated noise problem also occurs whenever a filter is based on a discrete time model that is derived from a continuous time model.

In this section, we provide the main ideas and tools used to show the following result,

Theorem 1. *Under appropriate assumptions on the coefficients of (1), for every $p \geq 1$ and $T \geq 0$ there exists $C > 0$, such that for every φ with sufficient regularity*

$$\left(\mathbb{E}_{\mathbb{Q}} \left[|\pi_T^{\varepsilon,x}(\varphi) - \pi_T^0(\varphi)|^p \right] \right)^{1/p} \leq \sqrt{\varepsilon}C(\varphi)$$

In particular, there exists a metric d on the space of probability measures, such that d generates the topology of weak convergence, and such that for every $T \geq 0$ there exists $C > 0$ such that

$$\mathbb{E}_{\mathbb{Q}}[d(\pi_T^{\varepsilon,x}, \pi_T^0)] \leq \sqrt{\varepsilon}C$$

The convergence result, with an exact rate of convergence, is an extension of [16]. The complete details of the proof are provided in [5].

2.2 Zakai Equation

To show the above result, we make use of probabilistic representations of stochastic partial differential equations, that then allows us to get estimates giving a rate of convergence. To begin, we perform a standard Girsanov change of measure using the exponential martingale D_t^ε ,

$$D_t^\varepsilon \equiv \frac{d\mathbb{P}^\varepsilon}{d\mathbb{Q}} \Big|_{\mathcal{F}_t} = \exp \left(- \int_0^t h^*(X_s^\varepsilon, Z_s^\varepsilon) dB_s - \frac{1}{2} \int_0^t \|h(X_s^\varepsilon, Z_s^\varepsilon)\|^2 ds \right),$$

where $dB_t \equiv \alpha dW_t + \gamma dU_t$. If $\alpha\alpha^* + \gamma\gamma^* = \text{Id.}$, then B_t is a standard BM; this we assume for now. Then by the Kallianpur-Striebel formula, we can express the normalized condition measure π_t^ε in terms of an unnormalized condition measure ρ_t^ε ,

$$\pi_t^\varepsilon(\varphi) = \frac{\mathbb{E}_{\mathbb{P}^\varepsilon} \left[\varphi(X_t^\varepsilon, Z_t^\varepsilon) \tilde{D}_t^\varepsilon \mid Y_t^\varepsilon \right]}{\mathbb{E}_{\mathbb{P}^\varepsilon} \left[\tilde{D}_t^\varepsilon \mid Y_t^\varepsilon \right]} = \frac{\rho_t^\varepsilon(\varphi)}{\rho_t^\varepsilon(1)}, \quad (5)$$

where $\tilde{D}_t^\varepsilon = (D_t^\varepsilon)^{-1}$. The advantage of working with ρ_t^ε is that its evolution is defined by linear dynamics, whereas π_t^ε is nonlinear. Similarly, define $\pi_t^0(\varphi) = \rho_t^0(\varphi)/\rho_t^0(1)$ and the x -marginals,

$$\pi_t^{\varepsilon,x}(\varphi) = \rho_t^{\varepsilon,x}(\varphi)/\rho_t^{\varepsilon,x}(1), \quad \rho_t^{\varepsilon,x}(\varphi) \equiv \int \varphi(x) \rho_t^\varepsilon(dx, dz).$$

Under the new measure \mathbb{P}^ε , the observation process is a standard Brownian motion (BM) and the SDEs for $(X_t^\varepsilon, Z_t^\varepsilon)$ are now of the form,

$$\begin{aligned} dX_t^\varepsilon &= [b(X_t^\varepsilon, Z_t^\varepsilon) - \sigma(X_t^\varepsilon, Z_t^\varepsilon) \alpha^* h(X_t^\varepsilon, Z_t^\varepsilon)] dt + \sigma(X_t^\varepsilon, Z_t^\varepsilon) d\tilde{W}_t \\ dZ_t^\varepsilon &= \frac{1}{\varepsilon} f(X_t^\varepsilon, Z_t^\varepsilon) dt + \frac{1}{\sqrt{\varepsilon}} g(X_t^\varepsilon, Z_t^\varepsilon) dV_t, \end{aligned}$$

where \tilde{W}_t is a standard BM under \mathbb{P}^ε . The unnormalized conditional measure, ρ_t^ε , satisfies a Zakai-type equation,

$$d\rho_t^\varepsilon(\varphi) = \rho_t^\varepsilon(\mathcal{G}^\varepsilon \varphi) dt + \rho_t^\varepsilon(h^* \varphi + \partial_x \varphi \sigma \alpha^*) dY_t^\varepsilon, \quad (6)$$

with generator \mathcal{G}^ε given by

$$\begin{aligned} \mathcal{G}^\varepsilon &\equiv \frac{1}{\varepsilon} \mathcal{G}_F + \mathcal{G}_S \\ \mathcal{G}_F &\equiv \sum_{i=1}^n f_i(x, z) \frac{\partial}{\partial z^i} + \frac{1}{2} \sum_{i,j}^n (g g^*)_{ij}(x, z) \frac{\partial^2}{\partial z^i \partial z^j} \\ \mathcal{G}_S &\equiv \sum_{i=1}^m b_i(x, z) \frac{\partial}{\partial x^i} + \frac{1}{2} \sum_{i,j}^m (\sigma \sigma^*)_{ij}(x, z) \frac{\partial^2}{\partial x^i \partial x^j}. \end{aligned}$$

Note that in the case of correlated noise, an additional stochastic forcing term, $\rho_t^\varepsilon(\partial_x \varphi \sigma \alpha^*) dY_t^\varepsilon$ is present in (6). Also of note, is that the correlated problem typically addressed in the literature is of the form:

$$\begin{aligned} dX_t^\varepsilon &= b(X_t^\varepsilon, Z_t^\varepsilon) dt + \psi(X_t^\varepsilon, Z_t^\varepsilon) dW_t + \xi(X_t^\varepsilon, Z_t^\varepsilon) dU_t \\ dY_t^\varepsilon &= h(X_t^\varepsilon, Z_t^\varepsilon) dt + dU_t \end{aligned} \quad (7)$$

(c.f. [6, 3]). Choosing $\xi = \sigma \alpha^*$ and $\psi \psi^* + \xi \xi^* = \sigma \sigma^*$ again yields the Zakai-type equation for ρ_t^ε , (6). This is useful since convergence proofs for branching particle algorithms, which provide a numerical solution to 6, already exist for this second case [6].

2.3 Dual Representations and Reduced Order Zakai Equation

For a given bounded test function φ and terminal time T , we follow [29] in introducing the associated dual process $v_t^{\varepsilon,T,\varphi}(x,z)$ of (6),

$$v_t^{\varepsilon,T,\varphi}(x,z) \equiv \mathbb{E}_{\mathbb{P}_{t,x,z}^\varepsilon} \left[\varphi(X_T^\varepsilon) \tilde{D}_{t,T}^\varepsilon \mid Y_{t,T}^\varepsilon \right], \quad (8)$$

where

$$\tilde{D}_{t,T}^\varepsilon \equiv \exp \left(\int_t^T h^*(X_s^\varepsilon, Z_s^\varepsilon) dY_s^\varepsilon - \frac{1}{2} \int_t^T \|h(X_s^\varepsilon, Z_s^\varepsilon)\|^2 ds \right).$$

$v_t^{\varepsilon,T,\varphi}(x,z)$ is dual in the sense that $\rho_T^{\varepsilon,x}(\varphi) = \rho_t^\varepsilon(v_t^{\varepsilon,T,\varphi})$ is almost surely constant for each bounded test function φ and time t . Therefore if we also introduce

$$\tilde{D}_{t,T}^0 \equiv \exp \left(\int_t^T \bar{h}^*(X_s^0) dY_s^\varepsilon - \frac{1}{2} \int_t^T \|\bar{h}(X_s^0)\|^2 ds \right),$$

$$v_t^{0,T,\varphi}(x) \equiv \mathbb{E}_{\mathbb{P}_{t,x}^\varepsilon} \left[\varphi(X_T^0) \tilde{D}_{t,T}^0 \mid Y_{t,T}^\varepsilon \right],$$

and

$$\bar{h}(x) \equiv \int h(x,z) \mu_\infty(dz; x), \quad (9)$$

then to show convergence of $\pi_T^{\varepsilon,x} \rightarrow \pi_T^0$ as $\varepsilon \rightarrow 0$, it suffices to show $v_0^{\varepsilon,T,\varphi} \rightarrow v_0^{0,T,\varphi}$ for sufficiently many φ . Specifically, let

$$\bar{v}^\varphi(x,z) \equiv v_0^{\varepsilon,T,\varphi}(x,z) - v_0^{0,T,\varphi}(x), \quad (10)$$

then in [5] we use the following inequality relations,

$$\begin{aligned} \mathbb{E} \left[|\rho_T^{\varepsilon,x}(\varphi) - \rho_T^0(\varphi)|^p \right] &= \mathbb{E} \left[\left| \int \bar{v}^\varphi(x,z) \mathbb{Q}_0(dx, dz) \right|^p \right] \\ &\leq \mathbb{E} \left[\int |\bar{v}^\varphi(x,z)|^p \mathbb{Q}_0(dx, dz) \right] \\ &= \int \mathbb{E} [|\bar{v}^\varphi(x,z)|^p] \mathbb{Q}_0(dx, dz), \end{aligned} \quad (11)$$

where

$$d\rho_t^0(\varphi) = \rho_t^0(\bar{\mathcal{G}}\varphi)dt + \rho_t^0(\bar{h}^*\varphi + \partial_x\varphi\bar{\sigma}\alpha^*)dY_t, \quad (12)$$

with $\rho_0^0(\varphi) = \mathbb{E}_{\mathbb{Q}}[\varphi(X_0^0)]$ and $\bar{\sigma}(x) = \int_{\mathbb{R}^n} \sigma(x, z) \mu_{\infty}(dz; x)$. Proving convergence of $\mathbb{E}[|\bar{v}^{\varphi}(x, z)|^p]$ at the rate $\varepsilon^{p/2}$ in the last line (11) and assuming the initial distribution \mathbb{Q}_0 is well-behaved, alongside supporting arguments, provides the desired result.

2.4 BDSDEs and Sketch of Convergence Proof

To show convergence of $\mathbb{E}[|\bar{v}^{\varphi}(x, z)|^p]$ at the rate $\varepsilon^{p/2}$, we use the following program. We fix $T > 0$ and a test function φ and omit them from the notation for $v_t^{\varepsilon, T, \varphi}$. The process v_t^{ε} solves the backward stochastic partial differential equation (BSPDE)

$$-dv_t^{\varepsilon} = \mathcal{G}^{\varepsilon} v_t^{\varepsilon} dt + (\bar{h}^* v_t^{\varepsilon} + \partial_x v_t^{\varepsilon} \bar{\sigma} \alpha^*) d\bar{Y}_t, \quad v_T^{\varepsilon} = \varphi \quad (13)$$

where \bar{Y}_t denotes the application of a backward Itô integral; Y_t a standard BM with backward filtration.

The main idea is to expand v_t^{ε} as a series expansion in ε ,

$$v_t^{\varepsilon}(x, z) = v_t^0(x) + \psi_t(x, z) + R_t(x, z), \quad (14)$$

ψ_t, R_t begin corrector and error terms respectively, and

$$-dv_t^0 = \bar{\mathcal{G}} v_t^0 dt + (\bar{h}^* v_t^0 + \partial_x v_t^0 \bar{\sigma} \alpha^*) d\bar{Y}_t, \quad v_T^0 = \varphi. \quad (15)$$

We pause at this point to re-iterate the significance of showing convergence to the reduced order filter. Specifically, the dual representation of the unnormalized conditional measure, given in (13), is a function valued process for a given test function $\varphi(x, z)$. The test function has domain $\mathbb{R}^m \times \mathbb{R}^n$ for some $m, n \in \mathbb{N}$. Hence the linear operator $\mathcal{G}^{\varepsilon}$ in (13) is defined by coefficients taking values in $\mathbb{R}^m, \mathbb{R}^n, \mathbb{R}^{m \times m}$ and $\mathbb{R}^{n \times n}$. In comparison, the dual of the reduced order filter, given in (15), has a linear operator $\bar{\mathcal{G}}$ defined only by coefficients taking values in \mathbb{R}^m and $\mathbb{R}^{m \times m}$. Typically, $m \ll n$ in multiscale problems, and hence performing filtering on the reduced order problem is computationally advantageous. Even in the case that m is equal to or only slightly less than $m + n$, it is still desirable to filter on the reduced order equation.

The critical tool to show convergence is to use the theory of backward doubly stochastic differential equations (BDSDEs) [30], so that we have a finite-dimensional representation for each of the terms in the expansion (14). For instance, consider the BSPDE given by (13). Applying the existence results for BSPDEs [33], let us try to find the dynamics of $\theta_s^{t,x,z} = v_s^{\varepsilon}(X_s^{\varepsilon,t,x}, Z_s^{\varepsilon,t,z})$, where $X_s^{\varepsilon,t,x}$ and $Z_s^{\varepsilon,t,z}$ are versions of X_s^{ε} and Z_s^{ε} starting at x and z respectively at time t . This implies that $\theta_t^{t,x,z} = v_t^{\varepsilon}(x, z)$ and according to [30], the dynamics of $\theta_s^{t,x,z}$ are given by a BDSDE:

$$-d\theta_s^{t,x,z} = (\bar{h}^*(X_s^{\varepsilon,t,x}, Z_s^{\varepsilon,t,z}) \theta_s^{t,x,z} + \eta_s^{t,x,z} \alpha^*) d\bar{Y}_s - \eta_s^{t,x,z} dW_s - \xi_s^{t,x,z} dV_s, \quad (16)$$

with boundary condition $\theta_T^{t,x,z} = \varphi(X_T^{\varepsilon,x,t}, Z_T^{\varepsilon,z,t})$, driven by the signal process with generator \mathcal{G}^ε ,

$$\begin{aligned} dX_s^\varepsilon &= b(X_s^\varepsilon, Z_s^\varepsilon)ds + \sigma(X_s^\varepsilon, Z_s^\varepsilon)dW_s & X_t^\varepsilon &= x \\ dZ_s^\varepsilon &= \frac{1}{\varepsilon}f(X_s^\varepsilon, Z_s^\varepsilon)ds + \frac{1}{\sqrt{\varepsilon}}g(X_s^\varepsilon, Z_s^\varepsilon)dV_s & Z_t^\varepsilon &= z, \end{aligned}$$

and $(\eta_s^{t,x,z}, \xi_s^{t,x,z})$ are pair process to $\theta_s^{t,x,z}$ such that

$$\begin{aligned} \eta_t^{t,x,z} &= \partial_x v_t^\varepsilon(x, z)\sigma(x, z) \\ \xi_t^{t,x,z} &= \frac{1}{\sqrt{\varepsilon}}\partial_z v_t^\varepsilon(x, z)g(x, z). \end{aligned}$$

The BDSDE (16) is a finite dimensional representation of (13) and therefore allows us to use standard tools, for example Gronwall's lemma, for calculating the necessary estimates.

To summarize this section, we use the existence results for BSPDE [33] for the terms in the expansion (14), producing a BDSDE representation [30] for each. Then using the ergodic property of the fast-process, and estimates on the semi-groups for the drift and diffusion coefficients [36, 31], we show in [5] that $\mathbb{E}[|\psi_t|^p], \mathbb{E}[|R_t|^p] \rightarrow 0$ as $\varepsilon \rightarrow 0$ at the rate of $\varepsilon^{p/2}$.

3 The Lorenz 1996 Model

The Lorenz '96 model was originally introduced in [23] to mimic multiscale mid-latitude atmospheric dynamics for an unspecified scalar meteorological quantity. A latitude circle is divided into K sectors, and each sector is subdivided into J sub-sectors. The model has two timescales with the slow-scale atmospheric variable at time t in the k -th sector given by $\{X_t^k\}$ and the fast-scale atmospheric variable in subsector (k, j) at time t given by $\{Z_t^{k,j}\}$. Mathematically, the dynamics are

$$\begin{aligned} dX_t^k &= (X_t^{k-1}(X_t^{k+1} - X_t^{k-2}) - X_t^k + F + \frac{h_x}{J} \sum_{j=1}^J Z_t^{k,j})dt \\ dZ_t^{k,j} &= \frac{1}{\varepsilon} \left(Z_t^{k,j+1}(Z_t^{k,j-1} - Z_t^{k,j+2}) - Z_t^{k,j} + h_z X_t^k \right) dt, \end{aligned} \quad (17)$$

where $k = 1, \dots, K$ and $j = 1, \dots, J$. Here, we use the version of the model in [10] and [17], in which the nonlinear, linear and slow scale effects in the fast dynamics are all of order 1. In this setting, [10] showed that (for a lower order version of the Lorenz '96 model) the fast scale dynamics display ergodic properties such that the averaging technique described in Section 3.1 can be used to average out the fast dynamics when we are only interested in the slow dynamics (coarse-grain process). In (17), F is a slow-scale forcing, and h_x, h_z are coupling terms.

The dynamics of unresolved modes can be represented by adding forcing in the form of stochastic terms (see, for example, [24, 25]). The use of stochastic terms to represent nonlinear self-interaction effects at short timescales in the unresolved modes is appropriate if we are only interested in the coarse-grain dynamics occurring in the long, slow timescale. This is called stochastic consistency in [25]. Considering (17), where only quadratic nonlinearity is present in the fast process, the motivation behind adding stochastic forcing is thus to model higher order self-interaction effects.

For the purpose of filtering, which requires some noise in the dynamics, as well as for the reason of stochastic consistency, let us write (17) in a standard form with additive stochastic forcing,

$$\begin{aligned} dX_t^\varepsilon &= b(X_t^\varepsilon, Z_t^\varepsilon)dt + \sigma_x dW_t, & X_t^\varepsilon &\in \mathbb{R}^K \\ dZ_t^\varepsilon &= \frac{1}{\varepsilon} f(X_t^\varepsilon, Z_t^\varepsilon)dt + \frac{1}{\sqrt{\varepsilon}} \sigma_z dV_t, & Z_t^\varepsilon &\in \mathbb{R}^J, \end{aligned} \quad (18)$$

with b, f given as in (17). In Section 6, we will consider the application of numerical filters to solve the filtering problem with the Lorenz '96 model as the signal of interest. We will in particular be interested in the estimation of the coarse-grain $\{X_t^k\}$ dynamics by way of using the homogenized dynamics $\{X_t^{0,k}\}$.

We now fix our model parameters, for the purpose of understanding the behavior of (17) and making comparisons with the numerical solution of the homogenized dynamics. Let the simulation parameters be: $\varepsilon = 1\text{E-}2$, $F = 10$, $(h_x, h_z) = (-1, 1)$, σ_x, σ_z sparse square matrices with 1 along the diagonal and 0.05 on the first two sub and super-diagonals, $K = 6$ and $J = 9$. Hence $(X_t, Z_t) \in \mathbb{R}^6 \times \mathbb{R}^{54}$ and the homogenized dynamics have $X_t^0 \in \mathbb{R}^6$, so a state space dimension a 10th of the original. Figure 1 illustrates the behavior of a generic slow state X_t^1 (shown in orange), the fast states in the 1st sector, that is $Z_t^{1,1}, \dots, Z_t^{1,9}$ (shown in gray), and the fast scale forcing that enters (17) for the X_t^1 component (shown in light blue); the fast scale forcing is $(h_x/J) \sum_{j=1}^J Z_t^{1,j}$.

Due to the symmetry of the model, it is sufficient to look at one sector to get a glimpse of the qualitative behavior of the dynamics. According to [23], the time scale used here, $T = 20$, is roughly equivalent to mimicking 100 days in 'real' time. The solution shown in Fig. 1 was produced by integrating the initial conditions with an Euler-Maruyama integration scheme with a step-size of $\delta = 1\text{E-}4$.

3.1 Heterogenous Multiscale Method (HMM)

Since we will be interested in filtering on the homogenized dynamics, given in general form by (2), we require a numerical method for generating \bar{b} and $\bar{a}^{1/2}$ for the signal dynamics, \bar{h} for the observation process, and depending on the numerical method chosen for filtering, $\bar{\sigma}$. The heterogenous multiscale method (HMM) [35, 10] outlines an algorithm for efficient multiscale integration with [9] providing

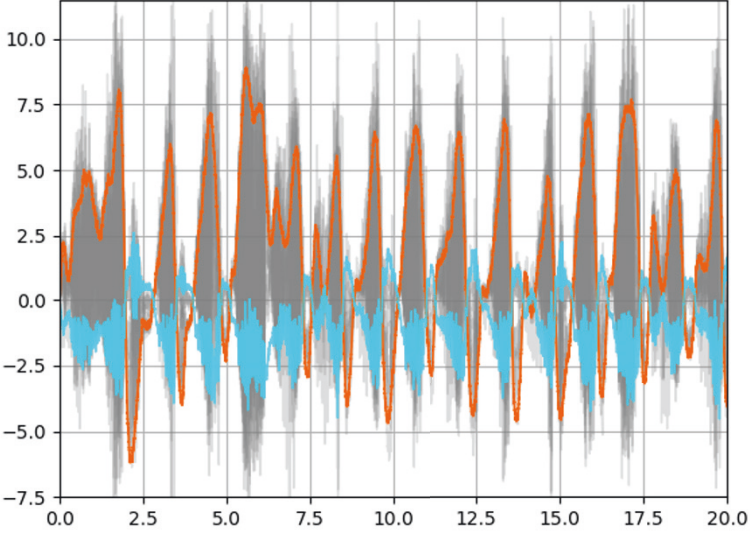


Fig. 1 Behavior of Lorenz '96: X_t^1 in orange, $Z_t^{1,1}, \dots, Z_t^{1,9}$ in gray, and the fast scale forcing $(h_x/J) \sum_{j=1}^J Z_t^{1,j}$ on X_t^1 in light blue.

the analysis proving relevant weak and strong convergence theorems for the algorithm. The driving numerical motivation for HMM, is that the stability of simulating the signal process $(X_t^\varepsilon, Z_t^\varepsilon)$ of (1) requires an integration step-size, δ_m , which must be smaller than ε . The general idea of HMM is to integrate $(x, Z_t^{\varepsilon, (z_0, s)})$, that is the fast process Z_t^ε starting at z_0 at time $s < t$, with fixed $X_t^\varepsilon = x$, with an integration step-size $\delta_m < \varepsilon$ for a small period of time $\Delta_m > \delta_m$ and with a finite number of realizations $\mathcal{R} \in \mathbb{N}$. We call Δ_m the fast-macro step-size. From this simulation, a transition density $\mu_{\Delta_m}(z; x, z_0)$ is constructed and should be close to $\mu_\infty(z; x)$. Then the averaged coefficients of (3) and (9) can be approximated, so that filtering can be applied to (12). Since $\mu_\infty(z; x)$ is dependent on x , the averaged coefficients must be recalculated, but on larger time-scales than Δ_m . We denote $\Delta_M \geq \Delta_m$, the slow-macro step-size, which is the interval of time upon which the transition density $\mu_{\Delta_m}(z; x, z_0)$ and hence coefficients $\bar{b}, \bar{a}, \bar{\sigma}, \bar{h}$ are assumed to hold accurately. During this time-interval a slow-integration step-size $\delta_M > \delta_m$ is used to integrate the averaged SDE (2).

The parameters we use in this paper are slightly different than that in [10]. Therefore we should confirm the applicability of HMM to our problem. We do this numerically, setting the relevant simulation parameters to $\mathcal{R} = 1$, $\delta_m = 1\text{E-}4$, $\delta_M = 1\text{E-}2$, $\Delta_m = 5\delta_m$, and $\Delta_M = 10\delta_M$. With these parameters, and the same initial conditions as used in Fig. 1, we get the result shown in Fig. 2 when integrating with HMM us-

ing Euler-Maruyama integration schemes for both the fast and slow-scale processes. After some time, the qualitative behavior seen in Fig. 1 and Fig. 2 are quite different, but Fig. 3 and Fig. 4 show us that on shorter time scales (the first 0.8 time units of the simulations), the dynamics of the numerically averaged X_t^0 are indeed near the original X_t .

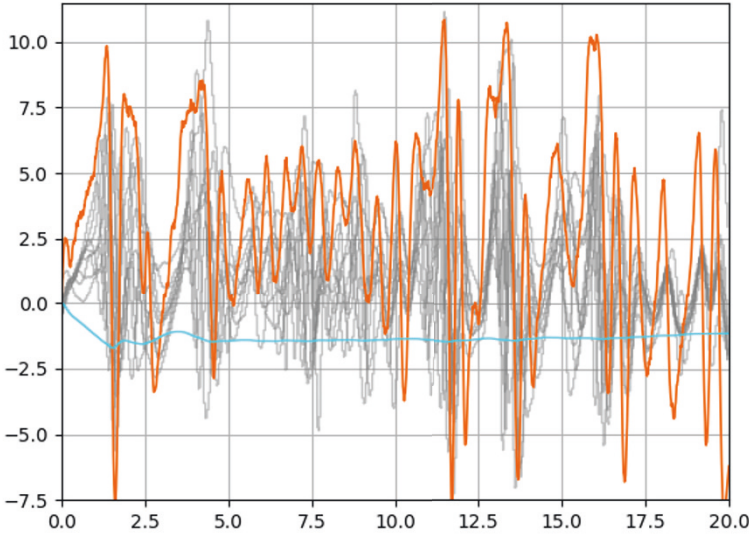


Fig. 2 Behavior of HMM solution of Lorenz '96: $X_t^{0,1}$ in orange, $\mathcal{R} = 1$ realizations of $Z_t^{1,1} \dots, Z_t^{1,9}$ with fixed X_t^0 in gray, and the averaged fast scale forcing on $X_t^{0,1}$ in light blue.

The fact that the numerically averaged is close on shorter time scales is sufficient in the filtering context of sparse in-time observations if the observations occur before the X_t and X_t^0 solutions separate too much. For instance, in Section 6 we will assume that the observations come every $\Delta t = 10\delta_M$, which is $\Delta t = 0.1$ for our parameters. The solutions X_t and X_t^0 are certainly visually close in Fig. 3 and Fig. 4 over the time interval $[0, 0.1]$. The update step at $t = 0.1$ in filtering will then improve the estimate of X_t at time $t = 0.1$ and therefore limit the separation that may have occurred over the interval $[0, 0.1]$.

A more rigorous numerical verification that HMM is appropriate for our model comes from two investigations: 1. comparing the effective (stationary) density associated with X_t^0 with the x -marginal of the transition density of X_t^ε with our time scale separation parameter $\varepsilon = 1E-2$, and 2. showing that the (x, Z_t^ε) process converges exponentially to its invariant distribution, regardless of initial condition on $Z_t^\varepsilon = z \in \mathbb{R}^n$. Technically, we should see this last result occur within a time interval

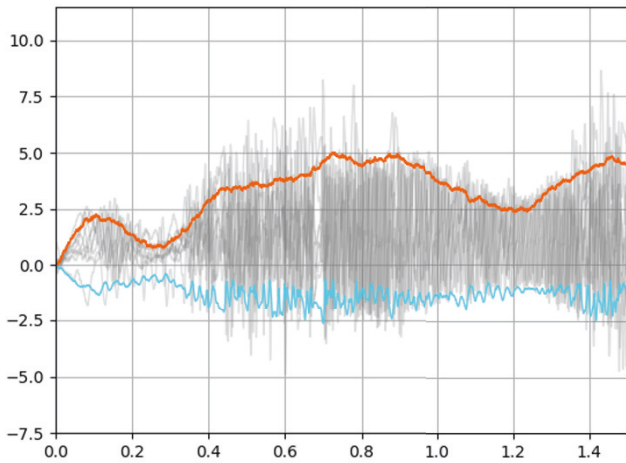


Fig. 3 Behavior of Lorenz '96: X_t^1 in orange, $Z_t^{1,1}, \dots, Z_t^{1,9}$ in gray, and the fast scale forcing $(h_x/J) \sum_{j=1}^J Z_t^{1,j}$ on X_t^1 in light blue.

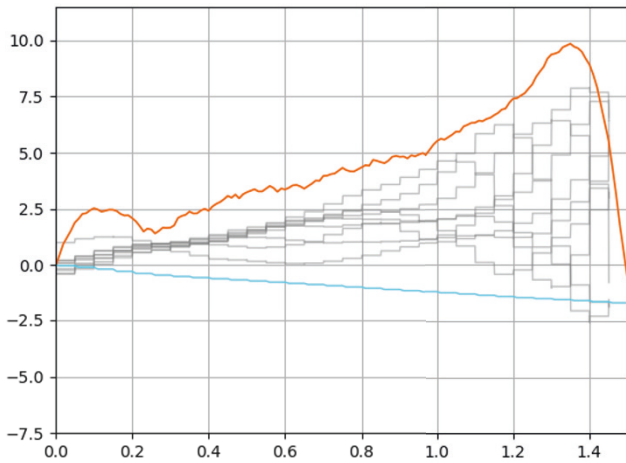


Fig. 4 Behavior of HMM solution of Lorenz '96: $X_t^{0,1}$ in orange, $\mathcal{R} = 1$ realizations of $Z_t^{1,1}, \dots, Z_t^{1,9}$ with fixed X_t^0 in gray, and the averaged fast scale forcing on $X_t^{0,1}$ in light blue.

of length Δ_m , which implies that the application of HMM is well founded for our

choice of parameter. Again, in the context of filtering, we are able to relax this last requirement and still effectively filter.

For the first investigation, Fig. 5 provides a numerical confirmation that (17) with $\varepsilon = 1\text{E-}2$ produces X_t^ε with a marginal density close to the effective density for X_t^0 . Specifically, Fig. 5 compares the marginal density of the first slow component X_t^1 for $\varepsilon = 1\text{E-}2$ and $1\text{E-}3$, which are nearly the same, implying that the statistics for the dynamics of X_t^ε with $\varepsilon = 1\text{E-}2$ is close to X_t^0 . Because of the symmetry of the signal model (17), all slow components X_t^k have the same marginal density, hence comparing only for the X_t^1 marginal density is appropriate.

In Fig. 6 we show the marginal density of the first component of the fast process for four different simulations. For this analysis, we simulate the full model from a randomly generated initial condition to eliminate transient effects. Then we fix the slow process $X_t^\varepsilon = x$ and simulate the fast process for randomly generated Z_0^ε , where each component of Z_0^ε is chosen according to $\mathcal{N}(0, 1)$; normal distribution with mean zero and variance one. Fig. 6 shows the convergence of the transition densities $\mu_{15\Delta_m}(z; X_0^\varepsilon = x, Z_0^\varepsilon)$ for the first component of Z_t^ε ; showing that on a macro step of $15\Delta_m$ we have sufficient convergence from most initial states of Z_0^ε . When using HMM in our estimation implementation of HHPF, we can in fact relax the condition for convergence of the transition densities and still effectively filter. Hence why we will use a macro step of only Δ_m in Section 6 analysis.

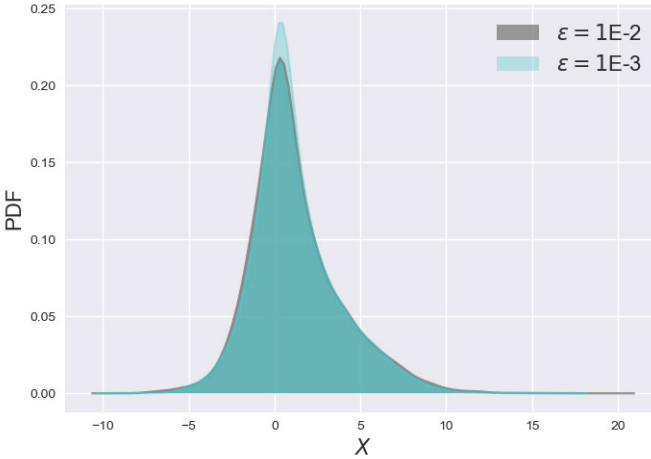


Fig. 5 Simulation of (18). Shown in gray is the X_t^1 (i.e. first component) marginal density when $\varepsilon = 1\text{E-}2$ and similarly in light blue the X_t^1 marginal density when $\varepsilon = 1\text{E-}3$.

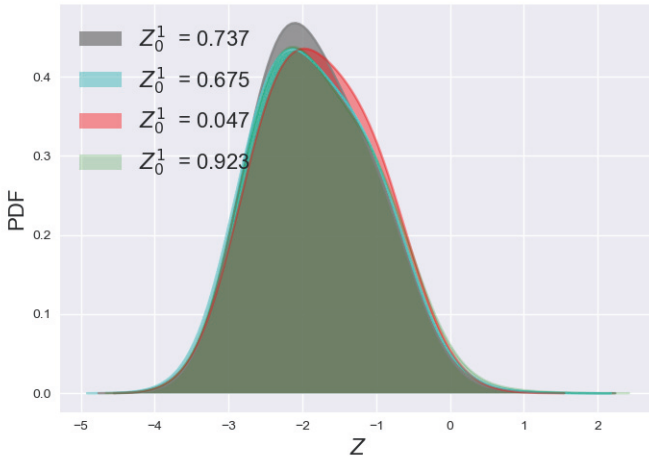


Fig. 6 Transition densities of the first component of Z_t^ϵ : $\mu_{15, \Delta_m}(z; X_0^\epsilon = x, Z_0^\epsilon)$, for randomly generated Z_0^ϵ ; the first component of Z_0^ϵ is shown as Z_0^1 in the legend for four different simulations. $X_0^\epsilon = x$ is a fixed slow state.

4 Numerical Solution of the Reduced Order Filter

The convergence result shown in Section 2 provides the theoretical foundation upon which efficient numerical methods can be developed to solve the multiscale correlated filtering problem. Our objective in this section, is to explain how this theory can be used to create sequential Monte Carlo methods for the homogenized filtering equation. In particular, we will show how a standard sequential importance sampling (SIS) particle filter (PF) can be combined with the HMM to create the Homogenized Hybrid Particle Filter (HHPF); an SIS PF adapted to the homogenized dynamics. We will consider the case of a continuous-time signal and discrete-time observation processes. A similar methodology can be pursued for other ensemble based filtering methods, for instance the ensemble Kalman filter [17, 38].

4.1 Sequential Importance Sampling Particle Filter (PF)

Standard particle filtering procedures (particle approximation of distributions, propagation and weighting) can be used in approximating the distribution of the signal process conditioned on observations. Ultimately, we are interested in approximating integrals of certain test functions φ ,

$$\pi_t^\varepsilon(\varphi) = \int_{\mathbb{R}^m \times \mathbb{R}^n} \varphi(x, z) \pi_t^\varepsilon(dx, dz).$$

Hence we would like to be able to sample from π_t^ε . Although we cannot do this directly, we can use the principle of importance sampling to approximate this action.

The main idea in importance sampling is that π is a distribution that is difficult to sample from, yet it is proportional to a function that is easy to evaluate: $\pi \propto \psi$. One also assumes that there exists a distribution q , called the *importance* or *proposal distribution*, that is absolutely continuous with respect to π ; that is $q \ll \pi$, and which is easy to sample from. Then define the *importance weights* $w \equiv \psi/q$, so that

$$\pi \propto wq. \quad (19)$$

Returning to our filtering problem in the discrete-time observation setting, let $\{t_k \in \mathbb{R}\}$, $k \in \mathbb{N}$ be a finite collection of observation times, that are strictly increasing. And for brevity, we will simply use k as a time index for processes instead of the full t_k . Let $\xi \in \mathbb{R}^m \times \mathbb{R}^n$ be a state in the range space of the signal process $(X_t^\varepsilon, Z_t^\varepsilon)$ and $y \in \mathbb{R}^d$ for Y_t^ε in (1). Then by Bayes' theorem, the posterior distribution $p(\xi_k | y_k)$ is proportional to multiplication of the likelihood and prior distributions,

$$p(\xi_k | y_k) \propto p(y_k | \xi_k) p(\xi_k) \quad (20)$$

Therefore applying the principle of importance sampling in this setting, we have

$$p(\xi_k | y_k) \propto w_k q(\xi_k) \quad \text{where} \quad w_k = \frac{p(y_k | \xi_k) p(\xi_k)}{q(\xi_k)}.$$

Now assume we can approximate the posterior distribution as a weighted collection of Dirac distributions. In particular, consider an ensemble of independent particles indexed by a set $\mathcal{A} = \{1, \dots, N\}$, $N \in \mathbb{N}$, with the particles evolving according to the signal process in (1). Each particle represents a stochastic realization of the signal process; we denote the set of values that the particles take in the signal state space at time t_k as \mathcal{A}_k^ξ and the values by individual particles with similarly notation $\mathcal{A}_k^\xi(j)$ for $j \in \mathcal{A}$. The probability of each particle representing the true signal process is given by the set of time-varying weights $\{w_k^j\}_{j \in \mathcal{A}}$. Then the posterior distribution is approximated at time t_k by a weighted sum of Dirac distributions,

$$p(\xi_k | y_k) = \sum_{j \in \mathcal{A}} w_k^j \delta_k^j(\xi_k),$$

where δ_k^j has support on the singleton given by $\mathcal{A}_k^\xi(j)$. And $w_k^j \in [0, 1]$ with $\sum_{j \in \mathcal{A}} w_k^j = 1$ for each t_k .

For convenience, let us make a common choice for the importance distribution, by setting $q(\xi_k)$ equal to the prior distribution. Then given a posterior distribution $p(\xi_k | y_k)$ at time t_k , the importance (prior) distribution is simply

$$q(\xi_{k+1}) = p(\xi_{k+1}) = \sum_{j \in \mathcal{A}} w_k^j \delta_{k+1}^j(\xi_{k+1}). \quad (21)$$

And the posterior distribution at t_{k+1} is,

$$\begin{aligned} p(\xi_{k+1}|y_{k+1}) &= \sum_{j \in \mathcal{A}} w_{k+1}^j \delta_{k+1}^j(\xi_{k+1}) \\ &\propto \sum_{j \in \mathcal{A}} w_k^j p(y_{k+1}|\xi_{k+1}) \delta_{k+1}^j(\xi_{k+1}). \end{aligned} \quad (22)$$

Therefore, when new observation data is available, the weights are updated according to

$$w_{k+1}^j \propto w_k^j p(y_{k+1}|\delta_{k+1}^j). \quad (23)$$

Since $\sum_{j \in \mathcal{A}} w_k^j = 1$ for each t_k , these new weights must be normalized. Lastly, note that in the case where our observation is a Gaussian process; that is

$$Y_k = h(\xi_k) + U_k \quad \text{with} \quad U_k \sim \mathcal{N}(0, R),$$

then the weights are updated according to,

$$w_{k+1}^j \propto w_k^j \exp\left(-\frac{1}{2}(y_{k+1} - h(\xi_{k+1}))^* R^{-1}(y_{k+1} - h(\xi_{k+1}))\right). \quad (24)$$

Particle filters are known to suffer from certain degeneracy conditions. The main issue is that it is not uncommon to have one particle with nearly all the weight after a small number of observations; that is for one $j \in \mathcal{A}$, $w_k^j \simeq 1$ and $w_k^i \ll 1$ for $\mathcal{A} \ni i \neq j$. The a priori selection of an optimal proposal distribution is helpful as a remedy to this problem, but often difficult. Another technique that can be effective in combating degeneracy is *resampling*; intuitively, this just means that particles with large weights are multiplied and those with small weights are eliminated. We refer the reader to [13, 8, 7, 2] for more details regarding resampling, importance sampling and other concepts associated with basic particle filters. In the numerical simulations presented in this paper, we will use the universal (systematic) resampling technique (c.f. [2, p.180]). The resampling technique is used when the *effective particle number* $N_{\text{eff},k} \equiv 1/\sum_{j \in \mathcal{A}_k} (w_k^j)^2$ falls below some user specified threshold; and occurs after updating the weights, but before normalization.

In summary, our standard particle filter algorithm, that will be used in our numerical simulations has the following recursive structure:

Particle Filter (PF) Algorithm

1. At time t_k , set $w_k^j = 1/N, \forall j \in \mathcal{A}$ and

$$p(\xi_k|y_k) = \sum_{j \in \mathcal{A}} w_k^j \delta_k^j(\xi_k).$$

2. Generate the prior at t_{k+1} by advecting each particle under the signal dynamics given by (1),

$$p(\xi_{k+1}) = \sum_{j \in \mathcal{A}} w_k^j \delta_{k+1}^j(\xi_{k+1}).$$

3. Update the particle weights according to

$$w_{k+1}^j \propto w_k^j p(y_{k+1} | \xi_{k+1}).$$

- 4a. If $N_{\text{eff},k}$ is below a threshold (indicating particle degeneracy), then apply universal resampling and set

$$w_{k+1}^j = 1/N, \quad \forall j \in \mathcal{A}.$$

- 4b. Otherwise, compute the l_2 norm of the weights and re-normalize each

$$w_{k+1}^j \leftarrow w_{k+1}^j / |w|_2.$$

4.2 Homogenized Hybrid Particle Filter (HHPF)

HHPF differs from regular particle filtering in the sense that particles are used to represent X_t^0 instead of $(X_t^\varepsilon, Z_t^\varepsilon)$. Hence the particles and their weights approximate the reduced order filter π_t^0 . The numerical integration of the particles $\mathcal{A}_t^x(j)$ under the SDE (2) requires multiscale integration techniques to approximate the averaged drift and diffusion coefficients. Similarly, multiscale techniques are needed for the averaged observation coefficient (9); for observation and updating of the particle weights $\{w_t^j\}_{j \in \mathcal{A}}$. These coefficients are calculated using the HMM described in Section 3.1. We simply summarize the algorithm steps here, and refer the reader to the papers [32, 39, 22] for full details on HHPF.

Homogenized Hybrid Particle Filter (HHPF) Algorithm

1. Same as (PF) step 1.
 2. Apply the HMM multiscale integration technique and compute the averaged coefficients $\bar{b}, \bar{a}^{1/2}, \bar{h}$.
 3. Generate the prior at t_{k+1} by advecting each particle under the signal dynamics given by (2).
 4. Same as (PF) step 3., but using \bar{h} in the likelihood distribution.
 5. Same as (PF) steps 4a. and 4b.
-

4.3 Likelihood for Correlated Sparse Observations

Neither the PF nor HHPF algorithms just described detail how we should account for correlation between the sensor and signal noise in the discrete-time observation

process; that is an observation process of the form,

$$Y_{t_k} = h(X_{t_k}) + \int_{t_{k-1}}^{t_k} \alpha dW_s + \gamma U_{t_k},$$

with $\alpha, \gamma \neq 0$. Following the SIS PF algorithm, when we select the proposal distribution as the prior distribution, then we must derive the likelihood distribution for updating the particle weights. To do this, consider a discrete-time signal and observation process of the form,

$$\begin{aligned} x_j &= f_j(x_{j-1}) + G_j v_{j-1}, \\ &\vdots \\ x_{k-1} &= f_{k-1}(x_{k-2}) + G_{k-1} v_{k-2}, \\ x_k &= f_k(x_{k-1}) + G_k v_{k-1}, \\ y_k &= h_k(x_k) + e_k. \end{aligned} \tag{25}$$

As before, subscript indices indicate times, x_k is the signal process, y_k the observation process, and $\{v_j\}$ is a sequence of independent Gaussian random variables. The sequence $\{e_j\}$ are also Gaussian, but correlated with $\{v_j\}$; specifically, the random variable e_k is correlated with v_{j-1}, \dots, v_{k-1} . Figure 7 provides a pictorial representation of (25).

The noises $v_{j-1}, \dots, v_{k-2}, v_{k-1}, e_k$ are jointly Gaussian,

$$\begin{pmatrix} v_{j-1} \\ \vdots \\ v_{k-2} \\ v_{k-1} \\ e_k \end{pmatrix} \in \mathcal{N} \left(0, \begin{bmatrix} Q_{j-1} & \dots & 0 & 0 & S_j \\ 0 & \ddots & 0 & 0 & \vdots \\ 0 & 0 & Q_{k-2} & 0 & S_{k-1} \\ 0 & 0 & 0 & Q_{k-1} & S_k \\ S_j^T & \dots & S_{k-1}^T & S_k^T & R_k \end{bmatrix} \right),$$

with Q_j the covariance matrix associated with v_j , R_k with e_k and S_j the covariance of v_{j-1} and e_k for instance.

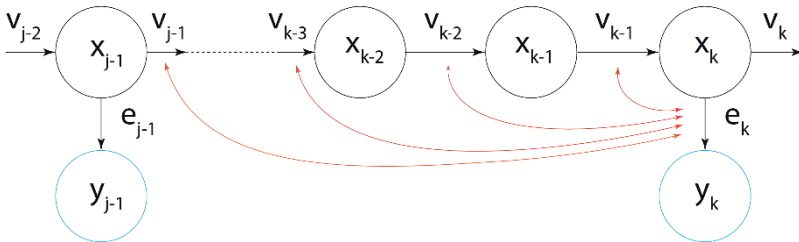


Fig. 7 A pictorial representation of (25) with arrows between v_j and e_k indicating sensor-signal correlation.

For simplicity of discussion, let us assume $f = f_j$ and $h = h_j, \forall j$. Also, define $\widehat{G}_j \equiv G_j Q_{j-1} G_j^T, \widehat{S}_j \equiv G_j S_j$, and

$$\mathfrak{R} \equiv \begin{bmatrix} \widehat{G}_j & \dots & 0 & 0 & \widehat{S}_j \\ 0 & \ddots & 0 & 0 & \vdots \\ 0 & 0 & \widehat{G}_{k-1} & 0 & \widehat{S}_{k-1} \\ 0 & 0 & 0 & \widehat{G}_k & \widehat{S}_k \\ \widehat{S}_j^T & \dots & \widehat{S}_{k-1}^T & \widehat{S}_k^T & R_k \end{bmatrix} = \begin{bmatrix} \widetilde{Q} & \widetilde{S} \\ \widetilde{S}^T & R_k \end{bmatrix}.$$

The probabilistic description of the state space model is then given by,

$$\begin{aligned} p \left(\begin{pmatrix} x_j \\ x_{j+1} \\ \vdots \\ x_k \\ y_k \end{pmatrix} \middle| x_{j-1} \right) &= \mathcal{N} \left(\begin{pmatrix} f(x_{j-1}) \\ \mathbb{E}[f(x_j)] \\ \vdots \\ \mathbb{E}[f(x_{k-1})] \\ \mathbb{E}[h(x_k)] \end{pmatrix}, \mathfrak{R} \right) \\ &= \mathcal{N} \left(\begin{pmatrix} \mathbb{E}[f(X_{j-1:k})] \\ \mathbb{E}[h(x_k)] \end{pmatrix}, \begin{bmatrix} \widetilde{Q} & \widetilde{S} \\ \widetilde{S}^T & R_k \end{bmatrix} \right), \end{aligned}$$

where we used the notation $X_{j-1:k}$ to be the vector (x_k, \dots, x_{j-1}) . We will need the following lemma on conditional Gaussian distributions.

Lemma 1. *Let X, Y be two vectors with jointly Gaussian distribution:*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{bmatrix} P_{xx} & P_{xy} \\ P_{yx} & P_{yy} \end{bmatrix} \right)$$

Then the conditional Gaussian distribution for Y given $X = x$ is Gaussian distributed,

$$(Y|X = x) \sim \mathcal{N}(\mu_y + P_{yx} P_{xx}^{-1}(x - \mu_x), P_{yy} - P_{yx} P_{xx}^{-1} P_{xy}).$$

Using Lemma 1, the likelihood $p(y_k | x_k, x_{k-1}, \dots, x_{j-1})$ is,

$$p(y_k | x_k, x_{k-1}, \dots, x_{j-1}) \propto \mathcal{N}(h(x_k) + \widetilde{S}^T \widetilde{Q}^{-1}(X_{j-1:k} - f(X_{j-1:k})), R_k - \widetilde{S}^T \widetilde{Q}^{-1} \widetilde{S}).$$

Although we presented the results of Section 2 for a continuous-time signal, once we apply a numerical filtering algorithm to the problem, we are forced to numerically integrate and therefore the continuous-time signal becomes a discrete-time process. For instance, our application of an Euler-Maruyama scheme means that each Euler step can be thought of as one line from (25). It is in this sense that we will apply the results of this subsection with PF and HHPF to solve the correlated filtering problem. For a simple two-dimensional example demonstrating correlated

filtering with the aforementioned methods, see [4]. For the reader interested in algorithms for the continuous-time signal and observation case with correlated sensor noise, see [6], where a branching particle filter is presented for solution of the Zakai equation.

5 Nudged Particle Filter

In the previous section, we presented the PF, HHPF, and correlated variants of each derived from sequential importance sampling when the proposal distribution was selected to be the prior distribution. The selection of the prior as the proposal is typical; the optimal proposal distribution, the one that minimizes the variance of the weights of the particles after observations, is usually not known in closed form, and selection of the prior as the proposal provides a closed form for the update of the weights when the observation process is Gaussian. But the typical sub-optimality of selecting the prior as the proposal further encourages particle degeneracy. Particle degeneracy can be particularly pronounced in the case where the signal dynamics are chaotic; possessing positive Lyapunov exponents, for example the Lorenz 1996 model. In this section, we use optimal control methods to construct an improved prior proposal distribution that helps to improve the filtering quality with fixed number of particles when the system is of high dimension and may be chaotic.

5.1 Nudging in Particle Filters

We will construct a better prior proposal, and hence posterior, by incorporating information from the observation into the particle propagation. For example, let us consider the time interval $[t_k, t_{k+1}]$, where observations are given at the discrete times t_k and t_{k+1} . We introduce an additive control, u_t , in the dynamics of each particle $\widehat{X}_t^i \in \mathcal{A}_t^X(i)$,

$$d\widehat{X}_t^i = b(\widehat{X}_t^i)dt + u_t^i dt + \sigma(\widehat{X}_t^i)dW_t, \quad t \in (t_k, t_{k+1}). \quad (26)$$

The control is chosen to minimize the cost functional:

$$J(t_k, x, u) \equiv \mathbb{E}_{t_k, \widehat{X}_{t_k}^{i,x}} \left[\frac{1}{2} \int_{t_k}^{t_{k+1}} u^i(s)^* Q(\widehat{X}_s^i)^{-1} u^i(s) ds + g \left(Y_{t_{k+1}}, \widehat{X}_{t_{k+1}}^i \right) \right], \quad (27)$$

where $\mathbb{E}_{t_k, \widehat{X}_{t_k}^{i,x}}$ is the expectation with respect to the probability measure of the process that starts at $\widehat{X}_{t_k}^{i,x} = x$ at time t_k , $Q(x) = \sigma\sigma^*(x)$ and

$$g(y, x) \equiv \frac{1}{2} (y - h(x))^* R^{-1} (y - h(x)).$$

The choice of cost functional, is such that the control that minimizes the cost functional will have the effect of steering the particle towards a location most representative of the observation $Y_{t_{k+1}}$. For brevity, in the remainder of the paper we suppress the x -dependence in the notation for Q .

Covariance matrices Q and R in the cost indicate that the subspaces of the signal and observation that have larger noise variance contribute less to the total cost. The matrix Q^{-1} allows for more control in the directions of large signal noise by penalizing the energy of the control less in those directions. The terminal cost g incurs a large cost component when $|Y_{t_{k+1}}^j - h_j(\hat{X}_{t_{k+1}}^i)|$ is large, but R^{-1} reduces the contribution of $|Y_{t_{k+1}}^j - h_j(\hat{X}_{t_{k+1}}^i)|$ to the total cost if quality of observation in direction j is poor, so that particles are controlled less based on information from $Y_{t_{k+1}}^j$.

5.2 Optimal Control Solution

We follow the standard procedure [12] and let $V(t, x)$ be the value function defined by

$$V(t, x) \equiv \inf_u J(t_k, x, u), \quad t \in [t_k, t_{k+1}].$$

Then $V(t, x)$ is the solution of the Hamilton-Jacobi-Bellman (HJB) equation,

$$-\frac{\partial V}{\partial t} + H(t, x, D_x V, D_x^2 V) = 0, \quad V(t_{k+1}, x) = g(Y_{t_{k+1}}, x), \quad (28)$$

where the Hamiltonian of the associated control problem is

$$H(t, x, p, P) \equiv \sup_u \left[-(b(x) + u)^* p - \frac{1}{2} u^* Q^{-1} u - \frac{1}{2} \text{tr}(QP) \right] \quad (29)$$

$$= \left[-b(x)^* p + \frac{1}{2} p^* Q p - \frac{1}{2} \text{tr}(QP) \right], \quad (30)$$

where the supremum in the above equation is achieved with $u = -Qp$. Hence the optimal control is

$$u(t) = -Q \nabla_x V(t, \hat{X}_t), \quad (31)$$

V being the solution of (28).

Using the form of the optimal control (31) in the Hamiltonian, (28) can be written as

$$\frac{\partial V}{\partial t} + b(x)^* \nabla_x V + \frac{1}{2} \text{tr}(Q \nabla_x^2 V) - \frac{1}{2} \nabla_x^* V Q \nabla_x V = 0, \quad t \in [t_k, t_{k+1}], \quad (32)$$

$$V(t_{k+1}, x) = g(Y_{t_{k+1}}, x). \quad (33)$$

Equation (32) is nonlinear due to the $\frac{1}{2} \nabla_x^* V Q \nabla_x V$ term. The nonlinearity can be removed by employing a log-transformation as in [12, 11]: $V(t, x) = -\log \Phi(t, x)$. The expression for the optimal control (31) becomes

$$u(t, x) = \frac{1}{\Phi(t, \hat{X}_t)} Q \nabla_x \Phi(t, \hat{X}_t), \quad (34)$$

where Φ satisfies

$$\begin{aligned} \frac{\partial \Phi}{\partial t} + b(x)^* \nabla_x \Phi + \frac{1}{2} \text{tr}(Q \nabla_x^2 \Phi) &= \frac{\partial \Phi}{\partial t} + \mathcal{L} \Phi = 0, \quad t \in [t_k, t_{k+1}], \\ \Phi(t_{k+1}, x) &= e^{-g(Y_{t_{k+1}}, x)}. \end{aligned} \quad (35)$$

Equation (35) is a linear second order PDE. Hence, by the Feynman-Kac formula (c.f. Theorem 4.2 of [18]), the solution to (35) can be represented as

$$\Phi(t, x) = \mathbb{E}_{t,x} \left[e^{-g(Y_{t_{k+1}}, \eta_{t_{k+1}}^{t,x})} \right], \quad (36)$$

where $\mathbb{E}_{t,x}$ is the expectation with respect to the sample paths generated by the uncontrolled diffusion equation. That is, the probability measure induced by a process $\eta^{t,x}$ evolving according to

$$\begin{aligned} d\eta_s^{t,x} &= b(\eta_s^{t,x}) ds + \sigma(\eta_s^{t,x}) d\tilde{W}_s, \quad s \in [t, t_{k+1}], \\ \eta_t^{t,x} &= x, \end{aligned} \quad (37)$$

where \tilde{W} is a standard Brownian motion.

For the optimal control (34), the gradient of (36) is needed. In [21], the gradient is obtained using the Clark-Ocone formula in Malliavin calculus [26]. Here, we give the result in [38], where additive noise in the signal dynamics can be exploited to yield the gradient by way of another Feynman-Kac formula. In particular, this will apply for the Lorenz 1996 model investigated in Section 6.

Let $\Phi_x \equiv \nabla_x \Phi$. Taking the gradient of (35),

$$\begin{aligned} \frac{\partial \Phi_x}{\partial t} + \mathcal{L} \Phi_x + (\nabla_x b(x))^* \Phi_x &= 0, \quad t \in [t_k, t_{k+1}], \\ \Phi_x(t_{k+1}, x) &= -e^{-g(Y_{t_{k+1}}, x)} \nabla_x g(Y_{t_{k+1}}, x). \end{aligned} \quad (38)$$

Using the Feynman-Kac formula,

$$\Phi_x(t, x) = -\mathbb{E}_{t,x} \left[e^{-g(Y_{t_{k+1}}, \eta_{t_{k+1}}^{t,x})} e^{\int_t^{t_{k+1}} (\nabla_x b(\eta_s^{t,x}))^* ds} \nabla_x g(Y_{t_{k+1}}, \eta_{t_{k+1}}^{t,x}) \right], \quad (39)$$

where $\mathbb{E}_{t,x}$ is expectation with respect to the sample paths η generated by (37).

5.3 Updating Particle Weights

By applying control to the particles, the particle system is deviating from the true signal dynamics. This deviation has to be compensated for in the particle weights (23) when constructing the posterior. However, since the particles are evolved with control according to (26), the weights at observation times should be updated according to

$$w_{t_{k+1}}^i \propto \exp\left(-g(Y_{t_{k+1}}, \hat{X}_{t_{k+1}}^i)\right) \frac{d\mu_i}{d\hat{\mu}_i}(t_{k+1}, \hat{X}_{[t_k, t_{k+1}]}^i) w_{t_k}^i, \quad (40)$$

where $\frac{d\mu_i}{d\hat{\mu}_i}(t_k, \hat{X}_{[t_k, t_{k+1}]}^i)$ is the Radon-Nikodym derivative of:

- μ_i , the measure on the path space $C([t_k, t_{k+1}], \mathbb{R}^m)$ generated by a process that evolves according to the uncontrolled signal dynamics (37) in $[t_k, t_{k+1}]$, with starting point $(t_k, \hat{X}_{t_k}^i)$,

with respect to

- $\hat{\mu}_i$, the measure generated by the process that evolves according to the controlled dynamics (26), with starting point $(t_k, \hat{X}_{t_k}^i)$.

According to (31), we have $u(t, \hat{X}_t^i) = -\sigma \sigma^* \nabla_x V(t, \hat{X}_t^i)$. Let $u = \sigma v$, where $v(t, \hat{X}_t^i) \equiv -\sigma^* \nabla_x V(t, \hat{X}_t^i)$. Then, the particle evolution equation (26) becomes

$$d\hat{X}_t^i = b(\hat{X}_t^i)dt + \sigma (dW_t + v(t, \hat{X}_t^i)dt), \quad \text{for } t \in (t_k, t_{k+1}]. \quad (41)$$

Girsanov's theorem can now be used to perform a measure change that makes $B \equiv W + \int v dt$, a Brownian motion under the new measure. Doing so, we obtain

$$\frac{d\mu_i}{d\hat{\mu}_i}(t_{k+1}, \hat{X}_{[t_k, t_{k+1}]}^i) = \exp\left(-\int_{t_k}^{t_{k+1}} v(s, \hat{X}_s^i)^* dW_s - \frac{1}{2} \int_{t_k}^{t_{k+1}} v(s, \hat{X}_s^i)^* v(s, \hat{X}_s^i) ds\right). \quad (42)$$

With the derivations just presented, we can state the algorithm for a nudged particle filter adapted to the homogenized dynamics. For additional remarks on the nudged particle filter, including insight into the optimality and behavior of the particles, see [38]. Additional application and theory of this algorithm to reduced order filters, but without correlated sensor-signal noise, is given in [22, 21, 38]. Other related works are [20, 37].

Nudged Homogenized Hybrid Particle Filter (HHPF_c) Algorithm

1. At time t_k , set $w_{t_k}^j = 1/N, \forall j \in \mathcal{A}$ and

$$p(\xi_{t_k}^j | y_{t_k}) = \sum_{j \in \mathcal{A}} w_{t_k}^j \delta_{t_k}^j(\xi_{t_k}^j).$$

2. Apply the HMM multiscale integration technique and compute the averaged coefficients $\bar{b}, \bar{a}^{1/2}, \bar{h}$.
3. Generate the prior at t_{k+1} by advecting each particle under the averaged signal dynamics; that is (26) but with b, σ replaced by $\bar{b}, \bar{a}^{1/2}$,

$$p(\xi_{t_{k+1}}) = \sum_{j \in \mathcal{A}} w_{t_{k+1}}^j \delta_{t_{k+1}}^j(\xi_{t_{k+1}}).$$

- The optimal feedback control has been chosen according to (34), requiring the solution of the Feynman-Kac formulas (36) and (39).
 - And the weights have been updated by multiplication of (42).
 - For (36), (39), and (42), use $\bar{b}, \bar{a}^{1/2}, \bar{h}$ instead of b, σ, h .
4. Update the particle weights at the observation time according to

$$w_{t_{k+1}}^j \leftarrow w_{t_{k+1}}^j p(y_{t_{k+1}} | \xi_{t_{k+1}}).$$

- 4a. If $N_{\text{eff},k}$ is below a threshold (indicating particle degeneracy), then apply universal resampling and set

$$w_{t_{k+1}}^j = 1/N, \quad \forall j \in \mathcal{A}.$$

- 4b. Otherwise, compute the l_2 norm of the weights and re-normalize each

$$w_{t_{k+1}}^j \leftarrow w_{t_{k+1}}^j / |w|_2.$$

6 Application to Lorenz 1996

In this section we apply the PF, HHPF, and HHPF_c algorithms to a correlated sensor-signal noise filtering problem of the Lorenz 1996 model with continuous-time signal and discrete-time observation. We use the same Lorenz 1996 model as given in Section 3 (17, 18), with the system parameters defined in Section 3, and the parameters for the HMM defined in Section 3.1. For the observation process, we use the following

$$Y_{t_k}^\varepsilon = h(X_{t_k}^\varepsilon) + \int_{t_{k-1}}^{t_k} \alpha dW_s + \gamma U_{t_k}, \quad (43)$$

where $\{t_k\}$ are observation times and in the non-correlated case,

$$\alpha \equiv \mathbf{0}_{m \times m}, \quad \gamma \equiv \sigma_x,$$

and in the correlated case,

$$\alpha \equiv \frac{1}{\sqrt{2}} \sigma_x, \quad \gamma \equiv \frac{1}{\sqrt{2}} \sigma_x.$$

The choice of α, γ in (43) means that in both the non-correlated or correlated case, the observation has the same statistics. In (43), $h \equiv \text{Id}_{m \times m}$; an $m \times m$ identity matrix and acts on $x \in \mathbb{R}^m$ by matrix multiplication. In the case of the homogenized hybrid filters, the sensor function $\tilde{h}(\cdot)$, is a function of X_k^0 .

In the simulations that follow, we use an observation step-size $\Delta t = 10\delta_M = 0.1$ and total simulation time of $T = 20$, which roughly corresponds to 0.5 and 100 ‘real’ days according to [23]. The deterministic Lorenz ’96 model investigated in [23] has an error doubling time of roughly 1.6 ‘real’ days. In all simulations, the true signal is correlated, but we will conduct one experiment with the HHPF filter assuming a sensor-signal model of the non-correlated type; that is $\alpha = 0_{m \times m}$ and $\gamma = \sigma_x$. In all but one simulation, we use 16 particles ($N = 16$), with an effective number of 8 ($N_{\text{eff}} = 8$); for one HHPF_c experiment we will use $N = 8$ and $N_{\text{eff}} = 4$.

In total, we consider 5 experiments with their defining parameters given in Table 1. Each experiment consisting of 24 simulations. The average RMSE, calculated as follows,

$$\text{RMSE} = \sqrt{\sum_{k=1}^{T/\Delta t} |X_{t_k}^\varepsilon - \mathbb{E}[X_{t_k}^\varepsilon]|_2^2},$$

is shown for each experiment in Table 1, as well as the average simulation run-time.

Figure 8 shows the result of the PF applied to the Lorenz ’96 problem over the time interval $[0, 20]$. The average RMSE was 1.52 and average simulation time 1,019 seconds. With the exception of the interval $[1, 2.5]$, the PF with 16 particles is able to track well $X_t^{\varepsilon,1}$; the first component of X_t^ε , but at the expense of long simulation times. In Figs. 9 and 10 we show the corresponding result for the HHPF experiments. And in Figs. 12 and 11 the results when nudging is used.

As one might expect, the use of HHPF results in a slight degradation in the accuracy of the estimate of the signal in comparison to the PF for a fixed number of particles, but with a significant reduction in simulation run-time. For instance, Table 1 shows that the HHPF simulations result in more than a ten time speed-up over the PF. Figure 9 produces the least accurate tracking of the signal out of all experiments. This is expected, since this experiment does not model the correlated sensor-signal noise and filters on the homogenized dynamics. Figure 10 shows that an improvement in accuracy for the same run-time can be made by using the correlated algorithm in Section 4.3.

Figure 11 depicts the type of improvement in tracking that using nudging provides over HHPF. The HHPF_c solution in Fig. 11 uses the same number of particles as the HHPF simulations, $N = 16$, and uses four realizations for calculation of the Feynman-Kac formulas in (36) and (39). The calculation of the control, which is calculated once over each observation interval - and held fixed, results in a slower average run-time of 159 seconds per simulation, but with a much improved RMSE of the coarse-grain states. Figure 12 is also an HHPF_c simulation, but with the number of particles reduced to $N = 8$, which still results in good tracking due to the nudging of the particles, and a reasonable run-time of 110 seconds per simulation

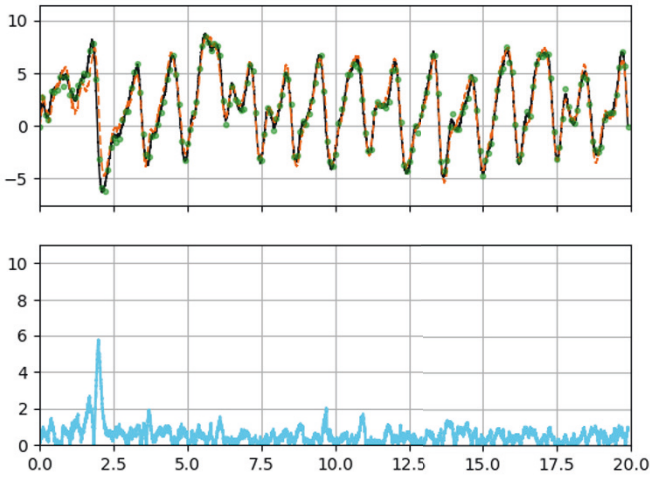


Fig. 8 PF, $\alpha = \gamma = \sigma_x / \sqrt{2}, N = 16, N_{\text{eff}} = 8$. Top graph: the signal $X_t^{\varepsilon,1}$ (first component) in black, the estimate $\mathbb{E}X_t^{\varepsilon,1}$ in orange, observations in green. Bottom graph: RMSE in light blue.

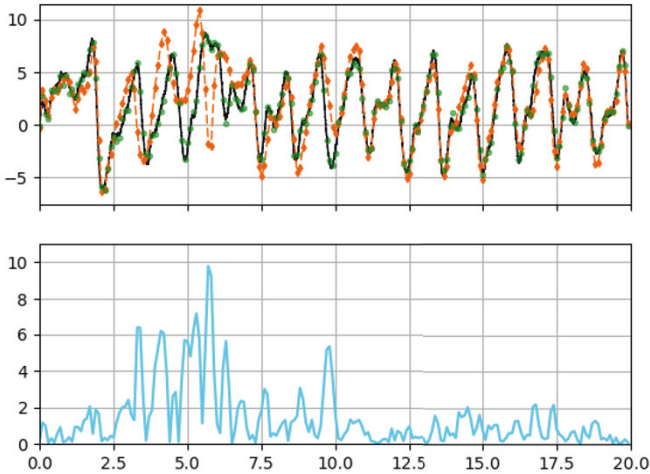


Fig. 9 HHPF, $\alpha = 0, \gamma = \sigma_x, N = 16, N_{\text{eff}} = 8$. Top graph: the signal $X_t^{\varepsilon,1}$ (first component) in black, the estimate $\mathbb{E}X_t^{\varepsilon,1}$ in orange, observations in green. Bottom graph: RMSE in light blue.

on average; the RMSE average of 1.30 is still lower than that of the PF average RMSE.

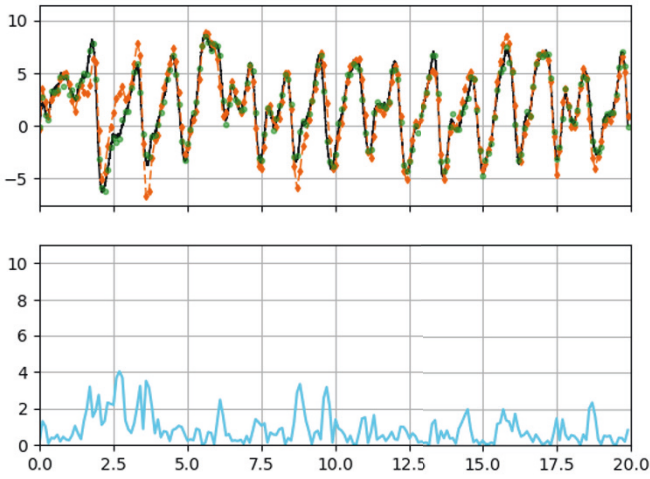


Fig. 10 HHPF with $\alpha = \gamma = \sigma_x/\sqrt{2}, N = 16, N_{\text{eff}} = 8$. Top graph: the signal $X_t^{\varepsilon,1}$ (first component) in black, the estimate $\mathbb{E}X_t^{\varepsilon,1}$ in orange, observations in green. Bottom graph: RMSE in light blue.

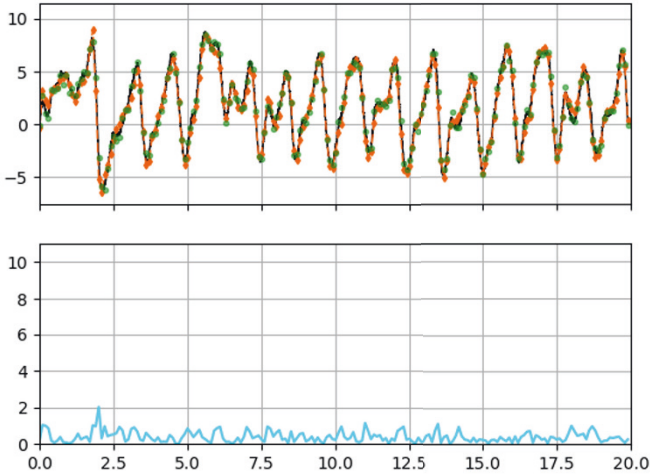


Fig. 11 HHPF_c with $\alpha = \gamma = \sigma_x/\sqrt{2}, N = 16, N_{\text{eff}} = 8$. Top graph: the signal $X_t^{\varepsilon,1}$ (first component) in black, the estimate $\mathbb{E}X_t^{\varepsilon,1}$ in orange, observations in green. Bottom graph: RMSE in light blue.

In Figs. 13 - 16, we provide a zoomed-in view of the interval $[2.5, 7.5]$ for the estimate of the signal in Figs. 8 - 12. Besides showing the signal, estimate of the

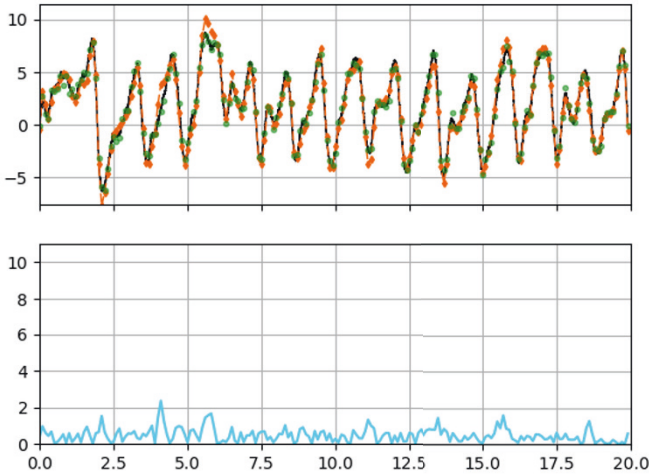


Fig. 12 HHPF_C with $\alpha = \gamma = \sigma_x/\sqrt{2}, N = 8, N_{\text{eff}} = 4$. Top graph: the signal $X_1^{\varepsilon,1}$ (first component) in black, the estimate $\mathbb{E}X_1^{\varepsilon,1}$ in orange, observations in green. Bottom graph: RMSE in light blue.

Table 1 Filtering results for various filter algorithms applied to the Lorenz 1996 model. RMSE integrated over time, and filter run-time (per simulation) averaged over 24 experiments.

Experiment	1st	2nd	3rd	4th	5th
Filter	PF	HHPF	HHPF	HHPF _C	HHPF _C
N_{eff}	8	8	8	8	4
N	16	16	16	16	8
α	$\sigma_x/\sqrt{2}$	0	$\sigma_x/\sqrt{2}$	$\sigma_x/\sqrt{2}$	$\sigma_x/\sqrt{2}$
γ	$\sigma_x/\sqrt{2}$	σ_x	$\sigma_x/\sqrt{2}$	$\sigma_x/\sqrt{2}$	$\sigma_x/\sqrt{2}$
RMSE	1.52	2.47	2.10	1.11	1.30
Run-Time	1019 s	85 s	85 s	159 s	110 s

signal, and observations in these figures, we also show the history of the particles (shown in light blue). The error in the observation of the signal is more apparent in these figures. One can also see when re-sampling occurs; a rapid collapse of particles far from the observations to locations closer to the observation at observation times. The diffusion of the particles between observation times, partly exacerbated by the chaotic property of the model, is also apparent. In Figs. 17 and 16, the particle traces in light blue show that although we apply control to the particles to nudge them towards observations, the running cost associated with applying control in (26) means that the control is not allowed to over-power the true dynamics by too much.

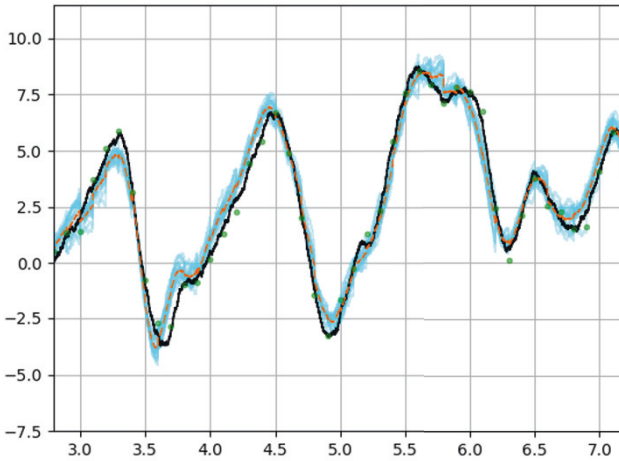


Fig. 13 PF with $\alpha = \gamma = \sigma_x / \sqrt{2}$, $N = 16$, $N_{\text{eff}} = 8$. The signal $X_t^{\epsilon,1}$ (first component) in black, the estimate $\mathbb{E}X_t^{\epsilon,1}$ in orange, observations in green, particles in light blue.

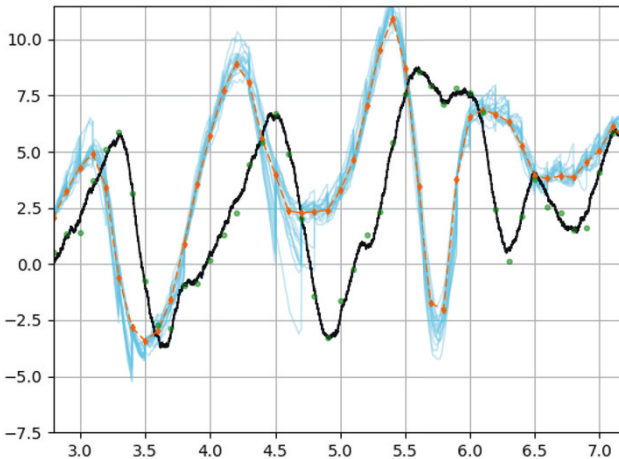


Fig. 14 HHPF with $\alpha = 0$, $\gamma = \sigma_x$, $N = 16$, $N_{\text{eff}} = 8$. The signal $X_t^{\epsilon,1}$ (first component) in black, the estimate $\mathbb{E}X_t^{\epsilon,1}$ in orange, observations in green, particles in light blue.

The last figure that we include is of the effective number, N_{eff} of the solutions shown in Figs. 8, 10, 11; having simulation parameters corresponding to the 1st, 3rd, and 4th experiments in Table 1 respectively. Figure 18 shows the effective number at

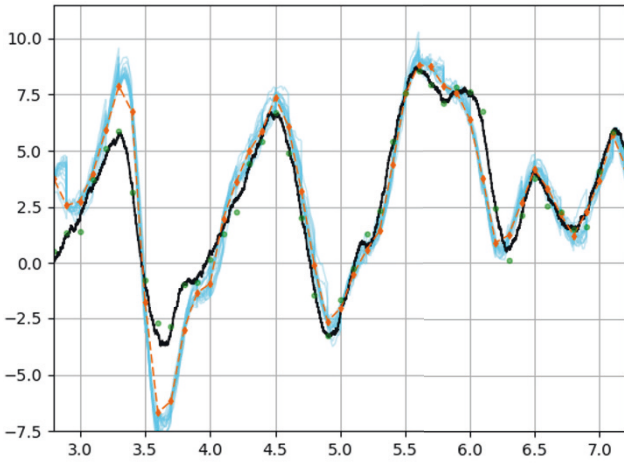


Fig. 15 HHPF with $\alpha = \gamma = \sigma_x/\sqrt{2}, N = 16, N_{\text{eff}} = 8$. The signal $X_t^{\epsilon,1}$ (first component) in black, the estimate $\mathbb{E}X_t^{\epsilon,1}$ in orange, observations in green, particles in light blue.

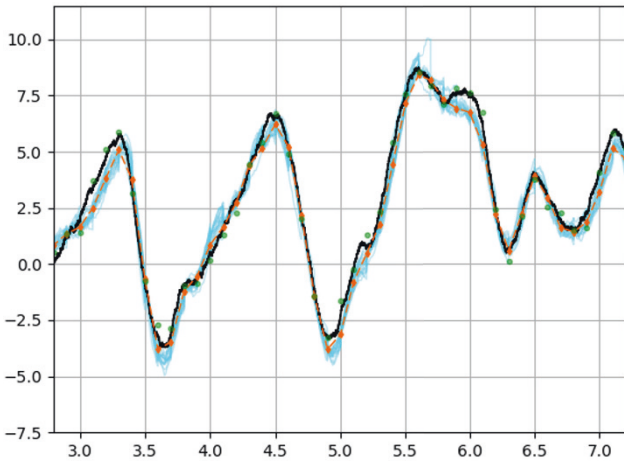


Fig. 16 HHPF_c with $\alpha = \gamma = \sigma_x/\sqrt{2}, N = 16, N_{\text{eff}} = 8$. The signal $X_t^{\epsilon,1}$ (first component) in black, the estimate $\mathbb{E}X_t^{\epsilon,1}$ in orange, observations in green, particles in light blue.

observation times for these simulations. Since we set the threshold $N_{\text{eff}} \leq 8$ to induce re-sampling in the simulation, this implies that for all three of these simulations, re-sampling occurred after every observation. The other simulations, corresponding

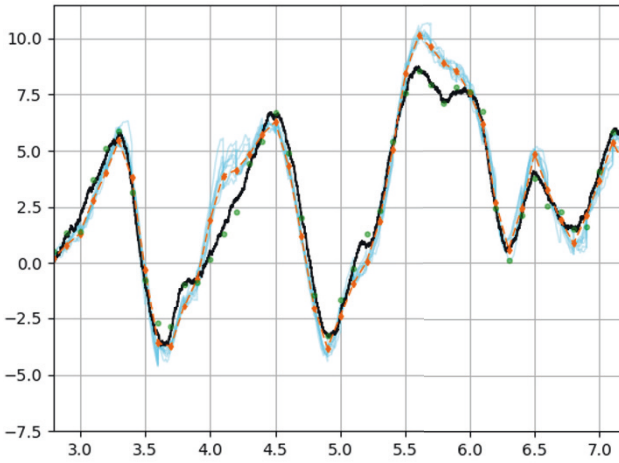


Fig. 17 HHPF_c with $\alpha = \gamma = \sigma_x/\sqrt{2}$, $N = 8$, $N_{\text{eff}} = 4$. The signal $X_t^{\epsilon,1}$ (first component) in black, the estimate $\mathbb{E}X_t^{\epsilon,1}$ in orange, observations in green, particles in light blue.

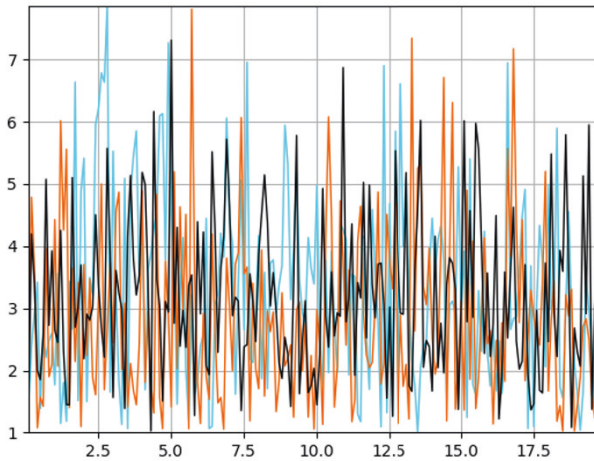


Fig. 18 The effective number N_{eff} at observation times versus time. PF shown in black, HHPF_c in orange, and HHPF in light blue. Values below 8 indicate re-sampling occurs.

to the results in Figs. 9 and 12, re-sampled on most, but not every observation. It is interesting that even with nudging, Fig. 18 shows that the HHPF_c approach on Lorenz '96 still results in significant resampling.

7 Conclusion

In this work we presented a wide range of numerical algorithms for filtering multiscale, high-dimensional, chaotic stochastic processes that may have a correlated sensor-signal observation process. The numerical algorithms are built from rigorous mathematical results. For the multiscale correlated observation case, we describe the mathematical techniques needed to show convergence of the x -marginal of the filter to a homogenized filter with a rate of $\varepsilon^{1/2}$ for a metric that generates the weak topology on the space of probability measures [5]. Using this result, the HHPF algorithm can be extended to the correlated case [32, 39, 22, 4]. The dynamics and properties of the Lorenz 1996 model were presented and then the algorithms for the PF and HHPF detailed. After describing the calculation of the likelihood distribution for the correlated case considered here, as well as the theory and algorithm for HHPF_c, then a number of experiments and their results were presented in Section 6.

The results of the experiments make clear the computational benefit of filtering on the homogenized dynamics. The addition of nudging in the particle dynamics, solving an optimal control problem that steers the particles toward areas where the observation is more likely, helps to further combat the high-dimensionality that may remain after reducing the state space size by using the HHPF instead of the PF. The nudging of particles is also helpful in combating the exponential error growth in chaotic systems. Lastly, the experiments showed the degradation in solution when correlation of the sensor-signal in the observation process is not correctly accounted for and modeled.

Current research is aimed at extending the aforementioned theoretical and computational results to the case of more than two time-scales and correlation in the observation process with fast time-scales. The results of this research can enable a general, efficient and flexible framework for data assimilation of a wide range of problems typically encountered in the engineering and physical sciences, as well as other fields.

Acknowledgements The authors acknowledge the support of the AFOSR under grant number FA9550-16-1-0390, NSERC under Discovery grant number 50503-10802, Fields-CQAM grant and TECSIS corporation.

References

1. Anderson, J.L.: An Ensemble Adjustment Kalman Filter for Data Assimilation. *Monthly Weather Review* **129**, 2884–2903 (2001)
2. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing* **50**(2), 174–188 (2002). DOI 10.1109/78.978374
3. Bain, A., Crisan, D.: *Fundamentals of Stochastic Filtering*. Springer (2009)

4. Beeson, R.: Reduced order nonlinear filters for multi-scale systems with correlated sensor noise. In: 21st International Conference on Information Fusion (FUSION) 2018 (FUSION 2018). Cambridge, United Kingdom (Great Britain) (2018)
5. Beeson, R., Namachchivaya, N.S., Perkowski, N.: Dimensional reduction in nonlinear filtering: Multi-scale systems and correlated sensor noise. In Preparation
6. Crisan, D.: Particle approximations for a class of stochastic partial differential equations. *Applied Mathematics and Optimization* **54**, 293–314 (2006)
7. Del Moral, P., Miclo, L.: Branching and Interacting Particle Systems Approximations of Feynman-Kac Formulae with Applications to Non-linear Filtering. In: *Séminaire de Probabilités XXXIV, Lecture Notes in Mathematics*, vol. 1729, pp. 1–145. Springer-Verlag Berlin (2000)
8. Doucet, A.: On sequential simulation-based methods for Bayesian filtering. Tech. rep., Cambridge University (1998)
9. E, W., Liu, D., Vanden-Eijnden, E.: Analysis of multiscale methods for stochastic differential equations. *Communication on Pure and Applied Mathematics* **58**, 1544–1585 (2005)
10. Fatkullin, I., Vanden-Eijnden, E.: A computational strategy for multiscale systems. *Journal of Computational Physics* **200**(2), 605–638 (2004)
11. Fleming, W.H.: Exit probabilities and optimal stochastic control. *Applied Mathematics and Optimization* **4**, 329–346 (1978)
12. Fleming, W.H.: PLogarithmic transformations and stochastic control, pp. 131–141. Springer Berlin Heidelberg, Berlin, Heidelberg (1982). DOI 10.1007/BFb0004532. URL <http://dx.doi.org/10.1007/BFb0004532>
13. Gordon, N.J., Salmond, D.J., Smith, A.F.M.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings F* **140**(2), 107–113 (1993)
14. Gustafsson, F., Saha, S.: Particle filtering with dependent noise. In: 2010 13th International Conference on Information Fusion, pp. 1–4 (2010). DOI 10.1109/ICIF.2010.5712052
15. Harlim, J., Kang, E.L.: Filtering partially observed multiscale systems with heterogeneous multiscale methods-based reduced climate models. *Mon. Wea. Rev.* **140**(3), 860–873 (2012)
16. Imkeller, P., Namachchivaya, N.S., Perkowski, N., Yeong, H.C.: Dimensional reduction in nonlinear filtering: A homogenization approach. *Ann. Appl. Probab.* **23**(6), 2290–2326 (2013). DOI 10.1214/12-AAP901. URL <http://dx.doi.org/10.1214/12-AAP901>
17. Kang, E.L., Harlim, J.: Filtering partially observed multiscale systems with heterogeneous multiscale methods-based reduced climate models. *Monthly Weather Review* **140**, 860–873 (2012)
18. Karatzas, I., Shreve, S.E.: *Brownian Motion and Stochastic Calculus*. Springer-Verlag New York Inc. (1988)
19. Kushner, H.J.: *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*. Birkhäuser, Boston (1990)
20. van Leeuwen, P.J.: Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quart. J. Royal Meteor. Soc.* **136**, 1991–1999 (2010)
21. Lingala, N., Perkowski, N., Yeong, H.C., Namachchivaya, N.S., Rapti, Z.: Probabilistic Engineering Mechanics. *Probabilistic Engineering Mechanics* **37**(C), 160–169 (2014)
22. Lingala, N., Sri Namachchivaya, N., Perkowski, N., Yeong, H.C.: Particle filtering in high-dimensional chaotic systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **22**(4), 047,509 (2012)
23. Lorenz, E.N.: Predictability: A problem partly solved. In: *Proceedings of the ECMWF Seminar on Predictability*, vol. 1, pp. 1–18. ECMWF (1996)
24. Majda, A., Timofeyev, I., Vanden-Eijnden, E.: A Mathematical Framework for Stochastic Climate Models. *Comm. Pure Appl. Math.* **54**, 891–974 (2001)
25. Majda, A.J., Timofeyev, I., Vanden-Eijnden, E.: Systematic Strategies for Stochastic Mode Reduction in Climate. *J. of Atmospheric Sciences* **60**, 1705–1722 (2003)
26. Øksendal, B.: *An introduction to Malliavin calculus with applications to economics* (1997). Lecture notes

27. Ott, E., Hunt, B.R., Szunyogh, I., Zimin, A.V., Kostelich, E.J., Corazza, M., Kalnay, E., Patil, D.J., Yorke, J.A.: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus* **56A**, 415–428 (2004)
28. Papanicolaou, G.C., Stroock, D., Varadhan, S.R.S.: Martingale approach to some limit theorems. In: *Papers from the Duke Turbulence Conference*. Duke University, Durham, North Carolina (1976)
29. Pardoux, E.: Stochastic partial differential equations and filtering of diffusion processes. *Stochastics* **3**(1-4), 127–167 (1980). DOI 10.1080/17442507908833142. URL <http://www.tandfonline.com/doi/abs/10.1080/17442507908833142>
30. Pardoux, E., Peng, S.: Backward doubly stochastic differential equations and systems of quasi-linear SPDEs. *Probability Theory and Related Fields* (98), 209–227 (1994)
31. Pardoux, E., Veretennikov, A.Y.: On Poisson Equation and Diffusion Approximation 2. *The Annals of Probability* **31**(3), 1166–1192 (2003)
32. Park, J.H., Namachchivaya, N.S., Yeong, H.C.: Particle filters in a multiscale environment: Homogenized hybrid particle filter. *Journal of Applied Mechanics* **78** (2011)
33. Rozovskii, B.L.: *Stochastic Evolution System: Linear Theory and Applications to Non-linear Filtering*. Kluwer Academic Publishers, Dordrecht (1990)
34. Saha, S., Gustafsson, F.: Marginalized particle filter for dependent gaussian noise processes. In: *2012 IEEE Aerospace Conference*, pp. 1–6 (2012). DOI 10.1109/AERO.2012.6187212
35. Vanden-Eijnden, E.: Numerical techniques for multi-scale dynamical systems with stochastic effects. *Communications in Mathematical Sciences* **1**(2), 385–391 (2003)
36. Veretennikov, A.Y.: On polynomial mixing bounds for stochastic differential equations. *Stochastic Processes and their Applications* **70**, 115–127 (1997)
37. Yang, T., Laugesen, R.S., Mehta, P.G., Meyn, S.P.: Multivariable feedback particle filter. *Automatica* **71**, 10–23 (2016)
38. Yeong, H.C., Beeson, R., Namachchivaya, N.S., Perkowski, N.: Particle filters with nudging in multiscale chaotic system: with application to the lorenz-96 atmospheric model. In *Submission* (2018)
39. Yeong, H.C., Park, J.H., Namachchivaya, N.S.: Particle filters in a multiscale environment: with application to the lorenz-96 atmospheric model. *Stochastics and Dynamics* **11**(02n03), 569–591 (2011). DOI 10.1142/S0219493711003450. URL <http://www.worldscientific.com/doi/abs/10.1142/S0219493711003450>



Postponing Collapse: Ergodic Control with a Probabilistic Constraint

Vivek S. Borkar and Jerzy A. Filar

Abstract We consider the long run average or ‘ergodic’ control of a discrete time Markov process with a probabilistic constraint in terms of a bound on the exit rate from a bounded subset of the state space. This is a natural counterpart of the more common probabilistic constraints in the finite horizon control problems. Using a recent characterization by Anantharam and the first author of risk-sensitive reward as the value of an average cost ergodic control problem, this problem is mapped to a constrained ergodic control problem that seeks to maximize an ergodic reward subject to a constraint on another ergodic reward. However, unlike the classical constrained ergodic reward/cost problems, this problem has some non-classical features due to a non-standard coupling between between the primary ergodic reward and the one that gets constrained. This renders the problem inaccessible to standard solution methodologies. A brief discussion of possible ways out is included.

1 Introduction

We live in an era where certain human development activities such as greenhouse gas emissions, clearing of forests or harvesting of fish could bring about a collapse of critical life support systems. Indeed, it is already widely believed that while the risks of such collapses cannot be entirely eliminated, they can be mitigated in the sense of postponing the onset of the most undesirable impacts until new technologies enable more sustainable development. Mathematically, this essential feature can be captured in terms of the first exit time from some desirable domain in the system’s state

Vivek S. Borkar

Department of Electrical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India e-mail: borkar.vs@gmail.com

Jerzy A. Filar

Centre for Applications in Natural Resource Mathematics, School of Mathematics and Physics, University of Queensland, St Lucia, QLD 4072, Australia e-mail: j.filar@uq.edu.au

space. Arguably, the associated risk mitigation problem concerns the time varying probability of such an exit and can be modelled as a control problem with a probabilistic constraint. Independently of the above motivation, there has been a considerable interest in stochastic control with probabilistic constraints wherein a suitable classical reward criterion on a finite horizon is sought to be maximized with a lower bound on the probability of the process remaining in a prescribed subset of the state space throughout this time horizon [2], [9]. Such a constraint, however, does not usually make sense for an infinite horizon control problem, as the aforementioned probability can be zero due to irreducibility properties of the process and the constraint is violated with certainty. A natural counterpart of the constraint then is to put a lower bound on the rate of exit from the prescribed set, given by the exponential rate of decay of the tail probability of the exit time from the set. We take this viewpoint here and map this constraint into a constraint on a risk-sensitive reward. Using a variational formulation of the risk-sensitive reward problem recently developed in [1], this is then converted to a constrained average reward (or ‘ergodic’ reward) control problem wherein one seeks to maximize a prescribed limiting average reward functional subject to a constraint on another limiting average reward functional. Such constrained Markov control problems have been extensively studied and one has in particular a linear programming formulation thereof in terms of the so called ergodic occupation measures. The present situation, however, turns out to be more complicated than the classical constrained Markov decision process framework (see, e.g., [3], [5]) because of a non-standard ‘running cost’ function that couples the primary and secondary objectives in a complicated manner. We give an equivalent formulation in terms of a static optimization problem over suitably defined sets of measures, with an additional constraint that reflects the above issue. By invoking a Lagrange multiplier, we can convert it into an optimization problem on product of two sets of measures specified by linear constraints, with a reward function that is separately strictly concave in its arguments. This turns out to be equivalent to a ‘team’ problem wherein two agents seek to maximize the same reward, albeit in a non-cooperative fashion.

While this contribution is strictly theoretical, we were inspired by the practical problem of sustainable management of a commercial fishery. In continuing this line of research, we intend to develop not only numerical methods for solving the non-trivial optimization problems that emerge, but also to adapt these methods to manage the risks of collapse of certain fisheries in Queensland.

We shall use the following notation throughout. Denote by $\mathcal{P}(\mathcal{X})$ the Polish space of probability measures on a Polish space \mathcal{X} with Prohorov topology ([4], Chapter 2), and by $C(\mathcal{X}), C_b(\mathcal{X})$ the space of continuous, resp. bounded continuous functions $\mathcal{X} \mapsto \mathcal{R}$.

Let \mathcal{N} denote the set of natural numbers, that is, $\{0, 1, 2, \dots\}$. Consider a controlled Markov chain $\{X_t, t \geq 0\}$ taking values in a Polish space S , controlled by a control process $\{Z_t, t \geq 0\}$ taking values in a compact metric space U , with con-

trolled transition kernel $(x, u) \in S \times U \mapsto p(dy|x, u) \in \mathcal{P}(S)$. Thus

$$P(X_{t+1} \in A | X_m, Z_m, m \leq t) = p(A | X_t, Z_t) \quad \forall t \text{ and } \forall A \subset S \text{ Borel.}$$

The $\{Z_t\}$ for which the above holds will be called an admissible control (process). We shall further say that $\{Z_t\}$ is a Markov policy if $Z_t = v(X_t, t)$ for a measurable $v : S \times \mathcal{N} \mapsto U$ and a randomized Markov policy if

$$P(Z_t \in B | X_m, Z_m, m \leq t) = \phi(B | X_t, t) \quad \forall B \subset S \text{ Borel}$$

for a measurable $\phi : S \times \mathcal{N} \mapsto \mathcal{P}(U)$. By standard practice of abuse of terminology, we use v , resp. ϕ to denote this control policy. A further special case is when there is no explicit time dependence of v, ϕ on the time variable, in other words, $Z_t = v(X_t)$ (respectively, $P(Z_t \in B | X_m, Z_m, m \leq t) = \phi(B | X_t)$) for a measurable $v : S \mapsto U$ (respectively, $\phi : S \mapsto \mathcal{P}(U)$). In this case, it will be called a stationary Markov (respectively, randomized stationary Markov) policy, abbreviated as SMP, RSMP respectively. Let π_0 be the law of X_0 .

We further assume the existence of a transition density $\varphi(y|x, u) > 0$ with respect to a positive measure λ on S with full support, that is,

$$p(dy|x, u) = \varphi(y|x, u)\lambda(dy). \quad (1)$$

2 The finite horizon control problem

We briefly describe the classical finite horizon control problem with probabilistic constraint, in order to motivate its infinite horizon version later.

Let S_0 be a proper subset of S which is the closure of its interior. Let

$$\tau = \min\{n \geq 0 : X_n \notin S_0\}$$

denote the first exit time from S_0 . We shall be interested in processes killed at τ , that is, on exit from S_0 . So without loss of generality, we set $S = S_0 \cup \{\Delta\}$ where Δ is a ‘coffin’ state, and correspondingly set

$$p(\{\Delta\}|x, u) := 1 - p(S_0|x, u), \quad p(\{\Delta\}|\Delta, u) = 1.$$

Correspondingly, redefine τ as $\tau = \min\{n \geq 0 : X_n = \Delta\}$. Without any loss of generality, we also assume that π_0 is supported on S_0 .

Let $E_{\kappa, t}[\cdot], P_{\kappa, t}(\cdot)$ denote resp. the expectation and probability for initial time t and initial law κ and let $t \wedge \tau := \min(t, \tau)$. If $\kappa = \delta_x :=$ the Dirac measure at x , we write $E_{x, t}[\cdot], P_{x, t}(\cdot)$ for simplicity. Consider the finite horizon constrained control

problem of maximizing the reward

$$E_{\pi_{0,0}} \left[\sum_{m=0}^{T \wedge \tau} r(X_m, Z_m) \right] \quad (2)$$

for a prescribed reward function $r \in C(S \times U)$, subject to the probabilistic constraint

$$P_{\pi_{0,0}}(\cup_{m=0}^T \{X_m \in S_0\}) \geq 1 - \nu \quad (3)$$

for a prescribed $1 > \nu > 0$. We assume feasibility, i.e., (3) holds under some control policy and the maximum reward under the stated constraint is finite. The point we want to highlight here is the fact that the above constraint may not make sense for $T = \infty$ because under common reachability/irreducibility conditions on the state process, the above probability may be zero.

3 Infinite horizon control problem

3.1 Problem formulation

In this section, we assume that S_0 and therefore $S = S_0 \cup \{\Delta\}$, are compact.

We assume that $P(\tau < \infty) = 1$. If we consider the above problem on an infinite time horizon, the probabilistic constraint (3) is then always violated as observed above, because with probability 1, state Δ will be reached in finite time. A natural extension then is to minimize the decay rate of the exit probability from S_0 , that is, maximize the quantity

$$\Gamma := \limsup_{T \uparrow \infty} \frac{1}{T} \log P(\tau > T) \quad (4)$$

over admissible control policies under consideration. For the time being, we shall confine our attention to randomized *stationary* Markov policies. Then the limsup above is a limit and is given by the principal eigenvalue of the positive and positively 1-homogeneous operator \mathcal{A} which we shall define soon. For an RSMP ϕ , we let $p_\phi(dy|x) := \int \phi(du|x) p(dy|x, u)$ denote the time-homogeneous transition kernel under ϕ . Define another transition probability kernel $x \in S_0 \mapsto q_\phi(dy|x) \in \mathcal{P}(S_0)$ by

$$q_\phi(dy|x) := (p_\phi(S_0|x))^{-1} p_\phi(dy|x)$$

and set

$$c_\phi(x) := \log p_\phi(S_0|x).$$

Define the operator $\mathcal{A}_\phi : C_b(S_0) \mapsto C_b(S_0)$ by: for $f \in C_b(S_0)$,

$$\begin{aligned} \mathcal{A}_\phi f &:= \int_{S_0} p_\phi(dy|x) f(y) \\ &= \int_{S_0} q_\phi(dy|x) e^{c_\phi(x)} f(y). \end{aligned}$$

As in the arguments leading to Theorem 2.2 of [1], we conclude that this is a strongly positive, strongly continuous, positively 1-homogeneous linear operator such that the Γ of (4) under RSMP ϕ , denoted by Γ_ϕ , is the principal eigenvalue of \mathcal{A}_ϕ guaranteed by the Krein-Rutman theorem [10]. We then replace (3) by the new constraint

$$-\Gamma_\phi \leq \eta. \tag{5}$$

Note that $\Gamma_\phi \leq 0$.

We replace the finite horizon reward by the infinite horizon average reward

$$\mathcal{W}_\phi := \liminf_{T \uparrow \infty} \frac{1}{T} E \left[\sum_{m=0}^T r(X_m, Z_m) \right]. \tag{6}$$

With this backdrop, our control problem is:

$$(P_0) \quad \text{Maximize over } \phi \text{ the quantity } \mathcal{W}_\phi \text{ subject to } -\Gamma_\phi \leq \eta.$$

3.2 An equivalent formulation

We recall a variational formula from [1] (Theorem 3.3 of *ibid.* specialized to the constant control case). Define

$$\mathcal{G} := \left\{ \zeta(dx, dy) = \zeta_0(dx) \zeta_1(dy|x) \in \mathcal{P}(S_0 \times S_0) : \int_{S_0} \zeta_0(dx) \zeta_1(dy|x) = \zeta_0(dy) \right\}, \tag{7}$$

that is, $\zeta_0(dx)$ is invariant under the transition kernel $\zeta_1(dy|x)$. Then Theorem 3.3 of [1] states that

$$\Gamma_\phi = \max_{\zeta \in \mathcal{G}} \left[\int \zeta(dx, dy) \left(c_\phi(x) - D(\zeta_1(dy|x) \| q_\phi(dy|x)) \right) \right], \tag{8}$$

where $D(\cdot \| \cdot)$ is the Kullback-Leibler divergence or ‘relative entropy’ defined by

$$\begin{aligned} D(\mu \| \mu') &:= \int \mu(dz) \log \left(\frac{d\mu}{d\mu'}(z) \right) \text{ if } \mu \ll \mu', \\ &= \infty \quad \text{otherwise.} \end{aligned}$$

The quantity being maximized above is seen to be the average reward for another controlled Markov chain with state space S_0 , control space $\mathcal{P}(S_0)$, and controlled transition kernel $Q(dy'|y, \zeta) = \zeta(dy')$, that is, the control itself is the law of the next state, with the cost per unit time for state-control pair (x, ζ) being given by

$$c_\phi(x) - D(\zeta(dy) || q_\phi(dy|x)).$$

Thus an SMP corresponds to a measurable map $y \in S_0 \mapsto \zeta(dy'|y)$, that is, a transition kernel. We consider a pair of controlled Markov chains $\{X_t\}, \{Y_t\}$ such that:

- the state space for $\{X_t\}$ is S_0 , the state space for $\{Y_t\}$ is also S_0 ,
- the control space for $\{X_t\}$ is $\mathcal{P}(U)$ and the control space for $\{Y_t\}$ is $\mathcal{P}(S_0)$,
- $\{X_t\}$ is governed by the RSMP $x \in S_0 \mapsto \phi(du|x) \in \mathcal{P}(U)$ and $\{Y_t\}$ is governed by the SMP $y \in S_0 \mapsto \zeta(dy'|y) \in \mathcal{P}(S_0)$,
- the transition kernel for $\{X_t\}$ under the above RSMP is $(x, \phi) \in S \times \mathcal{P}(U) \mapsto \int p(dx'|x, u)\phi(du|x) \in \mathcal{P}(S_0)$ and the transition kernel for $\{Y_t\}$ under the above SMP is $(y, \zeta) \in S_0 \times \mathcal{P}(S_0) \mapsto \zeta(dy'|y) \in \mathcal{P}(S_0)$.

Here ζ, γ factorize as

$$\zeta(dx, dy) = \zeta^0(dx)\zeta^1(dy|x), \quad \gamma(dx, du) = \gamma_0(dx)\phi(du|x).$$

We then have the static optimization (*not* a linear program) formulation for our problem as follows.

(P) Maximize over the pair $(\phi(\cdot|\cdot), \zeta_1(\cdot|\cdot))$ the reward $\int r(x, y)\gamma(dx, du)$ subject to

$$\gamma_0(dy) = \int \gamma(dx, du)p(dy|x, u), \quad (9)$$

$$\gamma(S_0 \times U) = 1, \quad (10)$$

$$\gamma \geq 0, \quad (11)$$

$$\zeta_0(dx) = \int \zeta^0(dx)\zeta^1(dy|x), \quad (12)$$

$$\zeta(S_0 \times S_0) = 1, \quad (13)$$

$$\zeta \geq 0, \quad (14)$$

and

$$\int \zeta(dx, dy)(c_\phi(x) - D(\zeta^1(dy|x) || q_\phi(dy|x))) \geq -\eta. \quad (15)$$

We state our main result as a theorem.

Theorem 1 The optimization problem **(P)** is equivalent to our constrained control problem (P_0) .

Proof (Sketch) We make the following observations:

- Constraints (9)-(11) characterize $\gamma(dx, du) = \gamma_0(dx)\phi(du|x)$ as an ‘ergodic occupation measure’ [5] wherein γ_0 is the unique stationary distribution under the RSMP ϕ . Note that the uniqueness follows from our assumption $\varphi(y|x, u) > 0$.
- Constraints (12)-(14) characterize ζ as an element of \mathcal{G} .
- Constraint (15) is equivalent to (5) for the RSMP ϕ .

The equivalence follows. \square

It is the last constraint (15) that makes this a hard problem, because the map $\gamma(dx, du) = \gamma_0(dx)\phi(du|x) \mapsto \phi(du|x)$ is anything but simple. A silver lining is the fact that the constraints on γ do not involve ζ , only the constraints on ζ involve γ though the dependence of (15) on ϕ . In particular, if we introduce a ‘Lagrange multiplier’ $\Lambda \geq 0$ associated with constraint (15), then the problem becomes:

(P*) Maximize

$$\int r(x, y)\gamma(dx, du) + \Lambda(\eta + \int \zeta(dx, dy)(c_\phi(x) - D(\zeta^1(\cdot|x)||q_\phi(\cdot|x)))) \quad (16)$$

subject to (9)-(14). Note that the constraints are now separated, that is, constraints (9)-(11) involve only γ whereas the constraints (12)-(14) involve only ζ . This is now an optimization problem associated with the ergodic control problem:

(P̄) Maximize over $\phi(du|\cdot), \zeta(dy|\cdot)$ the reward

$$\liminf_{T \uparrow \infty} \frac{1}{T} E \left[\sum_{m=0}^{T-1} \left(\int \phi(du|X_m) r(X_m, u) + \Lambda(\eta + (c_\phi(Y_m) - D(\zeta^1(dy|Y_m)||q_\phi(dy|Y_m)))) \right) \right]. \quad (17)$$

where $\{X_n\}, \{Y_n\}$ are controlled Markov chains introduced above. Note that this is an ergodic *team* problem because the two control choices ϕ and ζ are made non-cooperatively, albeit for a common objective. Also, there are further restrictions on the ‘information structure’ of the two controllers: the respective controls ϕ, ζ^1 are constrained to depend only on the corresponding current state X_n, Y_n respectively. This leads to difficulties which we discuss in the next section.

4 Remarks on computational schemes

Problem P* has linear constraints that are separate in the two variables γ, ζ , but the reward is not separable. The reward function is in fact strictly concave in each variable when the other variable is kept constant, but it is not jointly concave, which

makes the problem hard. Strict concavity implies that the maximizer in either variable with the other kept fixed is unique and depends continuously on the latter. Using this, it is easy to see that alternating maximization will lead to a local maximum.

Note that once ζ is fixed, the constrained maximization with respect to γ is a linear program for an ergodic control problem, whereas once γ is fixed, the constrained maximization with respect to ζ is a concave maximization problem which too can be made into a linear program by considering RSMP in place of SMP which means $\mathcal{P}(\mathcal{P}(S_0))$ -valued controls. Thus alternating maximization amounts to alternating linear programs. In principle one could replace these linear programs by alternative computational schemes such as policy or value iteration, see, e.g., [8], [11]. The difficulty here is that while this is possible for control of $\{X_n\}$ with ζ^1 frozen, it is not so easy for control of $\{Y_n\}$ with ϕ fixed, because there is no irreducibility type condition available that would be required for justifying such schemes. In fact this is so even when S_0 is finite, because the control space is not. If one approximates the control space by a finite set as well, then one has the extended linear and dynamic programming formulations that cover the general case ([13], Chapter 9). All this is for a fixed value of Λ , the Lagrange multiplier which is unknown.

To solve the overall constrained optimization problem, one has to also recursively learn the Lagrange multiplier using a ‘primal-dual’ philosophy. That is, run the iteration

$$\Lambda_{n+1} = [\Lambda_n - s(\eta + \Gamma_n)]^+. \quad (18)$$

Here $[x]^+ := \max(x, 0)$ is the projection to $[0, \infty)$ that ensures non-negativity of $\Lambda_n, n \geq 0$. This is then a constrained gradient descent for the Lagrange multiplier. The parameter $0 < s \ll 1$ is a small time step that renders this iteration ‘incremental’, that is, it moves on a slower time scale compared to the alternating maximization described above. Using the ‘two time scale’ approach of [6], Chapter 6, we treat Λ_n as ‘quasi-static’, that is, treat it as constant, whence the maximization over primal policy using value iteration or linear program tracks the optimal reward corresponding to λ_n ; see *ibid.* for a precise statement. Note that the objective function is linear in Λ . After the maximization over the primal variables, it is convex in Λ . Then by Danskin’s theorem [7], (18) will be a projected subgradient descent guaranteed to converge to a neighborhood of the global minimum, in other words, the Lagrange multiplier. The difficulty here is that the ‘primal’ alternating maximization only ensures a local maximum, so this will yield at best only a local constrained maximum.

Acknowledgements This work was initiated when the first author was visiting the Centre for Applications in Natural Resource Mathematics, University of Queensland. It is supported by the Australian Research Council Discovery Grant DP180101602. VSB was also supported in part by a J. C. Bose Fellowship from the Department of Science and Technology, Government of India.

References

1. Anantharam, V., Borkar, V. S.: A variational formula for risk-sensitive reward. *SIAM Journal of Control and Optimization*. **55(2)**, 961-988 (2017).
2. Andrieu, L., Cohen, J., Vázquez-Abad, F. J.: Gradient-based simulation optimization under probability constraints. *European Journal of Operational Research*. **212(2)**, 345-351 (2011).
3. Borkar, V. S.: *Topics in Controlled Markov Chains*. Pitman Research Notes in Math. No. 240, Longmans Scientific and Technical, Harlow, UK (1991).
4. Borkar, V. S.: *Probability Theory: An Advanced Course*. Springer Verlag, New York (1995).
5. Borkar, V. S.: Convex analytic methods in Markov decision processes. In: Feinberg E. A., Shwartz A. (eds.) *Handbook of Markov Decision Processes*, pp. 347-375. Kluwer Academic Publishers, Norwell, Mass. (2002).
6. Borkar, V. S.: *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Publishing Agency, New Delhi, and Cambridge University Press, Cambridge, UK (2008).
7. Danskin, J. M.: Theory of max-min, with applications. *SIAM Journal of Applied Mathematics*. **14(4)** (1966), 641-664.
8. Hernández-Lerma, O., Lasserre, J. B.: Policy iteration for average cost Markov control processes on Borel spaces. *Acta Applicandae Mathematicae*. **47**, 125-154 (1997).
9. Kang, B., Filar, J. A.: Time consistent dynamic risk measures. *Mathematical Methods of Operations Research* **63(1)** (2006), 169-186.
10. Krein, M. G., Rutman, M. A.: Linear operators leaving invariant a cone in Banach spaces. *Uspekhi Mat. Nauk*. **3(1)**, 3-95 (1948).
11. Meyn, S. P.: The policy iteration algorithm for average reward Markov decision processes with general state space. *IEEE Transactions on Automatic Control*. **42(12)**, 1663-1680 (1997).
12. Milgrom, P., Segal, I.: Envelope theorems for arbitrary choice sets. *Econometrica*. **70(2)**, 2002, 583-601.
13. Puterman, M.: *Markov Decision Processes: Discrete Dynamic Programming*. John Wiley and Sons, Hoboken, NJ, 1994.



Resource Sharing Networks and Brownian Control Problems

Amarjit Budhiraja* and Michael Conroy

Abstract We consider a family of resource sharing networks that were introduced in the work of Massoulié and Roberts (2000) as models for Internet flows and study an optimal stochastic control problem associated with the dynamic allocation of resource capacities to jobs in the system. Since these stochastic control problems are in general intractable, we analyze the system in a heavy traffic regime where one can formally approximate these control problems by certain Brownian control problems (BCP). It is shown that, both for a discounted cost and an ergodic cost criterion, an appropriate BCP gives a lower bound on the best achievable asymptotic cost under any sequence of admissible policies. The lower bounds established in this work show that the threshold control policies constructed in Budhiraja and Johnson (2017), which achieve the *Hierarchical Greedy Ideal* (HGI) performance (cf. Harrison et al. (2014)) in the heavy traffic limit, are in fact asymptotically optimal when certain monotonicity conditions on the cost function are satisfied.

1 Introduction

In this work we consider a family of resource sharing networks that were introduced in the work of Massoulié and Roberts [13] as models for Internet flows. A fundamental problem for such networks is to construct dynamic control policies that allocate resource capacities to jobs in the system, in an optimal manner. Optimality is typically formulated in terms of an appropriate cost function which turns the problem into that of optimal stochastic control. In general such control problems

Amarjit Budhiraja
University of North Carolina at Chapel Hill, e-mail: amarjit@unc.edu

Michael Conroy
University of North Carolina at Chapel Hill, e-mail: mconroy@live.unc.edu

* Research supported in part by the National Science Foundation (DMS- 1814894).

are intractable and therefore one considers an asymptotic formulation under a suitable scaling. When the cost function is of an infinite horizon discounted form, the papers [8, 11, 9] formulate certain Brownian control problems (BCP) that formally approximate the system manager's control under heavy traffic conditions. The goal of this work is to establish a rigorous asymptotic relationship between the network control problem and the corresponding Brownian control problem. We show that if a control policy is *admissible* in an appropriate sense (see Definition 3), the associated cost of using this policy is asymptotically bounded below by the value function of the corresponding BCP (namely, the optimal cost in the BCP). Thus the BCP gives a lower bound on the best achievable asymptotic cost under any sequence of admissible policies. A similar result for a broad family of unitary networks (cf. [3]) was established in [4]. A basic difference between unitary networks and resource sharing networks considered here is that in the former each job is processed by a single resource at any given time instant whereas in resource sharing networks a job may be processed *simultaneously* by several resources. Because of this basic structural difference the results obtained in [4] do not carry over to the setting of interest in the current work. In addition to a discounted cost problem, in this work, we also consider an ergodic cost criterion (see (11)). We formulate an appropriate Brownian control problem that governs the scaling limit of the network control problems under heavy traffic for this criterion. Under this criterion as well the cost of an admissible control policy is asymptotically bounded below by the value function of the corresponding BCP. Due to space limitations we only provide a sketch of the proof of this result. Detailed proof will be reported elsewhere.

A recent work [5] constructs, under several conditions on the system parameters, explicit threshold based control policies that achieve the so called *Hierarchical Greedy Ideal* (HGI) performance in the heavy traffic limit. We refer the reader to [10] for background and discussion of HGI policies. In general, the HGI, namely the asymptotic cost associated with HGI policies, may not be optimal in the associated Brownian control problem. However, using the minimality property of the one dimensional Skorohod map it can be seen that if the holding cost satisfies certain monotonicity properties, the HGI is indeed the optimal cost in the BCP both for the discounted and the ergodic cost. Therefore, together with the lower bounds established in the current work, the results of [5] say that the threshold control policies constructed in that work, under certain monotonicity conditions on the cost function (and under the conditions on system parameters assumed in [5]), are in fact asymptotically optimal. We discuss this point in Remark 1. Finally we remark that, although in the current work we only establish a lower bound, one expects that, under quite general conditions, the value functions of the network control problems should in fact converge to the value functions of the associated BCP. Such a result for unitary networks and with an infinite horizon discounted cost was established in [4]. This convergence problem for resource sharing networks considered here is currently open.

The chapter is organized as follows. In Section 2 we introduce the class of networks that will be studied. We also present the main conditions and the cost criteria of interest. Section 3 presents the equivalent workload formulations of the Brown-

ian control problem that arise on taking a formal heavy traffic limit of the network control problems of Section 2. In Section 4 we present the main result (Theorem 2) of this work. Finally, Section 5 gives the proof of Theorem 2.

Notation and Conventions. Inequalities for vectors will be interpreted componentwise. We will denote by $\mathbf{1}$ the vector of ones of an appropriate dimension. Convergence in probability will be denoted as \xrightarrow{P} and convergence in distribution will be denoted as \Rightarrow . All stochastic processes in this work will take values in \mathbb{R}^d for some d and will either be given on the time interval $[0, T]$ or $[0, \infty)$. All of these processes will have sample paths that are continuous from the right and have left limits (RCLL). We will denote by \mathcal{D}^m (resp. \mathcal{D}_+^m) the space of RCLL functions from $[0, \infty)$ to \mathbb{R}^m (resp. \mathbb{R}_+^m), equipped with the usual Skorohod topology, and by \mathcal{C}^m (resp. \mathcal{C}_+^m) the space of continuous functions from $[0, \infty)$ to \mathbb{R}^m (resp. \mathbb{R}_+^m), equipped with the local uniform topology. Unless specified otherwise all processes are given on the probability space (Ω, \mathcal{F}, P) . For a Polish space S , let $\mathcal{P}(S)$ denote the space of all probability measures on S equipped with the usual weak convergence topology. A collection of S -valued random variables is said to be tight if the corresponding collection of probability laws forms a relatively compact set in $\mathcal{P}(S)$. We will denote by δ_x the Dirac probability measure at the point x . A collection of \mathcal{D}^m -valued random variables is said to be \mathcal{C} -tight if every sequence in the collection has a convergent (in distribution) subsequence whose limit is in \mathcal{C}^m a.s.

2 Network Control Problems

Consider for each $r \in \mathbb{N}$ a stochastic processing network \mathcal{N}^r with J types of jobs and I resources for processing them. Here r is a scaling parameter and as $r \rightarrow \infty$, the system approaches criticality in a suitable sense. All the networks in the collection have a similar structure described through an $I \times J$ matrix K with $K_{ij} = 1$ if resource i works on job type j and $K_{ij} = 0$ otherwise. We will assume that for each subset of resources, there is at most one job type with it as the associated set of resources, or equivalently that no two columns of K are identical.

Let for $m \in \mathbb{N}$, $\mathbb{N}_m \doteq \{1, \dots, m\}$. In particular, $\mathbb{N}_I \doteq \{1, \dots, I\}$ and $\mathbb{N}_J \doteq \{1, \dots, J\}$. We will assume the following local traffic condition:

Condition 1 Let $\mathcal{R}_j = \{i \in \mathbb{N}_I : K_{ij} = 1\}$ be the set of resources that work on type j jobs, and let $\mathcal{J} = \{j \in \mathbb{N}_J : \sum_{i=1}^I K_{ij} = 1\}$ be the collection of all job types that use only one resource. Then, $\bigcup_{j \in \mathcal{J}} \mathcal{R}_j = \mathbb{N}_I$.

The above condition, which was first introduced in [12], says that for each resource there is a unique job-type that only requires service from that resource.

For job type $j \in \mathbb{N}_J$, let $\{u_j^r(k)\}_{k \in \mathbb{N}}$ be the i.i.d. inter-arrival times and $\{v_j^r(k)\}_{k \in \mathbb{N}}$ be the associated i.i.d. amounts of work. For each r , the random variables in the collection $\{u_j^r(k), v_j^r(k), k \in \mathbb{N}, j \in \mathbb{N}_J\}$ are taken to be mutually independent. We will assume that

$$P(u_j^r(1) > 0) = P(v_j^r(1) > 0) = 1 \quad \text{for all } r \text{ and } j, \quad (1)$$

and that

$$\{u_j^r(1)^2\}_r \text{ and } \{v_j^r(1)^2\}_r \text{ are uniformly integrable for each } j. \quad (2)$$

Let $\alpha_j^r = 1/E[u_j^r(1)]$ and $\beta_j^r = 1/E[v_j^r(1)]$. Also let $\sigma_j^{u,r}$ and $\sigma_j^{v,r}$ denote the standard deviations of $u_j^r(1)$ and $v_j^r(1)$, respectively. Define the collection of renewal processes

$$A_j^r(t) = \max \left\{ k \in \mathbb{N} : \sum_{i=1}^k u_j^r(i) \leq t \right\}, \quad j \in \mathbb{N}_J, t \in [0, \infty)$$

and

$$S_j^r(t) = \max \left\{ k \in \mathbb{N} : \sum_{i=1}^k v_j^r(i) \leq t \right\}, \quad j \in \mathbb{N}_J, t \in [0, \infty).$$

The capacity for each resource $i \in \mathbb{N}_I$ is denoted by C_i . This means that if at any time instant work of type $j \in \mathbb{N}_J$ is being processed at rate x_j then we must have $C \geq Kx$. A control policy in the r -th network is a J -dimensional stochastic process $B^r(t) = \{B_j^r(t)\}_{j \in \mathbb{N}_J}$, $0 \leq t < \infty$ which describes the amount of type- j work processed by time t . Associated with control process B^r is the I -dimensional capacity-utilization process $T^r = KB^r$, so that $T_i^r(t)$ represents the amount of work processed by resource i by time t . If we let $U^r(t) = tC - T^r(t)$, then $U_i^r(t)$ denotes the unused capacity of resource i by time t . We denote by $Q_j^r(t)$, $j \in \mathbb{N}_J$, the number of type- j jobs in the queue at time instant t . Then the state equation is given as

$$Q^r(t) = q^r + A^r(t) - S^r(B^r(t)), t \geq 0,$$

where $q^r \in \mathbb{N}_J$ denotes the initial queue length vector.

A control policy $\{B^r(t)\}$ is required to satisfy the following condition.

Condition 2 For every $r \in \mathbb{N}$ and P a.e. ω

- (i) (Monotonicity and Continuity) $t \mapsto B^r(t)$ is an absolutely continuous, nonnegative, nondecreasing function from $[0, \infty) \rightarrow \mathbb{R}^J$ with $B^r(0) = 0$.
- (ii) (Resource Constraint)

$$C \geq K \frac{d}{dt} B^r(t) \quad \text{for a.e. } t \geq 0. \quad (3)$$

- (iii) (Feasibility)

$$Q^r(t) \geq 0 \quad \text{for all } t \geq 0. \quad (4)$$

We will require one additional natural non-anticipativity condition on the control policies that will be introduced later below.

Let $\rho_j^r = \alpha_j^r / \beta_j^r$ and $\rho^r = (\rho_1^r, \dots, \rho_J^r)^T$. The following will be our main heavy traffic condition:

Condition 3 (*Heavy Traffic Condition*) For each $j \in \mathbb{N}_J$, there exist $\alpha_j, \beta_j \in (0, \infty)$ and $\bar{\alpha}_j, \bar{\beta}_j \in \mathbb{R}$ such that $\lim_{r \rightarrow \infty} r(\alpha_j^r - \alpha_j) = \bar{\alpha}_j$, $\lim_{r \rightarrow \infty} r(\beta_j^r - \beta_j) = \bar{\beta}_j$, and there exist $\sigma_j^u, \sigma_j^v \in (0, \infty)$ such that $\lim_{r \rightarrow \infty} \sigma_j^{u,r} = \sigma_j^u$, $\lim_{r \rightarrow \infty} \sigma_j^{v,r} = \sigma_j^v$.

Furthermore, with $\rho_j = \alpha_j/\beta_j$ for each $j \in \mathbb{N}_J$ and $\rho = (\rho_1, \dots, \rho_J)^T$, $C = K\rho$.

Note that Condition 3 in particular says that $\lim_{r \rightarrow \infty} r(\rho^r - \rho) = \eta$ where, for $j \in \mathbb{N}_J$, $\eta_j = \beta_j^{-2}(\bar{\alpha}_j\beta_j - \alpha_j\bar{\beta}_j)$.

Define the I -dimensional workload process W^r by

$$W^r(t) = KMQ^r(t), \quad (5)$$

where M is the $J \times J$ diagonal matrix with entries $1/\beta_j$.

When considering scaling limits, the following two types of scaled processes will be considered.

Definition 1. For $r \in \mathbb{N}$ the fluid-scaled versions of the processes $A^r, S^r, B^r, T^r, U^r, Q^r$, and W^r , are defined as

$$\begin{aligned} \bar{A}^r(t) &= r^{-2}A^r(r^2t), & \bar{S}^r(t) &= r^{-2}S^r(r^2t), \\ \bar{Q}^r(t) &= r^{-2}Q^r(r^2t), & \bar{W}^r(t) &= r^{-2}W^r(r^2t), \\ \bar{B}^r(t) &= r^{-2}B^r(r^2t), & \bar{T}^r(t) &= r^{-2}T^r(r^2t), \\ \bar{U}^r(t) &= r^{-2}U^r(r^2t), \end{aligned}$$

and the corresponding diffusion-scaled processes are given as

$$\begin{aligned} \hat{A}^r(t) &= r^{-1}(A^r(r^2t) - r^2t\alpha^r), & \hat{S}^r(t) &= r^{-1}(S^r(r^2t) - r^2t\beta^r), \\ \hat{Q}^r(t) &= r^{-1}Q^r(r^2t), & \hat{W}^r(t) &= r^{-1}W^r(r^2t), \\ \hat{B}^r(t) &= r^{-1}B^r(r^2t), & \hat{T}^r(t) &= r^{-1}T^r(r^2t), \\ \hat{U}^r(t) &= r^{-1}U^r(r^2t). \end{aligned}$$

Consider the processes

$$\hat{X}^r(t) = \hat{A}^r(t) - \hat{S}^r(\bar{B}^r(t)), \quad \hat{Y}^r(t) = rt\rho - \hat{B}^r(t), \quad t \geq 0. \quad (6)$$

Then with $\hat{\theta}_j^r(t) = rt(\alpha^r - \alpha) - r(\beta^r - \beta)\bar{B}^r(t)$, we have the relationship

$$\hat{Q}^r(t) = \hat{q}^r + \hat{X}^r(t) + \hat{\theta}^r(t) + M^{-1}\hat{Y}^r(t), \quad K\hat{Y}^r(t) = \hat{U}^r(t) \quad (7)$$

where $\hat{q}^r = \hat{Q}^r(0)$ and the second equality follows from Condition 3. We will make the following assumption on the initial condition.

Condition 4 For some $q \in \mathbb{R}_+^J$, $\hat{q}^r \rightarrow q$ as $r \rightarrow \infty$.

We now introduce one final requirement on control policies that will be considered here. Roughly speaking, this requirement says that a control policy must be nonanticipative, namely it can only use the information on events (e.g. arrival of jobs,

service completions) in the system that have occurred up to the current instant. In order to formulate the condition we begin by introducing certain multiparameter filtrations.

Definition 2. For $m = (m_1, \dots, m_J)$ and $n = (n_1, \dots, n_J) \in \mathbb{N}_J$, let

$$\mathcal{F}^r(m, n) = \sigma \{u_j^r(m'_j), v_j^r(n'_j) : m'_j \leq m_j, n'_j \leq n_j, j \in \mathbb{N}_J\}.$$

Then $\{\mathcal{F}^r(m, n), m, n \in \mathbb{N}^J\}$ is a multiparameter filtration generated by the interarrival and service times with the following partial ordering:

$$(m^1, n^1) \leq (m^2, n^2) \text{ if and only if } m_j^1 \leq m_j^2 \text{ and } n_j^1 \leq n_j^2 \text{ for all } j.$$

Let

$$\mathcal{F}^r = \sigma \left\{ \bigcup_{(m,n) \in \mathbb{N}^{2J}} \mathcal{F}^r(m, n) \right\}.$$

We can now define the class of admissible control policies.

Definition 3. For $r \in \mathbb{N}$, a J -dimensional process B^r is said to be an *admissible resource allocation policy* or an *admissible control policy* for network \mathcal{N}^r if it satisfies Condition 2 and the following two additional conditions:

(i) If for each r and t , we define the \mathbb{N}^{2J} -valued random variable

$$\tau^r(t) = (\tau^{r,A}(t), \tau^{r,S}(t)) = ((A^r(r^2t) + \mathbf{1})^T, (S^r(B^r(r^2t)) + \mathbf{1})^T),$$

where $\mathbf{1} = (1, 1, \dots, 1)^T \in \mathbb{N}_J$, then for each $t \geq 0$, $\tau^r(t)$ is an $\mathcal{F}^r(m, n)$ -stopping time.

(ii) If we define the filtration

$$\begin{aligned} \mathcal{G}^r(t) &= \mathcal{F}^r(\tau^r(t)) \\ &= \sigma \{A \in \mathcal{F}^r : A \cap \{\tau^r(t) \leq (m, n)\} \in \mathcal{F}^r(m, n) \text{ for all } (m, n) \in \mathbb{N}^{2J}\}, \end{aligned}$$

then the process $\hat{U}^r \doteq K\hat{Y}^r$ is $\{\mathcal{G}^r(t)\}$ -adapted.

A sequence of control policies $\{B^r\}_r$ is called admissible if, for each r , B^r is an admissible control policy for network \mathcal{N}^r . Denote the class of all admissible sequences as \mathcal{A} .

Parts (i) and (ii) of the above condition are satisfied for a very broad family of natural control policies (cf. [4, Theorem 5.4]).

We note that the requirement in (3), implies that for any admissible control policy B^r , the process $U^r(t) = tC - KB^r(t)$ is nonnegative and nondecreasing, so that for $s \leq t$,

$$0 \leq K(B^r(t) - B^r(s)) = (t - s)C - (U^r(t) - U^r(s)) \leq (t - s)C.$$

Hence for each $i \in \mathbb{N}_J$,

$$0 \leq \sum_{j=1}^J K_{ij}(B_j^r(t) - B_j^r(s)) \leq C_i(t-s). \quad (8)$$

Recall that, for each j there is an i such that $K_{ij} = 1$, so if for this (i, j) , $L_j \doteq C_i$, then

$$0 \leq B_j^r(t) - B_j^r(s) \leq L_j(t-s), \quad (9)$$

i.e. B^r is Lipschitz continuous with Lipschitz constant not depending on r .

We now introduce the cost function. We will consider linear holding cost given through a fixed strictly positive J -dimensional vector h . The following two types of costs will be considered:

Infinite Horizon Discounted Cost. For a “discount factor” $\gamma \in (0, \infty)$, the infinite horizon discounted cost in the r -th network \mathcal{N}^r , associated with a control policy B^r , is defined as

$$J_D^r(B^r) = \int_0^\infty e^{-\gamma t} E[h \cdot \hat{Q}^r(t)] dt. \quad (10)$$

Long-Term Cost Per Unit Time. Fix $\zeta \in \mathbb{R}_+^J$ such that $\zeta > 0$. In the r -th network \mathcal{N}^r the long-term cost per unit time (or the ergodic cost) associated with a control policy B^r is defined as

$$J_E^r(B^r) = \limsup_{T \rightarrow \infty} E \left[\frac{1}{T} \int_0^T h \cdot \hat{Q}^r(t) dt + \frac{\zeta \cdot \hat{U}^r(T)}{T} \right]. \quad (11)$$

For a sequence of control policies $\{B^r\}$, the associated discounted [resp. ergodic] asymptotic cost is defined as

$$J_D(\{B^r\}_r) = \liminf_{r \rightarrow \infty} J_D^r(B^r), \quad [\text{resp.}] \quad J_E(\{B^r\}_r) = \liminf_{r \rightarrow \infty} J_E^r(B^r). \quad (12)$$

The infimum of asymptotic discounted [resp. ergodic] costs over all admissible sequences of control policies will be referred to as the *asymptotic value function* for the discounted [resp. ergodic] control problem and is given as

$$J_D^* = \inf_{\{B^r\}_r \in \mathcal{A}} J_D(\{B^r\}_r), \quad [\text{resp.}] \quad J_E^* = \inf_{\{B^r\}_r \in \mathcal{A}} J_E(\{B^r\}_r). \quad (13)$$

3 Equivalent workload formulations of Brownian control problems

The main results of this work will give a lower bound on the asymptotic discounted and ergodic control value functions in terms of value functions of certain control problems for Brownian motions [9, 11]. We present below the *Equivalent Workload Formulations* (EWF) of these control problems. We refer the reader to [11]

for a discussion on equivalence between this formulation and the Brownian control problems as formulated in Harrison [8]. We begin by introducing the notion of an effective cost function. With our formulation of the workload process as in (5) in mind, let $G = KM$, where we recall that M is the $J \times J$ diagonal matrix with entries $1/\beta_j$, and let $\mathscr{W} \doteq \mathbb{R}_+^J$. For each $w \in \mathscr{W}$, define the effective cost function as

$$\hat{h}(w) = \min\{h \cdot q : Gq = w, q \geq 0\}. \quad (14)$$

Note that, from the local traffic condition (Condition 1), the set on the right side is nonempty for every $w \in \mathscr{W}$. It is known that we can select a continuous minimizer in the above linear program (cf. [2]), i.e. there is a continuous map $q^* : \mathscr{W} \rightarrow \mathbb{R}_+^J$ such that

$$q^*(w) \in \arg \min_q \{h \cdot q : Gq = w, q \geq 0\}.$$

Let $\theta = M^{-1}\eta$, and let Σ denote the $J \times J$ matrix

$$\Sigma = \Sigma^u + \Sigma^v R, \quad (15)$$

where Σ^u is the $J \times J$ diagonal matrix with entries $\alpha_j^3(\sigma_j^u)^2$, Σ^v is the $J \times J$ diagonal matrix with entries $\beta_j^3(\sigma_j^v)^2$, and R is the diagonal matrix with entries ρ_j . The EWF and the associated controls and state processes are defined as follows.

Definition 4. Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}(t)\})$ be a filtered probability space which supports a J -dimensional $\tilde{\mathcal{F}}(t)$ -Brownian motion \tilde{X} with drift 0 and covariance matrix Σ . An I -dimensional $\{\tilde{\mathcal{F}}(t)\}$ -adapted process \tilde{U} on this space is called an admissible control for the EWF if the following hold \tilde{P} -a.s.:

- (i) $\tilde{W}(t) \doteq w + G\theta t + G\tilde{X}(t) + \tilde{U}(t) \geq 0$ for all $t \geq 0$, where $w = Gq$,
- (ii) $t \mapsto \tilde{U}(t)$ is nondecreasing and $\tilde{U}(0) \geq 0$.

Denote the class of all such admissible controls as $\tilde{\mathcal{A}}$.

The discounted cost for a control $\tilde{U} \in \tilde{\mathcal{A}}$ in the EWF is defined as

$$\tilde{J}_D(\tilde{U}) = \int_0^\infty e^{-\gamma t} \tilde{E}[\hat{h}(\tilde{W}(t))] dt, \quad (16)$$

where γ is as in the last section and \hat{h} is as introduced in (14).

Some modifications are needed in order to define the EWF for the ergodic control problem.

Definition 5. Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}(t)\})$ and \tilde{X} be as in Definition 4. An I -dimensional $\{\tilde{\mathcal{F}}(t)\}$ -adapted process \tilde{U} on this space is called an admissible control for the ergodic EWF if there is a $\{\tilde{\mathcal{F}}(t)\}$ -adapted \mathbb{R}_+^I valued process \tilde{W} such that the following hold \tilde{P} -a.s.:

- (i) $\tilde{W}(t) = \tilde{W}(0) + G\theta t + G\tilde{X}(t) + \tilde{U}(t)$ for all $t \geq 0$,
- (ii) $\tilde{U}(t)$ is nondecreasing and $\tilde{U}(0) = 0$,

- (iii) For all $t \geq 0$, $(\tilde{W}(t + \cdot), \tilde{U}(t + \cdot) - \tilde{U}(t))$ has the same distribution on $\mathcal{D}([0, \infty) : \mathbb{R}^{2I})$ as $(\tilde{W}(\cdot), \tilde{U}(\cdot))$.

Denote the class of all such admissible controls as $\tilde{\mathcal{A}}_E$.

The ergodic cost for a control $\tilde{U} \in \tilde{\mathcal{A}}_E$ in the ergodic Brownian control problem (BCP) is defined as

$$\tilde{J}_E(\tilde{U}) = \tilde{E}[\hat{h}(\tilde{W}(0))]. \quad (17)$$

Define the value functions

$$\tilde{J}_D^* = \inf_{\tilde{U} \in \tilde{\mathcal{A}}} \tilde{J}_D(\tilde{U}), \quad \tilde{J}_E^* = \inf_{\tilde{U} \in \tilde{\mathcal{A}}_E} \tilde{J}_E(\tilde{U}). \quad (18)$$

Obtaining explicit simple form solutions for the control problems in Definitions 4 and 5 is in general impossible. However, there is one important setting, given in the next theorem, where explicit solutions are available. Proof of the next theorem relies on well known minimality properties of the Skorohod map with normal reflections on the domain \mathbb{R}_+^I (cf. [7]) and is omitted due to space constraints. We begin by recalling the definition of this Skorohod map.

Definition 6. Let $\psi \in \mathcal{D}^I$ such that $\psi(0) \in \mathbb{R}_+^I$. The pair $(\varphi, \eta) \in \mathcal{D}^{2I}$ is said to solve the *Skorohod problem* for ψ (in \mathbb{R}_+^I , with normal reflection) if $\varphi = \psi + \eta$; $\varphi(t) \in \mathbb{R}_+^I$ for all $t \geq 0$; $\eta(0) = 0$; η is nondecreasing and $\int_{[0, \infty)} 1_{\{\varphi_i(t) > 0\}} d\eta_i(t) = 0$ for all $i \in \mathbb{N}_I$. We write $\varphi = \Gamma(\psi)$ and refer to Γ as the I -dimensional *Skorohod map*.

It is known that there is a unique solution to the above Skorohod problem for every $\psi \in \mathcal{D}^I$ with $\psi(0) \in \mathbb{R}_+^I$.

Let $\nu \doteq G\theta$. Also we denote by ι the identity map on $[0, \infty)$.

Theorem 1. Suppose that \hat{h} is monotonically nondecreasing, namely if $w_1, w_2 \in \mathbb{R}_+^I$ satisfy $w_1 \leq w_2$ then $\hat{h}(w_1) \leq \hat{h}(w_2)$. Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}(t)\})$ and \tilde{X} be as in Definition 4. Let

$$W^*(t) \doteq \Gamma(w + \nu t + G\tilde{X})(t) = w + \nu t + G\tilde{X}(t) + U^*(t), t \geq 0. \quad (19)$$

Then $U^* \in \tilde{\mathcal{A}}$ and

$$\tilde{J}_D^* = \tilde{J}_D(U^*) = \int_0^\infty e^{-\gamma t} \tilde{E}[\hat{h}(W^*(t))] dt.$$

Suppose in addition that $\nu < 0$. Then there is a unique stationary distribution π for the Markov process described by (19). Assume without loss of generality that the filtered probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}(t)\})$ supports an $\tilde{\mathcal{F}}_0$ -measurable \mathbb{R}_+^I -valued random variable $W'(0)$ with distribution π , and let $\{W'(t)\}$ be the stationary process defined as

$$W'(t) \doteq \Gamma(W'(0) + \nu t + G\tilde{X})(t) = W'(0) + \nu t + G\tilde{X}(t) + U'(t), t \geq 0. \quad (20)$$

Then $U' \in \tilde{\mathcal{A}}_E$ and

$$\tilde{J}_E^* = \tilde{J}_E(U^l) = \int_{\mathbb{R}_+^l} \hat{h}(w) \pi(dw).$$

4 Main Result

We can now present the main result of this work.

Theorem 2. *The following inequalities hold: $J_D^* \geq \tilde{J}_D^*$ and $J_E^* \geq \tilde{J}_E^*$.*

Remark 1. In [5], under quite general conditions on the matrix K and assuming that network primitives (namely the interarrival times and amounts of work) are exponentially distributed, explicit threshold form admissible control policies $\{B^{r,*}\}$ are constructed for which $J_D^r(B^{r,*}) \rightarrow \tilde{J}_D(U^*)$ and, under the additional condition that $\nu < 0$, $J_E^r(B^{r,*}) \rightarrow \tilde{J}_E(U^l)$. In view of Theorems 1 and 2, we then have that under the conditions of [5] and with the additional assumption that \hat{h} is nondecreasing, the sequence of control policies $\{B^{r,*}\}$ of [5] is asymptotically optimal for the discounted cost problem, and if also $\nu < 0$, this sequence is also asymptotically optimal for the ergodic control problem, namely we have the following

$$\tilde{J}_D^* \leq J_D^* \leq \liminf_{r \rightarrow \infty} J_D^r(B^{r,*}) = \tilde{J}_D^*,$$

and

$$\tilde{J}_E^* \leq J_E^* \leq \liminf_{r \rightarrow \infty} J_E^r(B^{r,*}) = \tilde{J}_E^*.$$

In particular, $J_D^* = \tilde{J}_D^*$ and $J_E^* = \tilde{J}_E^*$.

The rest of this work is devoted to the proof of Theorem 2.

5 Proofs

In Section 5.1 we prove the result for the discounted cost, namely the first inequality in Theorem 2 and in Section 5.2 we consider the ergodic cost, namely the second inequality in Theorem 2. Due to space constraints we will only provide a sketch for the second inequality. Detailed proof will be reported elsewhere.

5.1 Discounted Cost

We need to show that for every admissible sequence of control policies $\{B^r\}_r \in \mathcal{A}$, $J_D(\{B^r\}_r) \geq \tilde{J}_D^*$. Now fix such a sequence and assume that

$$J_D(\{B^r\}_r) = \liminf_{r \rightarrow \infty} J_D^r(B^r) < \infty,$$

since otherwise the inequality is immediate. The subsequence $\{r'\} \subset \{r\}$ along which the \liminf is achieved will be labeled again as $\{r\}$. With this relabeling, $J_D(\{B^r\}_r) = \lim_{r \rightarrow \infty} J_D^r(B^r)$.

5.1.1 Preliminary results

We begin by establishing some useful asymptotic results and moment bounds.

Lemma 1. 1. *The following central limit theorem holds:*

$$(\hat{A}^r, \hat{S}^r) \Rightarrow (A, S) \quad \text{as } r \rightarrow \infty,$$

where A and S are independent J -dimensional Brownian motions with drift 0 and covariances Σ^u and Σ^v , respectively.

2. *The following law of large numbers holds:*

$$(\bar{Q}^r, \bar{B}^r, \bar{A}^r, \bar{S}^r) \xrightarrow{P} (0, \rho\mathbf{1}, \alpha\mathbf{1}, \beta\mathbf{1}) \quad \text{as } r \rightarrow \infty.$$

3. *If $X = A + S(\rho\mathbf{1})$, then X is a Brownian motion with drift 0 and covariance Σ , and*

$$(\hat{A}^r, \hat{S}^r(\bar{B}^r), \hat{X}^r) \Rightarrow (A, S(\rho\mathbf{1}), X) \quad \text{as } r \rightarrow \infty.$$

Proof. The first statement is just the functional central limit theorem for renewal processes (see Theorem 14.6 in [1]), and the independence of A and S follows from the independence of $\{u_j^r(k)\}$ and $\{v_j^r(k)\}$. The last statement is immediate from the first two. It remains to prove the law of large numbers in 2.

From the first statement, it follows that

$$(\bar{A}^r, \bar{S}^r) = \left(\frac{\hat{A}^r}{r} + \alpha^r \mathbf{1}, \frac{\hat{S}^r}{r} + \beta^r \mathbf{1} \right) \xrightarrow{P} (\alpha\mathbf{1}, \beta\mathbf{1}). \quad (21)$$

By (9), if $s \leq t$ then for each $j \in \mathbb{N}_J$,

$$\bar{B}_j^r(t) - \bar{B}_j^r(s) = \frac{1}{r^2} (B_j^r(r^2 t) - B_j^r(r^2 s)) \leq L_j(t - s),$$

so $\{\bar{B}^r\}_r$ is tight in \mathcal{C}^J . Since

$$\bar{Q}^r(t) = \frac{q^r}{r^2} + \bar{A}^r(t) - \bar{S}^r(\bar{B}^r(t)), \quad (22)$$

it follows that $(\bar{Q}^r, \bar{B}^r, \bar{A}^r, \bar{S}^r)$ is tight in \mathcal{D}_+^{AJ} . Suppose we have a subsequence that converges weakly to some $(\bar{Q}, \bar{B}, \bar{A}, \bar{S})$. Since $J_D(\{B^r\}_r) < \infty$, by Fatou's lemma we have, taking limit along the subsequence,

$$\int_0^\infty e^{-\gamma t} E[h \cdot \bar{Q}(t)] dt \leq \liminf_{r \rightarrow \infty} \int_0^\infty e^{-\gamma t} E[h \cdot \bar{Q}^r(t)] dt \leq \liminf_{r \rightarrow \infty} \frac{J^r(B^r)}{r} = 0.$$

Since $h > 0$, we have that $\bar{Q} = 0$ a.s. From (21), $\bar{A} = \alpha t$ and $\bar{S} = \beta t$ a.s., and hence from (22), for all $t \geq 0$, a.s. $0 = \alpha t - \beta \bar{B}(t)$, from which it follows that $\bar{B} = \rho t$ a.s. \square

Lemma 2. *There is a $c \in (0, \infty)$ such that for all $j \in \mathbb{N}_J$, $r \geq 1$, and $t \geq 0$,*

$$E \left[\sup_{0 \leq s \leq t} \hat{A}_j^r(s)^2 \right] + E \left[\sup_{0 \leq s \leq t} \hat{S}_j^r(\bar{B}_j^r(s))^2 \right] \leq c(t+1).$$

Furthermore, the process $\hat{U}^r = K\hat{Y}^r$ satisfies $\limsup_{r \rightarrow \infty} E[\hat{U}_i^r(t)] < \infty$ for all $i \in \mathbb{N}_I$ and $t \geq 0$.

Proof. The first estimate is proved exactly as Lemma 3.5 of [4]. We now consider the second statement in the lemma. Note that $\hat{W}^r = G\hat{Q}^r$. Since $h > 0$, there is a $c_1 \in (0, \infty)$ such that for all $r \in \mathbb{N}$ and $t \geq 0$

$$\|\hat{W}^r(t)\| \leq \|G\| \|\hat{Q}^r(t)\| \leq c_1 \|G\| h \cdot \hat{Q}^r(t).$$

It follows that

$$\limsup_{r \rightarrow \infty} \int_0^\infty e^{-\gamma t} E \|\hat{W}^r(t)\| dt \leq c_1 \|G\| J_D(\{B^r\}_r) < \infty.$$

Now, from (7),

$$\hat{W}^r(t) = G\hat{q}^r + G\hat{X}^r(t) + G\hat{\theta}^r(t) + \hat{U}^r(t), \quad (23)$$

and so from the first part of this lemma, there is a $c_2 \in (0, \infty)$ such that for all $t \geq 0$ and $r \in \mathbb{N}$

$$\begin{aligned} E \|\hat{U}^r(t)\| &\leq E \|\hat{W}^r(t)\| + \|G\hat{q}^r\| + E \|G\hat{X}^r(t)\| + E \|G\hat{\theta}^r(t)\| \\ &\leq E \|\hat{W}^r(t)\| + \|G\hat{q}^r\| + c_2(t+1). \end{aligned}$$

For each $i \in \mathbb{N}_I$, \hat{U}_i^r is nondecreasing, and so

$$\int_0^\infty e^{-\gamma s} E[\hat{U}_i^r(s)] ds \geq \int_t^{t+1} e^{-\gamma s} E[\hat{U}_i^r(s)] ds \geq e^{-\gamma(t+1)} E[\hat{U}_i^r(t)].$$

Finally, we have that

$$\begin{aligned} \limsup_{r \rightarrow \infty} E[\hat{U}_i^r(t)] &\leq \limsup_{r \rightarrow \infty} e^{\gamma(t+1)} \int_0^\infty e^{-\gamma s} E[\hat{U}_i^r(s)] ds \\ &\leq \limsup_{r \rightarrow \infty} e^{\gamma(t+1)} \int_0^\infty e^{-\gamma s} E \|\hat{U}^r(s)\| ds \\ &\leq e^{\gamma(t+1)} \limsup_{r \rightarrow \infty} \int_0^\infty e^{-\gamma s} E \|\hat{W}^r(s)\| ds \\ &\quad + e^{\gamma(t+1)} \gamma^{-1} (\|Gq\| + c_2(\gamma^{-1} + 1)) < \infty. \quad \square \end{aligned}$$

Define the process \hat{H}^r by

$$\hat{H}^r(t) = (\hat{H}^{r,A}(t), \hat{H}^{r,S}(t)) = (\hat{A}^r(t), \hat{S}^r(\bar{B}^r(t))). \quad (24)$$

Lemma 3. For each r , define the \mathbb{R}_+ -valued process $V^r(t) = t + \sum_{i=1}^I \hat{U}_i^r(t)$, $t \geq 0$. Then V^r has a.s. continuous paths and is strictly increasing. For $t \geq 0$ define $\check{V}^r(t) = \inf\{s \geq 0 : V^r(s) > t\}$, and consider the ‘time-transformed’ processes

$$\check{W}^r = \hat{W}^r(\check{V}^r), \quad \check{X}^r = \hat{X}^r(\check{V}^r), \quad \check{U}^r = \hat{U}^r(\check{V}^r), \quad \check{H}^r = \hat{H}^r(\check{V}^r), \quad \check{\theta}^r = \hat{\theta}^r(\check{V}^r).$$

Then,

(i) The sequence $(\check{W}^r, \check{X}^r, \check{H}^r, \check{U}^r, \check{V}^r, \check{\theta}^r)$ is tight in $\mathcal{D}^{I+3J} \times \mathcal{C}_+^{I+1+J}$.

(ii) If \check{V} is a weak limit point of \check{V}^r on some probability space, then

$$\lim_{t \rightarrow \infty} \check{V}(t) = \infty \quad a.s. \quad (25)$$

Let $V(t) = \inf\{s \geq 0 : \check{V}(s) > t\}$, then a.s. \check{V} is continuous, V is right continuous, and both \check{V} and V are nondecreasing maps on $[0, \infty)$.

(iii) If $(A, S, \check{V}, \check{H})$ is a weak limit point of $(\hat{A}^r, \hat{S}^r, \check{V}^r, \check{H}^r)$ on some probability space, then

$$\check{H} = (A(\check{V}), S(\rho\check{V})) \quad a.s. \quad (26)$$

(iv) If $(\check{V}, \check{\theta})$ is a weak limit point of $(\check{V}^r, \check{\theta}^r)$ on some probability space, then $\check{\theta}(\cdot) = \theta\check{V}(\cdot)$.

Proof. Since \hat{U}^r is nondecreasing, V^r is strictly increasing. Since $\hat{U}^r(t) = rC - K\hat{B}^r(t)$, the continuity of $\hat{B}^r(t)$ gives the continuity of $V^r(t)$. Now,

$$V^r(t) = t + \mathbf{1} \cdot \hat{U}^r(t) = t + rt \sum_{i=1}^I C_i - \sum_{i=1}^I \sum_{j=1}^J K_{ij} \hat{B}_j^r(t),$$

and so if $s \leq t$, then by (8), $V^r(t) - V^r(s) \geq t - s$. Furthermore, for each $i \in \mathbb{N}_I$,

$$0 \leq \hat{U}_i^r(t) - \hat{U}_i^r(s) \leq (V^r(t) - t) - (V^r(s) - s) \leq V^r(t) - V^r(s).$$

We then have that

$$0 \leq \check{V}^r(t) - \check{V}^r(s) \leq t - s, \text{ and } 0 \leq \check{U}_i^r(t) - \check{U}_i^r(s) \leq t - s. \quad (27)$$

which gives the tightness of \check{V}^r and \check{U}^r . The tightness of \check{X}^r , and \check{H}^r now follows from Lemma 1 and the tightness of \check{W}^r follows from (23) on noting that

$$\check{W}^r(t) = G\check{q}^r + G\check{X}^r(t) + \check{\theta}^r(t) + \check{U}^r(t).$$

In order to prove the first statement in (ii) it suffices to show that $\limsup_{r \rightarrow \infty} P(\check{V}^r(t) < m) \rightarrow 0$ as $t \rightarrow \infty$ for any $m > 0$. To see this, note that for $t > m$,

$$\begin{aligned}
P(\check{V}^r(t) < m) &= P(V^r(m) > t) = P\left(m + \sum_{i=1}^I \hat{U}_i^r(m) > t\right) \\
&\leq \sum_{i=1}^I P\left(\hat{U}_i^r(m) > \frac{t-m}{I}\right) \leq \frac{I}{t-m} \sum_{i=1}^I E[\hat{U}_i^r(m)], \quad (28)
\end{aligned}$$

and hence by Lemma 2, as $t \rightarrow \infty$,

$$\limsup_{r \rightarrow \infty} P(\check{V}^r(t) < m) \leq \frac{I}{t-m} \sum_{i=1}^I \limsup_{r \rightarrow \infty} E[\hat{U}_i^r(m)] \rightarrow 0.$$

Consider now the remaining statements in (ii). Continuity of \check{V} is a consequence of the continuity of \check{V}^r for every r . Also, since \check{V}^r is nondecreasing for every r , so is \check{V} . It then follows from the definition of V that it is right continuous and nondecreasing. Finally (iii) is an immediate consequence of Lemma 1 and the definitions of \check{H}^r and \hat{H}^r . Part (iv) is immediate from Lemma 1(2) and Condition 3. \square

The following result is a key step in the proof of the first inequality in Theorem 2. Proof is given in Section 5.1.3.

Theorem 3. *Let $(\check{W}, \check{X}, \check{U}, \check{H}, \check{V}, A, S)$ be weak limit points of $(\check{W}^r, \check{X}^r, \check{U}^r, \check{H}^r, \check{V}^r, \hat{A}^r, \hat{S}^r)$ on some probability space $(\check{\Omega}, \check{\mathcal{F}}, \check{P})$, and let V be defined as in Lemma 3. Let*

$$W = \check{W}(V), \quad X = \check{X}(V), \quad U = \check{U}(V), \quad H = \check{H}(V).$$

Then, $H = (A, S(\rho\iota))$ a.s. Furthermore, there is a filtration $\{\check{\mathcal{F}}(t), t \geq 0\}$ on $(\check{\Omega}, \check{\mathcal{F}}, \check{P})$ to which (W, X, U) is adapted and such that X is an $\check{\mathcal{F}}(t)$ -Brownian motion with drift 0 and covariance Σ defined in (15) and, with $w = Gq$, a.s.

$$W(t) = w + G\theta t + GX(t) + U(t), \quad t \geq 0. \quad (29)$$

5.1.2 Proof of Theorem 2: Discounted Cost

In this section we prove the first inequality in Theorem 2. From Lemma 1 and 3, the sequence $(\check{W}^r, \check{X}^r, \check{U}^r, \check{H}^r, \check{V}^r, \hat{A}^r, \hat{S}^r)$ is tight. Suppose without loss of generality that the sequence converges in distribution to some $(\check{W}, \check{X}, \check{U}, \check{H}, \check{V}, A, S)$ given on some probability space $(\check{\Omega}, \check{\mathcal{F}}, \check{P})$ as in Theorem 3. By appealing to Skorohod representation theorem we can assume that the convergence holds a.s. Note that \check{V}^r is continuous and the inequality in (28) and Lemma 2 show that $\check{V}^r(t) \uparrow \infty$ as $t \rightarrow \infty$. Thus using Fubini's theorem and changing variables

$$\begin{aligned}
J_D^r(B^r) &= \int_0^\infty e^{-\gamma t} E[h \cdot \check{Q}^r(t)] dt = E \left[\int_0^\infty e^{-\gamma \check{V}^r(t)} h \cdot \check{Q}^r(t) d\check{V}^r(t) \right] \\
&\geq E \left[\int_0^\infty e^{-\gamma \check{V}^r(t)} \hat{h}(\check{W}^r(t)) d\check{V}^r(t) \right],
\end{aligned}$$

where $\check{Q}^r = \hat{Q}^r(\check{V}^r)$ and \hat{h} is the effective cost function defined in (14). Since for fixed $N \in \mathbb{N}$ the map $(x, y) \mapsto e^{-\gamma y}(\hat{h}(x) \wedge N)$ is continuous and bounded, it follows that, for any $s > 0$

$$\int_0^s e^{-\gamma \check{V}^r(t)} (\hat{h}(\check{W}^r(t)) \wedge N) d\check{V}^r(t) \rightarrow \int_0^s e^{-\gamma \check{V}(t)} (\hat{h}(\check{W}(t)) \wedge N) d\check{V}(t) \quad \text{a.s.}$$

as $r \rightarrow \infty$. Thus, a.s.

$$\begin{aligned} \liminf_{r \rightarrow \infty} \int_0^\infty e^{-\gamma \check{V}^r(t)} \hat{h}(\check{W}^r(t)) d\check{V}^r(t) &\geq \liminf_{r \rightarrow \infty} \int_0^s e^{-\gamma \check{V}^r(t)} (\hat{h}(\check{W}^r(t)) \wedge N) d\check{V}^r(t) \\ &= \int_0^s e^{-\gamma \check{V}(t)} (\hat{h}(\check{W}(t)) \wedge N) d\check{V}(t). \end{aligned}$$

Letting $N, s \rightarrow \infty$, we obtain

$$\liminf_{r \rightarrow \infty} \int_0^\infty e^{-\gamma \check{V}^r(t)} \hat{h}(\check{W}^r(t)) d\check{V}^r(t) \geq \int_0^\infty e^{-\gamma \check{V}(t)} \hat{h}(\check{W}(t)) d\check{V}(t) \quad \text{a.s.}$$

Finally, using Fatou's lemma,

$$\begin{aligned} J_D(\{B^r\}_r) &= \liminf_{r \rightarrow \infty} J_D^r(B^r) \geq \liminf_{r \rightarrow \infty} E \left[\int_0^\infty e^{-\gamma \check{V}^r(t)} \hat{h}(\check{W}^r(t)) d\check{V}^r(t) \right] \\ &\geq E \left[\liminf_{r \rightarrow \infty} \int_0^\infty e^{-\gamma \check{V}^r(t)} \hat{h}(\check{W}^r(t)) d\check{V}^r(t) \right] \\ &\geq E \left[\int_0^\infty e^{-\gamma \check{V}(t)} \hat{h}(\check{W}(t)) d\check{V}(t) \right] \\ &= E \left[\int_0^\infty e^{-\gamma t} \hat{h}(W(t)) dt \right] = \tilde{J}_D(U) \geq \tilde{J}_D^*. \end{aligned}$$

where W and U are as in Theorem 3, the second equality on the last line is a consequence of Theorem 3, and the final inequality uses the observation that $U \in \mathcal{A}$. \square

5.1.3 Proof of Theorem 3

In this section we provide the proof of Theorem 3. Since B^r is admissible, \hat{U}^r is $\mathcal{G}^r(t)$ -adapted, where $\{\mathcal{G}^r(t), t \geq 0\}$ is as in Definition 3. Thus

$$\{\check{V}^r(s) \leq t\} = \{V^r(t) \geq s\} = \left\{ \sum_{i=1}^I \hat{U}^r(t) \geq s - t \right\} \in \mathcal{G}^r(t).$$

Thus for each fixed s , $\check{V}^r(s)$ is a $\{\mathcal{G}^r(t)\}$ -stopping time and consequently $\mathcal{H}^r(t) = \mathcal{G}^r(\check{V}^r(t))$, $t \geq 0$, defines a filtration. With τ^r as in Definition 3, let $\sigma^r(t) = \tau^r(\check{V}^r(t))$ for $r, t \geq 0$. We then have the following:

Lemma 4. (i) $\sigma^r(t)$ is a stopping time with respect to the multiparameter filtration

$$\{\mathcal{F}^r(m, n), (m, n) \in \mathbb{N}^{2J}\},$$

$$(ii) \mathcal{H}^r(t) = \mathcal{F}^r(\sigma^r(t)),$$

(iii) $\check{U}^r, \check{H}^r, \check{V}^r$ are $\mathcal{H}^r(t)$ -adapted.

Proof. For $r, t \geq 0$ and each $k \in \mathbb{N}$, let $\check{V}_k^r(t) \doteq \lfloor \check{V}^r(t)2^k + 1 \rfloor / 2^k$. Then $\check{V}_k^r(t)$ is a $\{\mathcal{G}^r(s), s \geq 0\}$ -stopping time that takes values in $\{j2^{-k}, j \geq 0\}$. Also $\check{V}_k^r(t) \downarrow \check{V}^r(t)$ as $k \rightarrow \infty$. Then $\sigma^r(t) = \inf_{k \in \mathbb{N}} \tau^r(\check{V}_k^r(t))$. For each $(m, n) \in \mathbb{N}^{2J}$,

$$\{\tau^r(\check{V}_k^r(t)) \leq (m, n)\} = \bigcup_{j \geq 0} \left\{ \tau^r(j2^{-k}) \leq (m, n) \right\} \cap \left\{ \check{V}_k^r(t) = j2^{-k} \right\} \in \mathcal{F}^r(m, n)$$

by the definition of the stopped σ -field $\mathcal{G}^r(t) = \mathcal{F}^r(\tau^r(t))$. Thus $\{\tau^r(\check{V}_k^r(t)), k \in \mathbb{N}\}$ is a sequence of $\mathcal{F}^r(m, n)$ -stopping times, and therefore $\sigma^r(t)$ is an $\mathcal{F}^r(m, n)$ -stopping time as well. This proves (i).

Next, it is easy to check that, for every k , $\mathcal{F}^r(\tau^r(\check{V}_k^r(t))) = \mathcal{G}^r(\check{V}_k^r(t))$. Since $\mathcal{F}^r(m, n)$ is indexed by a discrete set, $\tau^r(t)$ is right-continuous in t , and $\check{V}_k^r(t) \downarrow \check{V}^r(t)$, the left hand side above goes to $\mathcal{F}^r(\sigma^r(t))$ as $k \rightarrow \infty$ and the right hand side goes to $\mathcal{H}^r(t)$ as $k \rightarrow \infty$. This proves (ii).

By Definition 3, \hat{U}^r is $\mathcal{G}^r(t)$ -adapted, and so \check{U}^r is adapted to $\mathcal{G}^r(\check{V}^r(t)) = \mathcal{H}^r(t)$. Since $\tau^r(t)$ is a $\mathcal{F}^r(m, n)$ stopping time, it is $\mathcal{G}^r(t) = \mathcal{F}^r(\tau^r(t))$ -measurable and so \check{H}^r is $\mathcal{G}^r(t)$ -adapted. Thus \check{H}^r is $\mathcal{H}^r(t)$ -adapted. Also since $\check{V}^r(t)$ is a $\{\mathcal{G}^r(s)\}$ -stopping time, \check{V}^r is $\mathcal{H}^r(t)$ -adapted. This proves (iii) and completes the proof of the lemma. \square

We will need the following estimate. For a proof we refer to [4, Lemma 4.2].

Lemma 5. For all $t \geq 0$ and $k \in \mathbb{N}$,

$$\sup_{r \geq 0} \max_{j \in \mathbb{N}_J} \left(E[A_j^r(t)^k] + E[S_j^r(t)^k] \right) < \infty.$$

For $r \in \mathbb{N}$ and $(m, n) = (m_1, \dots, m_J, n_1, \dots, n_J) \in \mathbb{N}^{2J}$, define

$$\xi_j^r(m, n) = \xi_j^r(m_j) = \frac{1}{r} \sum_{k=1}^{m_j} (1 - \alpha_j^r u_j^r(k)),$$

and

$$\eta_j^r(m, n) = \eta_j^r(n_j) = \frac{1}{r} \sum_{k=1}^{n_j} (1 - \beta_j^r v_j^r(k))$$

for $j \in \mathbb{N}_J$. Then ξ_j^r and η_j^r are $\mathcal{F}^r(m, n)$ -martingales with quadratic variations

$$\langle \xi_j^r \rangle(m, n) = r^{-2} m_j (\alpha_j^r \sigma_j^{u, r})^2, \quad \langle \eta_j^r \rangle(m, n) = r^{-2} n_j (\beta_j^r \sigma_j^{v, r})^2,$$

$$\langle \xi_{j_1}^r, \xi_{j_2}^r \rangle(m, n) = \langle \eta_{j_1}^r, \eta_{j_2}^r \rangle(m, n) = \langle \xi_{j_3}^r, \eta_{j_4}^r \rangle(m, n) = 0,$$

for all $j, j_1, j_2, j_3, j_4 \in \mathbb{N}_J$ with $j_1 \neq j_2$. We will denote $\xi^r(m, n) = (\xi_1^r(m, n), \dots, \xi_J^r(m, n))$ and $\eta^r(m, n) = (\eta_1^r(m, n), \dots, \eta_J^r(m, n))$.

Lemma 6. For $r, t \geq 0$,

$$N^r(t) \doteq (N^{r,A}(t), N^{r,S}(t)) = (\xi^r(\sigma^r(t)), \eta^r(\sigma^r(t))).$$

is an $\mathcal{H}^r(t)$ -martingale.

Proof. This follows from a multiparameter version of the optional sampling theorem. For details we refer to [4, pages 1992-93]. \square

The following lemma shows that the martingale in Lemma 6 is close to \check{H}^r as r increases.

Lemma 7. For every $T < \infty$, $\sup_{0 \leq t \leq T} \|\check{H}^r(t) - N^r(t)\|_1 \xrightarrow{P} 0$ as $r \rightarrow \infty$.

Proof. For each $j \in \mathbb{N}_J$ and $t \geq 0$,

$$\begin{aligned} |\check{H}_j^{r,A}(t) - N_j^{r,A}(t)| &= |\hat{A}_j^r(\check{V}^r(t)) - \xi_j^r(\sigma^r(t))| \\ &= \left| \frac{1}{r} A_j^r(r^2 \check{V}^r(t)) - r \check{V}^r(t) \alpha_j^r - \xi_j^r(A_j^r(r^2 \check{V}^r(t)) + 1) \right| \\ &= \left| \frac{1}{r} A_j^r(r^2 \check{V}^r(t)) - r \check{V}^r(t) \alpha_j^r - \frac{1}{r} \sum_{k=1}^{A_j^r(r^2 \check{V}^r(t))+1} (1 - \alpha_j^r u_j^r(k)) \right| \\ &= \frac{\alpha_j^r}{r} \left| u_j^r(A_j^r(r^2 \check{V}^r(t)) + 1) + \sum_{k=1}^{A_j^r(r^2 \check{V}^r(t))} u_j^r(k) - r^2 \check{V}^r(t) - \frac{1}{\alpha_j^r} \right|. \end{aligned}$$

Since A_j^r is nondecreasing,

$$\sup_{0 \leq s \leq t} |\check{H}_j^{r,A}(s) - N_j^{r,A}(s)| \leq \frac{1}{r} \left(1 + \alpha_j^r \max_{k \leq A_j^r(r^2 t)+1} \left| u_j^r(k) - \frac{1}{\alpha_j^r} \right| \right). \quad (30)$$

A similar argument also gives

$$\sup_{0 \leq s \leq t} |\check{H}_j^{r,S}(s) - N_j^{r,S}(s)| \leq \frac{1}{r} \left(1 + \beta_j^r \max_{k \leq S_j^r(r^2 t)+1} \left| v_j^r(k) - \frac{1}{\beta_j^r} \right| \right). \quad (31)$$

Using the fact that $\{u_j^r(k) - 1/\alpha_j^r\}_{k \in \mathbb{N}}$ and $\{v_j^r(k) - 1/\beta_j^r\}_{k \in \mathbb{N}}$ are i.i.d. sequences of mean zero random variables and the uniform integrability property in (2), we have that for every $c > 0$ and $j \in \mathbb{N}_J$

$$\max_{1 \leq k \leq cr^2} |u_j^r(k) - 1/\alpha_j^r| \xrightarrow{P} 0 \quad \text{and} \quad \max_{1 \leq k \leq cr^2} |v_j^r(k) - 1/\beta_j^r| \xrightarrow{P} 0.$$

Also, by Lemma 1, $r^{-2} A_j^r(r^2 t) \xrightarrow{P} \alpha t$ and $r^{-2} S_j^r(r^2 t) \xrightarrow{P} \beta t$, uniformly on compacts. This shows that the right hand sides of (30) and (31) go to zero in probability as $r \rightarrow \infty$. \square

Now we return to the proof of Theorem 3. The first statement in the theorem is immediate from the definitions of various processes and properties of weak convergence. Also, from the definitions of \hat{H}^r, \hat{X}^r and \hat{W}^r in (24), (6), and (23), respectively,

$$\check{X}^r(t) = \hat{X}^r(\check{V}^r(t)) = \check{H}^{r,A}(t) - \check{H}^{r,S}(t),$$

and

$$\check{W}^r(t) = \hat{W}^r(\check{V}^r(t)) = G^r \check{q}^r + G^r \check{X}^r(t) + r \check{V}^r(t) K(\rho^r - \rho) + \check{U}^r(t).$$

Taking the limit as $r \rightarrow \infty$ along the weakly convergent subsequence of the processes in the statement of Theorem 3,

$$\begin{aligned} X(t) &= H^A(t) - H^S(t) = A(t) - S(\rho t), \\ W(t) &= Gq + G\theta t + GX(t) + U(t), \end{aligned}$$

where (H, X, W, U) are as in the statement of Theorem 3. This proves the identity in (29). Finally we prove the adaptedness statement and the Brownian motion property in Theorem 3. Define

$$\check{\mathcal{F}}(t) = \bigcap_{n \geq 1} \sigma \left\{ (\check{H}(s), \check{U}(s), \check{V}(s)), s \leq t + \frac{1}{n} \right\},$$

so that $\{\check{\mathcal{F}}(t), t \geq 0\}$ is a right-continuous filtration. By construction, for all $s, t \geq 0$,

$$\{V(s) < t\} = \{\check{V}(t) > s\} \in \check{\mathcal{F}}(t),$$

and so by right-continuity, $V(s)$ is a stopping time with respect to $\{\check{\mathcal{F}}(t), t \geq 0\}$ for all $s \geq 0$. Since V is nondecreasing, the stopped σ -fields $\check{\mathcal{F}}(t) = \check{\mathcal{F}}(V(t))$ form a filtration. Also by construction, \check{W}, \check{X} , and \check{U} are all $\check{\mathcal{F}}(t)$ -adapted, and thus (W, X, U) are $\check{\mathcal{F}}(t)$ -adapted.

To complete the proof of Theorem 3, it remains to show that X is an $\check{\mathcal{F}}(t)$ -Brownian motion with drift 0 and covariance Σ . For this, if \mathcal{L} is the differential operator defined as

$$\mathcal{L}f(x) = \frac{1}{2} \sum_{j=1}^J \alpha_j (\alpha_j \sigma_j^{\#})^2 \frac{\partial^2 f}{\partial x_j^2}(x) + \frac{1}{2} \sum_{j=1}^J \beta_j (\beta_j \sigma_j^{\vee})^2 \rho_j \frac{\partial^2 f}{\partial x_{j+J}^2}(x),$$

then it suffices to show that, for each $f \in \mathcal{C}_b^\infty(\mathbb{R}^{2J})$,

$$f(H(t)) - \int_0^t \mathcal{L}f(H(s)) ds \text{ is an } \check{\mathcal{F}}(t)\text{-martingale.} \quad (32)$$

We begin with the following lemma.

Lemma 8. *Suppose that for $0 \leq s \leq t$, $f \in \mathcal{C}_b^\infty(\mathbb{R}^{2J})$, and any bounded continuous function $g : \mathcal{C}([0, s] : \mathbb{R}^{2J} \times \mathbb{R}_+^{I+1}) \rightarrow \mathbb{R}$,*

$$E \left[g \left((\check{H}, \check{U}, \check{V}) \Big|_{[0,s]} \right) \left(f(\check{H}(t)) - f(\check{H}(s)) - \int_s^t \mathcal{L}f(\check{H}(u)) d\check{V}(u) \right) \right] = 0, \quad (33)$$

then (32) is satisfied.

Proof. Fix $f \in \mathcal{C}_b^\infty(\mathbb{R}^{2J})$. Let $Y(t) = f(\check{H}(t)) - \int_0^t \mathcal{L}f(\check{H}(u)) d\check{V}(u)$, $t \geq 0$. Then from (33) Y is an $\check{\mathcal{F}}(t)$ -martingale. Also note that $Y(V(t)) = f(H(t)) - \int_0^t \mathcal{L}f(H(s)) ds$, which is same as the expression in (32). Thus it suffices to show that $Y(t)$ and $V(t)$ satisfy hypotheses for the optional sampling theorem. Recall that $V(t)$ is a finite-a.s. $\{\check{\mathcal{F}}(s), s \geq 0\}$ -stopping time. Thus if for each $t \geq 0$, $E|Y(V(t))| < \infty$ and $E(|Y(T)|\mathbf{1}_{\{V(T) > T\}}) \rightarrow 0$ as $T \rightarrow \infty$, then we are done (see, for instance, Theorem 2.2.13 of [6]).

Since $f \in \mathcal{C}_b^\infty(\mathbb{R}^{2J})$, there is a $c \in (0, \infty)$ so that $|f| \vee |\mathcal{L}f| \leq c$, and so $|Y(s)| \leq c(1 + \check{V}(s))$ for each $s \geq 0$. Thus for all t $E|Y(V(t))| \leq c(1 + t) < \infty$, and using Lemma 3(ii), as $T \rightarrow \infty$,

$$E|Y(T)|\mathbf{1}_{\{V(T) > T\}} \leq E|Y(T)|\mathbf{1}_{\{\check{V}(T) < t\}} \leq c(1 + t)P(\check{V}(T) < t) \rightarrow 0.$$

This verifies the conditions for the optional sampling theorem and consequently completes the proof. \square

In order to finish the proof of Theorem 3 we will now show that (33) holds for all f, g as in Lemma 8. From Lemma 7, with various processes as in the statement of Theorem 3, by possibly taking a subsequence, as $r \rightarrow \infty$,

$$(\check{H}^r, N^r, \check{V}^r, \check{U}^r) \Rightarrow (\check{H}, \check{H}, \check{V}, \check{U}). \quad (34)$$

Recall from Lemma 4 that for each r , $(\check{H}^r, \check{U}^r, \check{V}^r)|_{[0,s]}$ is $\mathcal{H}^r(s)$ -measurable and using the continuous mapping theorem and dominated convergence theorem, it then suffices to show that for any $s \leq t$,

$$\limsup_{r \rightarrow \infty} \left| E \left[f(N^r(t)) - f(N^r(s)) - \int_s^t \mathcal{L}f(N^r(u)) d\check{V}^r(u) \Big| \mathcal{H}^r(s) \right] \right| = 0. \quad (35)$$

Partition the interval $[s, t]$ into the times $s = t_0^r < t_1^r < \dots < t_{r^2}^r \leq t$, where

$$t_i^r = s + \frac{i}{r^2}(t - s), \quad \text{for } i = 0, 1, \dots, r^2.$$

Then define the quantity

$$\begin{aligned} \Psi^r(s, t) = & \frac{1}{2} \sum_{j=1}^J \sum_{i=0}^{r^2} \left[\psi_j^{\mu, r} \frac{\partial^2 f}{\partial x_j^2}(N^r(t_i^r)) (\bar{A}_j^r(\check{V}^r(t_{i+1}^r)) - \bar{A}_j^r(\check{V}^r(t_i^r))) \right. \\ & \left. + \psi_j^{\nu, r} \frac{\partial^2 f}{\partial x_{j+J}^2}(N^r(t_i^r)) (\bar{S}_j^r(\bar{B}_j^r(\check{V}^r(t_{i+1}^r))) - \bar{S}_j^r(\bar{B}_j^r(\check{V}^r(t_i^r)))) \right], \end{aligned}$$

where $\Psi_j^{u,r} = (\alpha_j^r \sigma_j^{u,r})^2$ and $\Psi_j^{v,r} = (\beta_j^r \sigma_j^{v,r})^2$. Using the fact that $\check{V}^r(t) \leq t$, there are $c_1, c_2 \in (0, \infty)$ such that for $r \geq 1$ and $s \leq t$

$$E|\Psi^r(s, t)|^2 \leq c_1 \max_{j \in \mathbb{N}_J} (E[\bar{A}_j^r(\check{V}^r(t))^2] + E[\bar{S}_j^r(\bar{B}_j^r(\check{V}^r(t)))^2]) \leq c_2(t+1).$$

In particular, we have that $\{\Psi^r(s, t)\}_r$ is uniformly integrable. By Lemma 1 and (34),

$$(\bar{A}_j^r(\check{V}^r), \bar{S}_j^r(\bar{B}_j^r(\check{V}^r))) \Rightarrow (\alpha_j \check{V}, \beta_j \rho_j \check{V})$$

in \mathcal{D}_+^1 . It then follows that

$$\limsup_{r \rightarrow \infty} E \left| \Psi^r(s, t) - \int_s^t \mathcal{L}f(N^r(u)) d\check{V}^r(u) \right| = 0. \quad (36)$$

Now, by Taylor's theorem, for any $s \leq u < v \leq t$ write

$$\begin{aligned} f(N^r(v)) - f(N^r(u)) &= \sum_{j=1}^{2J} \frac{\partial f}{\partial x_j}(N^r(u))(N_j^r(v) - N_j^r(u)) \\ &\quad + \frac{1}{2} \sum_{j,k=1}^{2J} \frac{\partial^2 f}{\partial x_j \partial x_k}(L)(N_j^r(v) - N_j^r(u))(N_k^r(v) - N_k^r(u)), \end{aligned}$$

where L lies on the line segment between $N^r(v)$ and $N^r(u)$. From Lemma 6 N^r is an $\mathcal{H}^r(t)$ -martingale, and so

$$\begin{aligned} &E[f(N^r(v)) - f(N^r(u)) | \mathcal{H}^r(s)] \\ &= E \left[\frac{1}{2} \sum_{j,k=1}^{2J} \frac{\partial^2 f}{\partial x_j \partial x_k}(L)(N_j^r(v) - N_j^r(u))(N_k^r(v) - N_k^r(u)) \middle| \mathcal{H}^r(s) \right]. \end{aligned}$$

By partitioning $[s, t]$ as before and expanding $f(N^r(t)) - f(N^r(s))$ as a telescoping sum, we then get

$$\begin{aligned} &E[f(N^r(t)) - f(N^r(s)) - \Psi^r(s, t) | \mathcal{H}^r(s)] \\ &= \frac{1}{2} \sum_{j=1}^{2J} \sum_{i=0}^{r^2} E \left[\left(\frac{\partial^2 f}{\partial x_j^2}(L_i) - \frac{\partial^2 f}{\partial x_j^2}(N^r(t_i^r)) \right) (N_j^r(t_{i+1}^r) - N_j^r(t_i^r))^2 \middle| \mathcal{H}^r(s) \right], \quad (37) \end{aligned}$$

where L_i lies on the line segment between $N^r(t_{i+1}^r)$ and $N^r(t_i^r)$ for each i . Here we also used the following equalities, which can easily be checked from the definition of N^r : for $s \leq u < v$,

$$\begin{aligned}
& E[(N_j^r(v) - N_j^r(u))^2 | \mathcal{H}^r(s)] \\
& \quad = \psi_j^{u,r} E[\bar{A}_j^r(\check{V}^r(v)) - \bar{A}_j^r(\check{V}^r(u)) | \mathcal{H}^r(s)] \quad \text{for } 1 \leq j \leq J, \\
& E[(N_j^r(v) - N_j^r(u))^2 | \mathcal{H}^r(s)] \\
& \quad = \psi_j^{v,r} E[\bar{S}_j^r(\bar{B}_j^r(\check{V}^r(v))) - \bar{S}_j^r(\bar{B}_j^r(\check{V}^r(u))) | \mathcal{H}^r(s)] \quad \text{for } J+1 \leq j \leq 2J, \\
& E[(N_j^r(v) - N_j^r(u))(N_k^r(v) - N_k^r(u)) | \mathcal{H}^r(s)] = 0 \quad \text{for } j \neq k.
\end{aligned}$$

The following lemma will allow us to finish the proof of (35).

Lemma 9. *The sequence $\{r^2(N_j^r(t_{i+1}^r) - N_j^r(t_i^r))^2\}_{j \in \mathbb{N}_J, i \leq r^2, r \geq 0}$ is uniformly integrable conditional on $\mathcal{H}^r(s)$, namely given $\varepsilon > 0$, there is an $M < \infty$ such that for all $r \in \mathbb{N}$, $j \in \mathbb{N}_J$, and $i = 0, 1, \dots, r^2$, a.s.*

$$E \left[r^2(N_j^r(t_{i+1}^r) - N_j^r(t_i^r))^2 \mathbf{1}_{\{r^2(N_j^r(t_{i+1}^r) - N_j^r(t_i^r))^2 > M\}} \middle| \mathcal{H}^r(s) \right] \leq \varepsilon.$$

Proof. Let $u_j^r(A_j^r(r^2\check{V}^r(t_{i+1}^r)) + 1) \mathbf{1}_{\{A_j^r(r^2\check{V}^r(t_{i+1}^r)) - A_j^r(r^2\check{V}^r(t_i^r)) \geq 1\}} \doteq \zeta_{j,i}^r$. Using the definition of N^r , if $j \leq J$,

$$\begin{aligned}
& r(N_j^r(t_{i+1}^r) - N_j^r(t_i^r)) \\
& = \sum_{k=1}^{A_j^r(r^2\check{V}^r(t_{i+1}^r)) + 1} (1 - \alpha_j^r u_j^r(k)) - \sum_{k=1}^{A_j^r(r^2\check{V}^r(t_i^r)) + 1} (1 - \alpha_j^r u_j^r(k)) \\
& = A_j^r(r^2\check{V}^r(t_{i+1}^r)) - A_j^r(r^2\check{V}^r(t_i^r)) \\
& \quad - \alpha_j^r \left(\sum_{k=1}^{A_j^r(r^2\check{V}^r(t_{i+1}^r)) + 1} u_j^r(k) - \sum_{k=1}^{A_j^r(r^2\check{V}^r(t_i^r)) + 1} u_j^r(k) \right) \\
& \leq A_j^r(r^2\check{V}^r(t_{i+1}^r)) - A_j^r(r^2\check{V}^r(t_i^r)) + \alpha_j^r \zeta_{j,i}^r + \alpha_j^r (r^2\check{V}^r(t_{i+1}^r) - r^2\check{V}^r(t_i^r)) \\
& \leq A_j^r(r^2\check{V}^r(t_{i+1}^r)) - A_j^r(r^2\check{V}^r(t_i^r)) + \alpha_j^r \zeta_{j,i}^r + \alpha_j^r (t - s),
\end{aligned}$$

and so

$$\begin{aligned}
r^2(N_j^r(t_{i+1}^r) - N_j^r(t_i^r))^2 & \leq 4(A_j^r(r^2\check{V}^r(t_{i+1}^r)) - A_j^r(r^2\check{V}^r(t_i^r)))^2 \\
& \quad + 4(\alpha_j^r)^2 (\zeta_{j,i}^r)^2 + 4(\alpha_j^r)^2 (t - s)^2.
\end{aligned} \tag{38}$$

Since a similar inequality holds when $J+1 \leq j \leq 2J$, it suffices to prove the conditional uniform integrability of $\{(A_j^r(r^2\check{V}^r(t_{i+1}^r)) - A_j^r(r^2\check{V}^r(t_i^r)))^2\}_{j,i,r}$ and $\{(\zeta_{j,i}^r)^2\}_{j,i,r}$.

For the first, we have by a similar argument as in [4, page 2000] that, a.s., for any $c > 0$,

$$\begin{aligned}
& P(A_j^r(r^2\check{V}^r(t_{i+1}^r)) - A_j^r(r^2\check{V}^r(t_i^r)) > c | \mathcal{H}^r(s)) \\
& \leq E[P(A_j^r(t - s) + 1 > c | \mathcal{H}^r(t_i^r)) | \mathcal{H}^r(s)] \\
& \leq P(A_j^r(t - s) + 1 > c).
\end{aligned}$$

In particular,

$$\begin{aligned} E \left[\left(A_j^r(r^2\check{V}^r(t_{i+1}^r)) - A_j^r(r^2\check{V}^r(t_i^r)) \right)^2 \mathbf{1}_{\{(A_j^r(r^2\check{V}^r(t_{i+1}^r)) - A_j^r(r^2\check{V}^r(t_i^r)))^2 > c\}} \right] \Big| \mathcal{H}^r(s) \Big] \\ \leq E \left[(A_j^r(t-s) + 1)^2 \mathbf{1}_{\{(A_j^r(t-s) + 1)^2 > c\}} \right], \end{aligned}$$

and Lemma 5 then gives the required conditional uniform integrability.

For $c > 0$, let

$$\mu_j^r(c) = E \left[u_j^r(1)^2 \mathbf{1}_{\{u_j^r(1)^2 > c\}} \right],$$

and define the process

$$M_j^r(m, n) = M_j^r(m_j) = \sum_{k=1}^{m_j} u_j^r(k)^2 \mathbf{1}_{\{u_j^r(k)^2 > c\}} - m_j \mu_j^r(c).$$

Let $\mathcal{X} = \mathbf{1}_{\{A_j^r(r^2\check{V}^r(t_{i+1}^r)) - A_j^r(r^2\check{V}^r(t_i^r)) \geq 1\}}$. Since $\{u_j^r(k)\}_{k \in \mathbb{N}}$ are i.i.d., $\{M_j^r(m, n)\}$ is an $\{\mathcal{F}^r(m, n)\}$ -martingale, and therefore

$$\begin{aligned} & (\zeta_{j,i}^r)^2 \mathbf{1}_{\{(\zeta_{j,i}^r)^2 > c\}} \\ &= \mathcal{X} \left(\sum_{k=1}^{A_j^r(r^2\check{V}^r(t_{i+1}^r)) + 1} u_j^r(k)^2 \mathbf{1}_{\{u_j^r(k)^2 > c\}} - \sum_{k=1}^{A_j^r(r^2\check{V}^r(t_{i+1}^r))} u_j^r(k)^2 \mathbf{1}_{\{u_j^r(k)^2 > c\}} \right) \\ &\leq \sum_{k=1}^{A_j^r(r^2\check{V}^r(t_{i+1}^r)) + 1} u_j^r(k)^2 \mathbf{1}_{\{u_j^r(k)^2 > c\}} - \sum_{k=1}^{A_j^r(r^2\check{V}^r(t_i^r)) + 1} u_j^r(k)^2 \mathbf{1}_{\{u_j^r(k)^2 > c\}} \\ &= M_j^r(A_j^r(r^2\check{V}^r(t_{i+1}^r)) + 1) - M_j^r(A_j^r(r^2\check{V}^r(t_i^r)) + 1) \\ &\quad + ((A_j^r(r^2\check{V}^r(t_{i+1}^r)) - A_j^r(r^2\check{V}^r(t_i^r))) \mu_j^r(c). \end{aligned}$$

Then by optional sampling,

$$E \left[(\zeta_{j,i}^r)^2 \mathbf{1}_{\{(\zeta_{j,i}^r)^2 > c\}} \Big| \mathcal{H}^r(s) \right] \leq 2\mu_j^r(c) \sup_{r \geq 0} E[A_j^r(t) + 1].$$

By Lemma 5, $\sup_{r \geq 0} E[A_j^r(t) + 1] < \infty$ and by the assumption in (2) we have that $\sup_{r \geq 0} \mu_j^r(c) \rightarrow 0$ as $c \rightarrow \infty$. This establishes the desired uniform integrability and completes the proof. \square

Now we complete the proof of (35). It follows from the previous lemma that for each $\varepsilon > 0$, there is an $M = M(\varepsilon)$ such that a.s.

$$\sup_{r \geq 0} \sup_{i \in \{0, \dots, r^2\}} E \left[r^2 \|N^r(t_{i+1}^r) - N^r(t_i^r)\|_1^2 \mathbf{1}_{\{r^2 \|N^r(t_{i+1}^r) - N^r(t_i^r)\|_1^2 > M\}} \Big| \mathcal{H}^r(s) \right] < \varepsilon.$$

Since the second derivatives of f are bounded, we have from (37) that there is a $c < \infty$ so that for every $\varepsilon > 0$,

$$\begin{aligned} & |E[f(N^r(t)) - f(N^r(s)) - \Psi^r(s, t) | \mathcal{H}^r(s)]| \\ & \leq 2cJ\varepsilon + \frac{M(\varepsilon)}{2} \sum_{j=1}^{2J} \sup_{\|x-y\|_1^2 \leq \frac{M}{j^2}} \left| \frac{\partial^2 f}{\partial x_j^2}(x) - \frac{\partial^2 f}{\partial x_j^2}(y) \right|. \end{aligned}$$

Since $f \in \mathcal{C}_b^\infty(\mathbb{R}^{2J})$, letting $r \rightarrow \infty$ followed by $\varepsilon \rightarrow 0$ gives us that

$$\limsup_{r \rightarrow \infty} |E[f(N^r(t)) - f(N^r(s)) - \Psi^r(s, t) | \mathcal{H}^r(s)]| = 0. \quad (39)$$

Together with (36) this gives (35) and completes the proof of Theorem 3. \square

5.2 Ergodic Cost

In this section we sketch the proof of the second part of Theorem 2, namely the inequality $J_E^* \geq \tilde{J}_E^*$. As for the discounted case, it suffices to show that for every admissible sequence of control policies $\{B^r\}_r \in \mathcal{A}$, $J_E(\{B^r\}_r) \geq \tilde{J}_E^*$. Now fix such a sequence and assume that

$$J_E(\{B^r\}_r) = \liminf_{r \rightarrow \infty} J_E^r(B^r) < \infty,$$

since otherwise the inequality is immediate. The subsequence $\{r'\} \subset \{r\}$ along which the liminf is achieved will be labeled again as $\{r\}$. With this relabeling, $J_E(\{B^r\}_r) = \lim_{r \rightarrow \infty} J_E^r(B^r)$. Fix $\delta > 0$ and choose $r_0 \in \mathbb{N}$ such that for all $r \geq r_0$, $J_E(\{B^r\}_r) \geq J_E^r(B^r) - \delta$. Given $r \geq r_0$, choose $T_r \geq 1$ such that

$$J_E^r(B^r) \geq \frac{E\zeta \cdot \hat{U}^r(T)}{T} + \frac{1}{T} \int_0^T E[h \cdot \hat{Q}^r(t)] dt - \delta \text{ for all } T \geq T_r.$$

We assume without loss of generality that $T_r \rightarrow \infty$ as $r \rightarrow \infty$. Thus

$$\frac{E\zeta \cdot \hat{U}^r(T)}{T} + \frac{1}{T} \int_0^T E[h \cdot \hat{Q}^r(t)] dt \leq J_E(\{B^r\}_r) + 2\delta \text{ for all } r \geq r_0 \text{ and } T \geq T_r. \quad (40)$$

For the rest of this section we assume without loss of generality that $r_0 = 1$. For $r \in \mathbb{N}$ and $T > 0$ consider $\mathcal{P}(\mathcal{D}^{I+2J+I+1})$ -valued random variable $v^{r,T}$ defined as

$$v^{r,T}(A) \doteq \frac{1}{T} \int_0^{V^r(T)} \delta_{(\check{w}^r(t+\cdot), \check{x}^{r,t}(\cdot), \check{\theta}^{r,t}(\cdot), \check{u}^{r,t}(\cdot), \check{v}^{r,t}(\cdot))}(A) d\check{V}^r(t), \quad A \subset \mathcal{D}^{I+2J+I+1},$$

where for a process $\{\zeta^r(t)\}_{t \geq 0}$, $\zeta^{r,t}(\cdot) = \zeta^r(t + \cdot) - \zeta^r(t)$. The following lemma gives a key tightness property.

Lemma 10. $\{v^{r,T}, r \in \mathbb{N}, T \geq T_r\}$ is a tight collection of $\mathcal{P}(\mathcal{D}^{I+2J+I+1})$ -valued random variables.

Proof. (Sketch) We first show that

$$\text{the collection } \{\check{X}^r(t+\cdot) - \check{X}^r(t), r \in \mathbb{N}, t \geq 0\} \text{ is tight.} \quad (41)$$

For this first note that exactly as in the proof of Lemma 1 it can be shown that the collection $\{\hat{X}^r(t+\cdot) - \hat{X}^r(t), r \in \mathbb{N}, t \geq 0\}$ is \mathcal{C} -tight in \mathcal{D}^J . Also, as in the proof of Lemma 3, it can be shown that $\{\check{V}^r(t+\cdot) - \check{V}^r(t), r \in \mathbb{N}, t \geq 0\}$ is tight in \mathcal{C}_+^1 . From this it follows that

$$\{\hat{X}^r(\check{V}^r(t) + \check{V}^r(t+\cdot) - \check{V}^r(t)) - \hat{X}^r(\check{V}^r(t)), r \in \mathbb{N}, t \geq 0\}$$

is \mathcal{C} -tight in \mathcal{D}^J . The statement in (41) now follows on observing that

$$\hat{X}^r(\check{V}^r(t) + \check{V}^r(t+\cdot) - \check{V}^r(t)) - \hat{X}^r(\check{V}^r(t)) = \check{X}^r(t+\cdot) - \check{X}^r(t).$$

In a similar way it can be seen that $\{\check{\theta}^r(t+\cdot) - \check{\theta}^r(t), r \in \mathbb{N}, t \geq 0\}$ is \mathcal{C} -tight in \mathcal{D}^J . We now argue that with

$$v_1^{r,T}(A) \doteq \frac{1}{T} \int_0^{V^r(T)} \delta_{(\check{X}^{r,t}(\cdot), \check{\theta}^{r,t}(\cdot), \check{V}^{r,t}(\cdot))}(A) d\check{V}^r(t), A \subset \mathcal{D}^{2J+1},$$

we have that

$$\{v_1^{r,T}, r \in \mathbb{N}, T \geq T_r\} \text{ is tight collection of } \mathcal{P}(\mathcal{D}^{2J+1}) \text{ valued random variables.} \quad (42)$$

For this it is enough to show that the collection $\{E v_1^{r,T}, r \in \mathbb{N}, T \geq T_r\}$ is relatively compact in $\mathcal{P}(\mathcal{D}^{2J})$. Note that, by a similar calculation as in (28), for $r \in \mathbb{N}$, $M \geq 1$ and $T \geq 1$,

$$P(V^r(T) \geq MT) \leq \frac{I}{(M-1)T} \sum_{i=1}^I E \hat{U}_i^r(T). \quad (43)$$

From (40) there is a $c_1 \in (0, \infty)$ such that $E \hat{U}_i^r(T) \leq c_1 T$ for all $T \geq T_r$ and for all $r \in \mathbb{N}$. Fix $\varepsilon > 0$ and choose $M \geq 1$ such that $c_1 I^2 / (M-1) \leq \varepsilon/2$. From the tightness of $\{\check{X}^{r,t}(\cdot), \check{\theta}^{r,t}(\cdot), \check{V}^{r,t}(\cdot)\}$ we can find a compact $K \subset \mathcal{D}^{2J+1}$ such that

$$\sup_{r \in \mathbb{N}, t \geq 0} P((\check{X}^{r,t}(\cdot), \check{\theta}^{r,t}(\cdot), \check{V}^{r,t}(\cdot)) \in K^c) \leq \frac{\varepsilon}{2M}.$$

Then, for $r \in \mathbb{N}$ and $T \geq T_r$

$$\begin{aligned}
E v_1^{r,T}(K^c) &= E \frac{1}{T} \int_0^{V^r(T)} \delta_{(\check{x}^{rt}(\cdot), \check{\theta}^{rt}(\cdot), \check{v}^{rt}(\cdot))}(K^c) d\check{V}^r(t) \\
&\leq P(V^r(T) \geq MT) + E \frac{1}{T} \int_0^{MT} \delta_{(\check{x}^{rt}(\cdot), \check{\theta}^{rt}(\cdot), \check{v}^{rt}(\cdot))}(K^c) d\check{V}^r(t) \\
&\leq \frac{cI^2}{(M-1)T} + M \frac{\varepsilon}{2M} \leq \varepsilon/2 + \varepsilon/2 = \varepsilon.
\end{aligned}$$

This proves the tightness statement in (42).

Next, by an argument similar to that in Lemma 3, we have the tightness of the collection $\{\check{U}^r(t+\cdot) - \check{U}^r(t), r \in \mathbb{N}, t \geq 0\}$. Together with the tightness in (41), this gives the tightness of the collection $\{\check{W}^r(t+\cdot) - \check{W}^r(t), r \in \mathbb{N}, t \geq 0\}$. Now an argument similar to the one used in the proof of (42) gives the tightness of $\{v_2^{r,T}, r \in \mathbb{N}, T \geq T_r\}$, where

$$v_2^{r,T}(A) \doteq \frac{1}{T} \int_0^{V^r(T)} \delta_{(\check{W}^r(t+\cdot) - \check{W}^r(t), \check{U}^r(t+\cdot) - \check{U}^r(t))}(A) d\check{V}^r(t), \quad A \subset \mathcal{D}^{2I}.$$

In order to complete the proof of the lemma it now suffices to show that

$$\left\{ v_3^{r,T}, r \in \mathbb{N}, T \geq T_r \right\} \text{ is a tight collection of } \mathcal{P}(\mathbb{R}_+^I) \text{ valued random variables,} \quad (44)$$

where

$$v_3^{r,T} = \frac{1}{T} \int_0^{V^r(T)} \delta_{\check{W}^r(t)} d\check{V}^r(t).$$

For this, again, it suffices to show that the collection $\{E v_3^{r,T}, r \in \mathbb{N}, T \geq T_r\}$ is relatively compact in $\mathcal{P}(\mathbb{R}_+^I)$. However, this is immediate on observing that, for some $c_2, c_3 < \infty$, and all $r \in \mathbb{N}, T \geq T_r$

$$\begin{aligned}
E \int_{\mathbb{R}_+^I} z \cdot \mathbf{1} d v_3^{r,T}(dz) &= E \frac{1}{T} \int_0^{V^r(T)} \check{W}^r(t) \cdot \mathbf{1} d\check{V}^r(t) \\
&= E \frac{1}{T} \int_0^T \hat{W}^r(t) \cdot \mathbf{1} dt \\
&\leq c_2 E \frac{1}{T} \int_0^T h \cdot \hat{Q}^r(t) dt \leq c_3,
\end{aligned}$$

where the last inequality is from (40). This completes the proof of the lemma. \square

The above lemma says that the collection $\{v^{r,T_r}, r \in \mathbb{N}\}$ is tight. Now we characterize the weak limit points of v^{r,T_r} . The proof follows along the lines of Theorem 3. We omit the details here.

Lemma 11. *Let \mathbf{v} be a $\mathcal{P}(\mathcal{D}^{I+2J+I+1})$ -valued random variable defined on some probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ such that $v^{r,T_r} \Rightarrow \mathbf{v}$ as $r \rightarrow \infty$. Let*

$$\{(\check{w}(t), \check{x}(t), \check{\theta}(t), \check{u}(t), \check{v}(t)), t \geq 0\}$$

denote the canonical coordinate process on $\mathbf{S} = \mathcal{D}^{I+2J+I+1}$. Then for \tilde{P} a.e. ω , and $\nu(\omega)$ a.s., $t \mapsto \check{\nu}(t)$ is continuous and $\check{\nu}(t) \rightarrow \infty$ as $t \rightarrow \infty$. Define, for $t \geq 0$, $\bar{\nu}(t) \doteq \inf\{s \geq 0 : \check{\nu}(s) > t\}$ and $\nu(t) \doteq \bar{\nu}(t) - \bar{\nu}(0)$. Let $w(t) \doteq \check{w}(\nu(t))$, $x(t) \doteq \check{x}(\nu(t))$, $\theta(t) \doteq \check{\theta}(\nu(t))$, $u(t) \doteq \check{u}(\nu(t))$. Consider the map $\Upsilon : \mathbf{S} \rightarrow \mathbf{S}_0$, where $\mathbf{S}_0 \doteq \mathcal{D}^{I+2J+I}$, defined ($\nu(\omega)$ a.s. for \tilde{P} a.e. ω) as

$$\Upsilon(\check{w}, \check{x}, \check{\theta}, \check{u}, \check{\nu})(t) = (w(t), x(t), \theta(t), u(t)), t \geq 0., (\check{w}, \check{x}, \check{\theta}, \check{u}, \check{\nu}) \in \mathbf{S}.$$

Let $\tilde{\nu}(\omega) = \nu(\omega) \circ \Upsilon^{-1}$. Abusing notation, denote the canonical coordinate process on \mathbf{S}_0 as $\{(w(t), x(t), \theta(t), u(t)), t \geq 0\}$ and let $\mathcal{E}(t) = \sigma\{(w(s), x(s), \theta(s), u(s)), s \leq t\}$ denote the canonical filtration. Then, for \tilde{P} -a.e. $\omega \in \tilde{\Omega}$,

- (a) $\{x(t), t \geq 0\}$ is an $\mathcal{E}(t)$ -Brownian motion with covariance Σ , under $\tilde{\nu}(\omega)$,
- (b) $\theta(t) = \theta t$ for all $t \geq 0$, $\tilde{\nu}(\omega)$ a.s.
- (c) For all $t \geq 0$, a.s. $\tilde{\nu}(\omega)$,

$$w(t) = w(0) + G\theta t + Gx(t) + u(t).$$

- (d) $u \in \tilde{\mathcal{A}}_E$ where $\tilde{\mathcal{A}}_E$ is as introduced below Definition 5, with $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}(t)\}, \tilde{X}(\cdot))$ replaced with $(\mathbf{S}_0, \mathcal{B}(\mathbf{S}_0), \tilde{\nu}(\omega), \{\mathcal{E}(t)\}, x(\cdot))$.

Now we can complete the proof of the second statement in Theorem 2. We assume without loss of generality that the ν^{r,T_r} converges in distribution to some limit ν . This ν must satisfy the properties in Lemma 11. Let $\nu_*^{r,T}$ denote the marginal of $\nu^{r,T}$ given by

$$\nu_*^{r,T} = \frac{1}{T} \int_0^{V^r(T)} \delta_{\check{W}^r(t+\cdot)} d\check{V}^r(t). \quad (45)$$

By Fatou's lemma and Lemma 11, we have that for any $N \in (0, \infty)$,

$$\begin{aligned} \lim_{r \rightarrow \infty} J_E^r(B^r) + \delta &\geq \liminf_{r \rightarrow \infty} \left(\frac{1}{T_r} \int_0^{T_r} E[h \cdot \hat{Q}^r(t)] dt + \frac{E[\zeta \cdot \hat{U}^r(T_r)]}{T_r} \right) \\ &\geq \liminf_{r \rightarrow \infty} E \left[\frac{1}{T_r} \int_0^{T_r} \hat{h}(\hat{W}^r(t)) dt \right] \\ &\geq \liminf_{r \rightarrow \infty} E \left[\frac{1}{T_r} \int_0^{T_r} (\hat{h}(\hat{W}^r(t)) \wedge N) dt \right] \\ &= \liminf_{r \rightarrow \infty} E \left[\frac{1}{T_r} \int_0^{V^r(T_r)} (\hat{h}(\check{W}^r(t)) \wedge N) d\check{V}^r(t) \right] \\ &= \liminf_{r \rightarrow \infty} E \left[\int_{\mathcal{D}^I} (\hat{h}(\varphi(0)) \wedge N) \nu_*^{r,T} (d\varphi) \right]. \end{aligned}$$

Using the weak convergence of ν^{r,T_r} to ν , we have

$$\begin{aligned} \lim_{r \rightarrow \infty} J_E^r(B^r) + \delta &\geq E \left[\int_{\mathbb{R}_+^I} (\hat{h}(x) \wedge N) v \circ (\check{w}(0))^{-1}(dx) \right] \\ &= E \left[\int_{\mathbb{R}_+^I} (\hat{h}(x) \wedge N) \tilde{v} \circ (w(0))^{-1}(dx) \right], \end{aligned}$$

where the last equality follows on recalling that with v in Lemma 11 satisfies $v(0) = 0$, $v(\omega)$ a.s. for \tilde{P} a.e. ω . Finally recalling the definition of \tilde{J}_E^* and sending $N \rightarrow \infty$, we have

$$J_E(\{B^r\}_r) + \delta \geq E \left[\int_{\mathbb{R}_+^I} \hat{h}(x) v \circ (w(0))^{-1}(dx) \right] \geq \tilde{J}_E^*.$$

The result follows since $\delta > 0$ is arbitrary. \square

References

1. Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.
2. Volker Böhm. On the continuity of the optimal policy set for linear programs. *SIAM Journal on Applied Mathematics*, 28(2):303–306, 1975.
3. Maury Bramson and RJ Williams. Two workload properties for brownian networks. *Queueing Systems*, 45(3):191–221, 2003.
4. Amarjit Budhiraja, Arka Prasanna Ghosh, et al. Diffusion approximations for controlled stochastic networks: An asymptotic bound for the value function. *The Annals of Applied Probability*, 16(4):1962–2006, 2006.
5. Amarjit Budhiraja and Dane Johnson. Control policies approaching HGI performance in heavy traffic for resource sharing networks. *arXiv preprint arXiv:1710.09042*, 2017.
6. Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.
7. J. M. Harrison and R. J. Williams. Brownian models of open queueing networks with homogeneous customer populations. *Stochastics*, 22(2):77–115, 1987.
8. J Michael Harrison. Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic differential systems, stochastic control theory and applications*, pages 147–186. Springer, 1988.
9. J Michael Harrison. Brownian models of open processing networks: Canonical representation of workload. *The Annals of Applied Probability*, 13(1):390–393, 2003.
10. J Michael Harrison, Chinmoy Mandayam, Devavrat Shah, and Yang Yang. Resource sharing networks: Overview and an open problem. *Stochastic Systems*, 4(2):524–555, 2014.
11. J Michael Harrison and Jan A Van Mieghem. Dynamic control of brownian networks: state space collapse and equivalent workload formulations. *The Annals of Applied Probability*, pages 747–771, 1997.
12. WN Kang, FP Kelly, NH Lee, and RJ Williams. State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *The Annals of Applied Probability*, pages 1719–1780, 2009.
13. Laurent Massoulié and James W Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecommunication systems*, 15(1-2):185–201, 2000.



American Option Model and Negative Fichera Function on Degenerate Boundary

Xiaoshan Chen, Zhuo Jin, and Qingshuo Song

Abstract We study American put option with stochastic volatility whose value function is associated with a 2-dimensional parabolic variational inequality with degenerate boundaries. Given the Fichera function on the boundary, we first analyze the existences of the strong solution and the properties of the 2-dimensional manifold for the free boundary. Thanks to the regularity result of the underlying PDE, we can also provide the uniqueness of the solution by the argument of the verification theorem together with the generalized Itos formula even though the solution may not be second order differentiable in the space variable across the free boundary.

1 Introduction

Option pricing is one of the most important topics in the quantitative finance research. Although the Black-Scholes model has been well studied, empirical evidence suggests that the Black-Scholes model is inadequate to describe asset returns and the behavior of the option markets. One possible remedy is to assume that the volatility of the asset price also follows a stochastic process, see [9] and [10].

In the standard Black-Scholes model, a standard logarithmic change of variables transforms the Black-Scholes equation into an equation with constant coefficients which can be studied by a general PDE theory directly. Different from the standard Black-Scholes PDE, the general PDE methods does not directly apply to the PDE

Xiaoshan Chen

Department of Mathematics, South China Normal University, e-mail: xchen53@gmail.com

Zhuo Jin

Centre for Actuarial Studies, Department of Economics, The University of Melbourne, e-mail: zhuo.jin@unimelb.edu.au

Qingshuo Song

Department of Mathematics, Worcester Polytechnic Institute, City University of Hong Kong, e-mail: qsong@wpi.edu

associated with the option pricing underlying the stochastic volatility model in the following cases: 1) The pricing equation is degenerate at the boundary; 2) The drift and volatility coefficients may grow faster than linear, see [9].

In the related literatures, [9] derived a closed-form solution for the price of a European call option on some specific stochastic volatility models. [6] also studied the Black-Scholes equation of European option underlying stochastic volatility models, and showed that the value function was the unique classical solution to a PDE with a certain boundary behavior. Also, [2] showed a necessary and sufficient condition on the uniqueness of classical solution to the valuation PDE of European option in a general framework of stochastic volatility models. In contrast to the European option pricing on the stochastic volatility model, although there have been quite a few approximate solutions and numerical approaches, such as [1, 4], the study of the existence and uniqueness of strong solution for PDE related to American option price on the stochastic volatility model is rather limited. In particular, the unique solvability of PDE associated with American options of finite maturity with the presence of degenerate boundary and super-linear growth has not been studied in an appropriate Sobolev space.

Note that, American call options with no dividend is equivalent to the European call option. For this reason, we only consider a general framework of American put option model with stochastic volatility whose value function is associated by a 2-dimensional parabolic variational inequality. On the other hand, in the theory of linear PDE, boundary conditions along degenerate boundaries should not be needed if the Fichera function is nonnegative, otherwise it should be imposed, see [16]. Therefore, we only consider the case when Fichera function is negative on the degenerate boundary $y = 0$ in this paper, and leave the other case in the future study.

To resolve solvability issue, we adopt similar methodology of [3] to work on a PDE of truncated version backward in time using appropriate penalty function and mollification. The main difference is that [3] studies constant drift and volatility, while the current paper considers functions of drift and volatility and the negative Fichera function plays a crucial role in the proof.

Uniqueness issue is usually tackled by comparison result implied by Ishii's lemma with notions of viscosity solution, see [5]. However, this approach does not apply in this problem due to the fast growth of drift and volatility functions on unbounded domain. The approach to establish uniqueness in our paper is similar to the classical verification theorem conducted to classical PDE solution. In fact, a careful construction leads to a local regularity of the solutions in Sobolev space, and this enables us to apply generalized Ito's formula (see [12]) with weak derivatives. Note that this approach not only provides uniqueness of strong solution of PDE, but also provides that the value function of American option is exactly the unique strong solution.

In the next section, we first introduce the generalized stochastic volatility model. Section 3 shows the existence of strong solution to the truncated version of the variational inequality. We characterize the free boundary in section 4. Section 5 shows that the value function of the American option price is the unique strong

solution to the variational inequality with appropriate boundary datum. Concluding remarks are given in section 6.

2 Stochastic volatility model

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a filtered probability space with a filtration $(\mathcal{F}_t)_{t \geq 0}$ that satisfies the usual conditions, W_t and B_t be two standard Brownian motions with correlation ρ . Suppose the stock price follows

$$\text{(Stk)} \quad dX_s = X_s(rds + \sigma(Y_s)dW_s), \quad X_t = x > 0,$$

and volatility follows

$$\text{(Vol)} \quad dY_s = \mu(Y_s)ds + b(Y_s)dB_s, \quad Y_t = y > 0.$$

Let $X^{x,t}$ and $Y^{y,t}$ be dynamics satisfying (Stk) and (Vol) with respective initial conditions on superscripts.

We consider an American put option underlying the asset X_s with strike $K > 0$ and maturity T , which has the payoff $(K - X_\tau)^+$ at the exercise time $\tau \in \mathcal{T}_{t,T}$. Here $\mathcal{T}_{t,T}$ denotes the set of all stopping times in $[t, T]$. Define value function of the optimal stopping by

$$V(x, y, t) = \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}_{x,y,t} [e^{-r(\tau-t)}(K - X_\tau)^+], \quad (x, y, t) \in \bar{\mathbb{R}}^+ \times \bar{\mathbb{R}}^+ \times [0, T]. \quad (1)$$

Following assumptions will be imposed:

(A1) μ, σ^2, b^2 are locally Lipschitz continuous on \mathbb{R} with $\mu(0) = \sigma(0) = b(0) = 0$ and $\sigma(y), b(y) > 0, \sigma'(y) \geq 0$ for all $y > 0$.

(A2) $|\mu| + b$ is at most linear growth, and $(\sigma^2)'$ is at most polynomial growth.

In the above, $\sigma'(y)$ and $(\sigma^2)'(y)$ stand for the first and second derivatives of $\sigma(y)$, respectively. Under the assumptions (A1)–(A2), we have unique non-negative, non-explosive strong solutions for both (Stk) and (Vol). Furthermore, $X_s^{x,t} > 0$ for all $s > t$. Such a stock price model includes Heston model.

Provided that the value function $V(x, y, t)$ is smooth enough, applying dynamic programming principle and Itô's formula, the value of the option $V(x, y, t)$ formally satisfies the variational inequality

$$\begin{cases} \min \left\{ -\partial_t V - \mathcal{L}_x V, V - (K - x)^+ \right\} = 0, & (x, y, t) \in \mathcal{Q} := \mathbb{R}^+ \times \mathbb{R}^+ \times [0, T], \\ V(x, y, T) = (K - x)^+, & (x, y) \in \mathbb{R}^+ \times \mathbb{R}^+, \end{cases} \quad (2)$$

where

$$\mathcal{L}_x V = \frac{1}{2} x^2 \sigma^2(y) \partial_{xx} V + \rho x \sigma(y) b(y) \partial_{xy} V + \frac{1}{2} b^2(y) \partial_{yy} V + r x \partial_x V + \mu(y) \partial_y V - r V.$$

On the boundary $x = 0$, the Fichera condition on linear parabolic equation suggests us not to impose any boundary condition. On the boundary $y = 0$, the Fichera function is

$$F = \left[\mu(y) - \frac{1}{2} \rho \sigma(y) b(y) - b(y) b'(y) \right] \Big|_{y=0} = \mu(0) - \lim_{y \rightarrow 0} b(y) b'(y).$$

So

(F1) when $\mu(0) < \lim_{y \rightarrow 0} b(y) b'(y)$, one has to impose the boundary condition;

(F2) when $\mu(0) \geq \lim_{y \rightarrow 0} b(y) b'(y)$, one should not impose any boundary condition.

Although the problem (2) is a variational inequality instead of linear PDE, the Fichera condition in the linear PDE theory does not directly prove the existence of solution to the problem (2). Throughout this paper, we study the relation between value function (1) and PDE of (2) with an appropriate boundary data on $y = 0$ under the case (F1), while the case (F2) is left in the future study.

Then, what boundary condition should be imposed on the boundary $y = 0$? To proceed, we define $\nu := \inf\{s > t : Y_s = 0\}$ the first hitting time of the process Y_s to the boundary $y = 0$. Then for any stopping time $\tau > \nu$, we have

$$e^{-r(\tau-t)} (K - X_\tau)^+ \leq e^{-r(\nu-t)} (K - X_\nu)^+, \quad (3)$$

thus

$$V(x, 0, t) = \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}_{x,0,t} [e^{-r(\tau-t)} (K - X_\tau)^+] \leq (K - x)^+.$$

On the other hand, by taking $\tau = t$,

$$V(x, 0, t) = \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}_{x,0,t} [e^{-r(\tau-t)} (K - X_\tau)^+] \geq (K - x)^+.$$

Hence

$$V(x, 0, t) = (K - x)^+ = \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}_{x,0,t} [e^{-r(\tau-t)} (K - X_\tau)^+], \quad (x, t) \in \mathbb{R}^+ \times [0, T]. \quad (4)$$

Due to the non-linearity of the variational inequality (2), we may not expect the heuristic argument in the above assuming the enough regularity on V . In addition, Fichera condition on the boundary data is suitable only for linear second order PDE, see [16]. Our objective in this paper is to justify the regularity of the value function in the Sobolev space, so that the value function (1) can be characterized as the unique solution of the variational inequality (2) and an additional boundary data (4).

3 Solvability on the transformed problem

In order to obtain the existence of solution to the variational inequality (2) with boundary condition (4) in Sobolev space, we consider the existence of strong solution to the associated transformed problem of (2) with boundary condition (4) in this section.

To proceed, we take a simple logarithm transformation to the variational inequality. Let $s = \ln x$, $\theta = T - t$, $u(s, y, \theta) = V(x, y, t)$, then

$$\begin{aligned} \mathcal{L}_x V &= \frac{1}{2} \sigma^2(y) \partial_{ss} u + \rho \sigma(y) b(y) \partial_{sy} u \\ &\quad + \frac{1}{2} b^2(y) \partial_{yy} u + \left(r - \frac{1}{2} \sigma^2(y) \right) \partial_s u + \mu(y) \partial_y u - ru \\ &:= \mathcal{L}_s u. \end{aligned}$$

Thus $u(s, y, \theta)$ satisfies

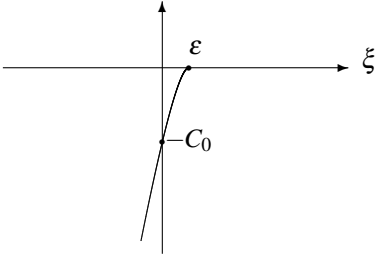
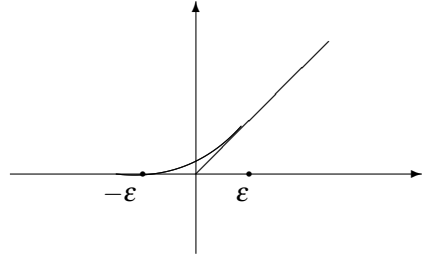
$$\begin{cases} \min \left\{ \partial_\theta u - \mathcal{L}_s u, u - (K - e^s)^+ \right\} = 0, & (s, y, \theta) \in \mathcal{Q} := \mathbb{R} \times \mathbb{R}^+ \times (0, T], \\ u(s, y, 0) = (K - e^s)^+, & s \in \mathbb{R}, y \in \mathbb{R}^+, \\ u(s, 0, \theta) = (K - e^s)^+, & s \in \mathbb{R}, \theta \in (0, T]. \end{cases} \quad (5)$$

Problem (5) is a variational inequality, we apply penalty approximation techniques to show the existence of strong solution to (5). Suppose $u_\varepsilon(s, y, \theta)$ satisfies

$$\begin{cases} \partial_\theta u_\varepsilon - \mathcal{L}_s^\varepsilon u_\varepsilon + \beta_\varepsilon(u_\varepsilon - \pi_\varepsilon(K - e^s)) = 0, & (s, y, \theta) \in \mathcal{Q}, \\ u_\varepsilon(s, y, 0) = \pi_\varepsilon(K - e^s), & s \in \mathbb{R}, y \in \mathbb{R}^+, \\ u_\varepsilon(s, 0, \theta) = \pi_\varepsilon(K - e^s), & s \in \mathbb{R}, \theta \in (0, T]. \end{cases} \quad (6)$$

where $\mathcal{L}_s^\varepsilon u = \frac{1}{2} (\sigma^2(y) + \varepsilon) \partial_{ss} u + \rho (\sigma(y) b(y) + \varepsilon) \partial_{sy} u + \frac{1}{2} (b^2(y) + \varepsilon) \partial_{yy} u + (r - \frac{1}{2} \sigma^2(y) - \frac{1}{2} \varepsilon) \partial_s u + \mu(y) \partial_y u - ru$, and $\beta_\varepsilon(\xi)$ (Fig. 1.), $\pi_\varepsilon(\xi)$ (Fig. 2.) satisfy

$$\begin{aligned} &\beta_\varepsilon(\xi) \in C^2(-\infty, +\infty), \beta_\varepsilon(\xi) \leq 0, \beta_\varepsilon(0) = -2r(K + 1), \beta'_\varepsilon(\xi) \geq 0, \beta''_\varepsilon(\xi) \leq 0. \\ \lim_{\varepsilon \rightarrow 0} \beta_\varepsilon(\xi) &= \begin{cases} 0, & \xi > 0, \\ -\infty, & \xi < 0, \end{cases} \quad \pi_\varepsilon(\xi) = \begin{cases} \xi, & \xi \geq \varepsilon, \\ \nearrow, & |\xi| \leq \varepsilon, \\ 0, & \xi \leq -\varepsilon, \end{cases} \\ \pi_\varepsilon(\xi) \in C^\infty, & 0 \leq \pi'_\varepsilon(\xi) \leq 1, \pi''_\varepsilon(\xi) \geq 0, \lim_{\varepsilon \rightarrow 0} \pi_\varepsilon(\xi) = \xi^+. \end{aligned}$$

Fig. 1. $\beta_\varepsilon(\xi)$ Fig. 2. $\pi_\varepsilon(\xi)$

Since $\mathcal{Q} = (-\infty, +\infty) \times (0, +\infty) \times (0, T]$ is infinitely, we consider the truncated version of (6), denote $\mathcal{Q}^N = (-N, N) \times (0, N) \times (0, T]$, let $u_\varepsilon^N(s, y, \theta)$ satisfy

$$\begin{cases} \partial_\theta u_\varepsilon^N - \mathcal{L}_s^\varepsilon u_\varepsilon^N + \beta_\varepsilon(u_\varepsilon^N - \pi_\varepsilon(K - e^s)) = 0, & (s, y, \theta) \in \mathcal{Q}^N, \\ u_\varepsilon^N(s, y, 0) = \pi_\varepsilon(K - e^s), & (s, y) \in (-N, N) \times (0, N), \\ u_\varepsilon^N(s, 0, \theta) = \pi_\varepsilon(K - e^s), & (s, \theta) \in (-N, N) \times (0, T], \\ \partial_y u_\varepsilon^N(s, N, \theta) = 0, & (s, \theta) \in (-N, N) \times (0, T], \\ u_\varepsilon^N(-N, y, \theta) = \pi_\varepsilon(K - e^{-N}), & (y, \theta) \in (0, N) \times (0, T], \\ \partial_s u_\varepsilon^N(N, y, \theta) = 0, & (y, \theta) \in (0, N) \times (0, T]. \end{cases} \quad (7)$$

Lemma 1. For any fixed $\varepsilon, N > 0$, there exists a unique solution $u_\varepsilon^N \in W_p^{2,1}(\mathcal{Q}^N)$ to the problem (7), and

$$\pi_\varepsilon(K - e^s) \leq u_\varepsilon^N(s, y, \theta) \leq K + 1, \quad (8)$$

$$\partial_\theta u_\varepsilon^N(s, y, \theta) \geq 0. \quad (9)$$

Proof. For any fixed $\varepsilon, N > 0$, it is not hard to show by the fixed point theorem that problem (7) has a solution $u_\varepsilon^N \in W_p^{2,1}(\mathcal{Q}^N)$, and

$$|u_\varepsilon^N|_0 \leq C(|\beta_\varepsilon(u_\varepsilon^N - \pi_\varepsilon(K - e^s))|_0 + |\pi_\varepsilon(K - e^s)|_0) \leq C(|\beta_\varepsilon(-K - 1)| + K + 1) \leq C.$$

The proof of uniqueness is a standard way as well.

Since

$$\begin{aligned} & \partial_\theta \pi_\varepsilon(K - e^s) - \mathcal{L}_s^\varepsilon(\pi_\varepsilon(K - e^s)) + \beta_\varepsilon(0) \\ &= -\frac{1}{2}(\sigma^2(y) + \varepsilon)(\pi_\varepsilon''(\cdot)e^{2s} - \pi_\varepsilon'(\cdot)e^s) + \left(r - \frac{1}{2}\sigma^2(y) - \frac{1}{2}\varepsilon\right)\pi_\varepsilon'(\cdot)e^s + r\pi_\varepsilon(K - e^s) + \beta_\varepsilon(0) \\ &\leq r(K + \varepsilon) + r(K + \varepsilon) + \beta_\varepsilon(0) \leq 0. \end{aligned}$$

Combining with the initial and boundary conditions, we know $\pi_\varepsilon(K - e^s)$ is a sub-solution of (7). Similarly we know $K + 1$ is a supersolution of (7).

Next we will prove (9). Set $u^\delta(s, y, \theta) := u_\varepsilon^N(s, y, \theta + \delta)$, then $u^\delta(s, y, \theta)$ satisfies

$$\begin{cases} \partial_\theta u^\delta - \mathcal{L}_s^\varepsilon u^\delta + \beta_\varepsilon(u^\delta - \pi_\varepsilon(K - e^s)) = 0, \\ u^\delta(s, y, 0) = u_\varepsilon^N(s, y, \delta) \geq \pi_\varepsilon(K - e^s) = u_\varepsilon^N(s, y, 0), \\ u^\delta(s, 0, \theta) = \pi_\varepsilon(K - e^s), \\ \partial_y u^\delta(s, N, \theta) = 0, \\ u^\delta(-N, y, \theta) = \pi_\varepsilon(K - e^{-N}), \\ \partial_s u^\delta(N, y, \theta) = 0. \end{cases}$$

Applying the comparison principle, we have

$$u^\delta(s, y, \theta) \geq u_\varepsilon^N(s, y, \theta), \quad (s, y, \theta) \in (-N, N) \times (0, N) \times (0, T - \delta],$$

which yields $\partial_\theta u_\varepsilon^N \geq 0$.

Letting $N \rightarrow +\infty$, the existence of strong solution to the penalty problem (6) with some estimates is given in the following theorem.

Lemma 2. *For any fixed $\varepsilon > 0$, there exists a unique solution $u_\varepsilon(s, y, \theta) \in W_{p,loc}^{2,1}(\mathcal{Q}) \cap C^1(\overline{\mathcal{Q}})$ to the problem (6) for any $1 < p < +\infty$, and*

$$\pi_\varepsilon(K - e^s) \leq u_\varepsilon(s, y, \theta) \leq K + 1, \tag{10}$$

$$\partial_\theta u_\varepsilon(s, y, \theta) \geq 0, \tag{11}$$

$$-e^s \leq \partial_s u_\varepsilon(s, y, \theta) \leq 0, \tag{12}$$

$$\partial_y u_\varepsilon(s, y, \theta) \geq 0. \tag{13}$$

Proof. For any fixed $\varepsilon > 0, R > 0$, applying $W_p^{2,1}$ interior estimate with part of boundary [13] to the problem (7) ($N > R$), then

$$|u_\varepsilon^N|_{W_p^{2,1}(\mathcal{Q}^R)} \leq C(|\beta_\varepsilon(u_\varepsilon^N - \pi_\varepsilon(K - e^s))|_{L^\infty(\mathcal{Q}^R)} + |\pi_\varepsilon(K - e^s)|_{W_p^{2,1}(\overline{\mathcal{Q}^R} \cap \{\theta=0\})}) \leq C,$$

where C depends on ε, R but is independent of N . Letting $N \rightarrow +\infty$, by the imbedding theorem, we know problem (6) has a solution $u_\varepsilon(s, y, \theta) \in W_{p,loc}^{2,1}(\mathcal{Q}) \cap C^1(\overline{\mathcal{Q}})$. (10)–(11) are consequences of (8)–(9).

Now we aim to prove (12). Differentiate the equation in (6) w.r.t. s and denote $w_1 = \partial_s u_\varepsilon$, then

$$\begin{cases} \partial_\theta w_1 - \mathcal{L}_s^\varepsilon w_1 + \beta'_\varepsilon(\cdot)w_1 = -\beta'_\varepsilon(\cdot)\pi'_\varepsilon(\cdot)e^s \leq 0, & (s, y, \theta) \in \mathcal{Q}, \\ w_1(s, y, 0) = w_1(s, 0, \theta) = -\pi'_\varepsilon(\cdot)e^s \leq 0. \end{cases} \tag{14}$$

Applying the maximum principle [17] we know $w_1 = \partial_s u_\varepsilon \leq 0$. In view of

$$(\partial_\theta - \mathcal{L}_s^\varepsilon)(-e^s) + \beta'_\varepsilon(\cdot)(-e^s) = -\beta'_\varepsilon(\cdot)e^s \leq -\beta'_\varepsilon(\cdot)\pi'_\varepsilon(\cdot)e^s.$$

Combining with the initial and boundary conditions, applying the comparison principle we have

$$-e^s \leq w_1(s, y, \theta) = \partial_s u_\varepsilon(s, y, \theta) \leq 0.$$

Finally we want to prove (13). We first differentiate (14) w.r.t. s , denote $w_2 = \partial_{ss} u_\varepsilon$, we obtain

$$\begin{cases} \partial_\theta w_2 - \mathcal{L}_s^\varepsilon w_2 + \beta'_\varepsilon(\cdot) w_2 = -\beta'_\varepsilon(\cdot) \pi'_\varepsilon(\cdot) e^s + \beta'_\varepsilon(\cdot) \pi''_\varepsilon(\cdot) e^{2s} - \beta''_\varepsilon(\cdot) [\pi'_\varepsilon(\cdot) e^s + w_1]^2, \\ w_2(s, y, 0) = w_2(s, 0, \theta) = -\pi'_\varepsilon(\cdot) e^s + \pi''_\varepsilon(\cdot) e^{2s}. \end{cases} \quad (15)$$

Set $w_3(s, y, \theta) := w_2(s, y, \theta) - w_1(s, y, \theta)$, in view of (14) and (15)

$$\begin{cases} \partial_\theta w_3 - \mathcal{L}_s^\varepsilon w_3 + \beta'_\varepsilon(\cdot) w_3 = \beta'_\varepsilon(\cdot) \pi''_\varepsilon(\cdot) e^{2s} - \beta''_\varepsilon(\cdot) [\pi'_\varepsilon(\cdot) e^s + w_1]^2 \geq 0, \\ w_3(s, y, 0) = w_3(s, 0, \theta) = \pi''_\varepsilon(\cdot) e^{2s} \geq 0. \end{cases}$$

Applying maximum principle we know $w_3(s, y, \theta) \geq 0$, i.e., $\partial_{ss} u_\varepsilon - \partial_s u_\varepsilon \geq 0$.

Differentiate (6) w.r.t. y , denote $w_4(s, y, \theta) = \partial_y u_\varepsilon(s, y, \theta)$. Then we get

$$\begin{cases} \partial_\theta w_4 - \mathcal{L}_s^\varepsilon w_4 - \rho(\sigma'(y)b(y) + \sigma(y)b'(y)) \partial_s w_4 - b(y)b'(y) \partial_y w_4 \\ \quad - \mu'(y) w_4 + \beta'_\varepsilon(\cdot) w_4 = \sigma(y) \sigma'(y) (\partial_{ss} u_\varepsilon - \partial_s u_\varepsilon), \\ w_4(s, y, 0) = 0, \\ w_4(s, 0, \theta) \geq 0. \end{cases}$$

Since $\sigma'(y) \geq 0$, $u_\varepsilon \in C^{2,1}(\mathcal{Q})$ and $\partial_{ss} u_\varepsilon - \partial_s u_\varepsilon \geq 0$, by maximum principle [17], we have $\partial_y u_\varepsilon(s, y, \theta) \geq 0$.

Now we are able to show the solvability on the variational inequality (5) in the Sobolev space by the approximation of a subsequence of $\{u_\varepsilon\}$.

Lemma 3. *There exists a solution $u \in W_p^{2,1}(\mathcal{Q}_\delta^N \setminus B_h)$ to the problem (5), where $\mathcal{Q}_\delta^N = (-N, N) \times (\delta, N) \times (0, T]$, $B_h = (\ln K - h, \ln K + h) \times (0, +\infty) \times (0, T]$ for any $N, \delta, h > 0$. Moreover,*

$$(K - e^s)^+ \leq u(s, y, \theta) \leq K + 1, \quad (16)$$

$$\partial_\theta u(s, y, \theta) \geq 0, \quad (17)$$

$$-e^s \leq \partial_s u(s, y, \theta) \leq 0, \quad (18)$$

$$\partial_y u(s, y, \theta) \geq 0. \quad (19)$$

Proof. Since $\sigma(y)$, $b(y)$ are continuous and $\sigma'(y) \geq 0$, in $\mathcal{Q}_{\frac{1}{2}}^N$, we have $\sigma^2(y) + \varepsilon \geq \sigma^2(\frac{1}{2}) > 0$, and $\lambda_1 |\xi|^2 \leq a^{ij} \xi_i \xi_j \leq \Lambda_1 |\xi|^2$, with Λ_1, λ_1 independent of ε . Applying $C^{\alpha, \alpha/2}$ estimate [14] and $W_p^{2,1}$ interior estimate with part of boundary [13], we have

$$|u_\varepsilon|_{C^{\alpha, \alpha/2}(\overline{\mathcal{Q}_{\frac{1}{2}}^N})} \leq C(|u_\varepsilon|_0 + |\beta_\varepsilon(u_\varepsilon - \pi_\varepsilon(K - e^s))|_0 + [\pi_\varepsilon(K - e^s)]_{C^{\gamma}(-N, N)}) \leq C_1,$$

$$|u_\varepsilon|_{W_p^{2,1}(\mathcal{Q}_{\frac{1}{2}}^N \setminus B_h)} \leq C(|u_\varepsilon|_{L^\infty} + |\beta_\varepsilon(u_\varepsilon - \pi_\varepsilon(K - e^s))|_{L^\infty} + |(K - e^s) \vee 0|_{W_p^{2,1}(\mathcal{Q}_{\frac{1}{2}}^N \setminus B_h)}) \leq C_2,$$

where C_1, C_2 are independent of ε due to the estimate (10) and the definitions of $\beta_\varepsilon, \pi_\varepsilon$. Thus there exists a subsequence of $\{u_\varepsilon\}$, denote $\{u_\varepsilon^{(1)}\}$, and $u^{(1)} \in W_p^{2,1}(\mathcal{Q}_{\frac{1}{2}}^N \setminus B_h) \cap C(\overline{\mathcal{Q}_{\frac{1}{2}}^N})$, such that

$$\begin{aligned} u_\varepsilon^{(1)}(s, y, \theta) &\rightharpoonup u^{(1)}(s, y, \theta) \quad \text{in } W_p^{2,1}(\mathcal{Q}_{\frac{1}{2}}^N \setminus B_h) \text{ weakly,} \\ u_\varepsilon^{(1)}(s, y, \theta) &\rightarrow u^{(1)}(s, y, \theta) \quad \text{in } C(\overline{\mathcal{Q}_{\frac{1}{2}}^N}) \text{ uniformly.} \end{aligned}$$

In a same way, in $\mathcal{Q}_{\frac{1}{3}}^N$, we have $\sigma^2(y) + \varepsilon \geq \sigma^2(\frac{1}{3})$, and $\lambda_2|\xi|^2 \leq a^{ij}\xi_i\xi_j \leq \Lambda_2|\xi|^2$, with Λ_2, λ_2 independent of ε . Thus there exists $\{u_\varepsilon^{(2)}\} \subseteq \{u_\varepsilon^{(1)}\}$, $u^{(2)} \in W_p^{2,1}(\mathcal{Q}_{\frac{1}{3}}^N \setminus B_h) \cap C(\overline{\mathcal{Q}_{\frac{1}{3}}^N})$, such that

$$\begin{aligned} u_\varepsilon^{(2)}(s, y, \theta) &\rightharpoonup u^{(2)}(s, y, \theta) \quad \text{in } W_p^{2,1}(\mathcal{Q}_{\frac{1}{3}}^N \setminus B_h) \text{ weakly,} \\ u_\varepsilon^{(2)}(s, y, \theta) &\rightarrow u^{(2)}(s, y, \theta) \quad \text{in } C(\overline{\mathcal{Q}_{\frac{1}{3}}^N}) \text{ uniformly.} \end{aligned}$$

Moreover,

$$u^{(2)}(s, y, \theta) = u^{(1)}(s, y, \theta), \quad (s, y, \theta) \in \mathcal{Q}_{\frac{1}{2}}^N.$$

Define $u(s, y, \theta) = u^{(k)}(s, y, \theta)$, if $(s, y, \theta) \in \mathcal{Q}_{\frac{1}{k+1}}^N$, abstracting diagram subsequence $\{u_{\varepsilon_k}^{(k)}\}$, for any $\delta, h, N > 0$, we have

$$\begin{aligned} u_{\varepsilon_k}^{(k)}(s, y, \theta) &\rightharpoonup u(s, y, \theta) \quad \text{in } W_p^{2,1}(\mathcal{Q}_\delta^N \setminus B_h) \text{ weakly,} \\ u_{\varepsilon_k}^{(k)}(s, y, \theta) &\rightarrow u(s, y, \theta) \quad \text{in } C(\overline{\mathcal{Q}_\delta^N}) \text{ uniformly,} \end{aligned}$$

thus $u(s, y, \theta) \in W_p^{2,1}(\mathcal{Q}_\delta^N \setminus B_h) \cap C(\overline{\mathcal{Q}} \setminus \{y = 0\})$ and $u(s, y, \theta)$ satisfies the variational inequality in (5) and the initial condition.

Next we will prove the continuity on the degenerate boundary $y = 0$. For any $s_0 \in \mathbb{R} \setminus \{\ln K\}$, then there exists $\varepsilon_0 > 0$ such that $\pi_{\varepsilon_0}(K - e^{s_0}) = (K - e^{s_0})^+$, denote $w_0(s, y, \theta) = \pi_{\varepsilon_0}(K - e^s) + Ay^\alpha \geq 0$, with $0 < \alpha < 1$, and $A \geq 1$ to be determined, then for any $\varepsilon < \varepsilon_0$

$$\begin{aligned} &\partial_\theta w_0 - \mathcal{L}_s^\varepsilon w_0 + \beta_\varepsilon(w_0 - \pi_\varepsilon(K - e^s)) \\ &= -\frac{1}{2}(\sigma^2(y) + \varepsilon)\pi_{\varepsilon_0}''(\cdot)e^{2s} - \frac{1}{2}(b^2(y) + \varepsilon)A\alpha(\alpha - 1)y^{\alpha-2} + r\pi_{\varepsilon_0}'(\cdot)e^s \\ &\quad - \mu(y)\alpha Ay^{\alpha-1} + rw_0 + \beta_\varepsilon(\pi_{\varepsilon_0}(K - e^s) - \pi_\varepsilon(K - e^s) + Ay^\alpha) \\ &\geq -\frac{1}{2}(\sigma^2(y) + \varepsilon)\pi_{\varepsilon_0}''(\cdot)e^{2s} + \frac{1}{2}(b^2(y) + \varepsilon)A\alpha(1 - \alpha)y^{\alpha-2} - \mu(y)\alpha Ay^{\alpha-1} + \beta_\varepsilon(0), \end{aligned}$$

since the negative Fichera function indicates $\mu(0) < \lim_{y \rightarrow 0} b(y)b'(y)$, in addition, $b^2(y) = O(y)$ or $b^2(y) = o(y)$ when $y \rightarrow 0$, hence there exists $\delta_0 > 0$ small enough and independent of ε such that

$$-\frac{1}{2}(\sigma^2(y) + \varepsilon)\pi_{\varepsilon_0}''(\cdot)e^{2s} + \frac{1}{2}(b^2(y) + \varepsilon)A\alpha(1 - \alpha)y^{\alpha-2} - \mu(y)\alpha Ay^{\alpha-1} + \beta_\varepsilon(0) \geq 0$$

for any $y \in (0, \delta_0)$. Moreover, we can choose A large enough such that $A\delta_0^\alpha \geq K + 1$. Combining

$$\pi_{\varepsilon_0}(K - e^s) + Ay^\alpha \geq \pi_\varepsilon(K - e^s), \quad \varepsilon < \varepsilon_0,$$

applying comparison principle, we have

$$\pi_\varepsilon(K - e^s) \leq u_\varepsilon(s, y, \theta) \leq \pi_{\varepsilon_0}(K - e^s) + Ay^\alpha, \quad (s, y, \theta) \in \mathbb{R} \times (0, \delta_0) \times (0, T].$$

Letting $\varepsilon \rightarrow 0^+$ we have

$$(K - e^s)^+ \leq u(s, y, \theta) \leq \pi_{\varepsilon_0}(K - e^s) + Ay^\alpha, \quad (s, y, \theta) \in \mathbb{R} \times (0, \delta_0) \times (0, T].$$

In particular

$$(K - e^{s_0})^+ \leq u(s_0, y, \theta) \leq (K - e^{s_0})^+ + Ay^\alpha, \quad y \in (0, \delta_0).$$

Letting $y \rightarrow 0^+$, we obtain

$$u(s_0, 0, \theta) = (K - e^{s_0})^+, \quad \theta \in (0, T],$$

since s_0 is arbitrary, then $u(s, 0, \theta) = (K - e^s)^+$, $s \in \mathbb{R} \setminus \{\ln K\}$. Therefore $u(s, y, \theta)$ is a solution to the problem (5). (16)–(19) are consequences of (10)–(13).

4 Characterization of free boundary to the problem (5)

Variational inequality (5) is an obstacle problem, this section aims to characterize the free boundary arise from (5).

Lemma 4. *The solution to the problem (5) satisfies*

$$u(s, y, \theta) > 0, \quad (s, y, \theta) \in \mathbb{R} \times \mathbb{R}^+ \times (0, T].$$

Proof. For any fixed $y_0 > 0$, we have

$$\begin{cases} \partial_\theta u - \mathcal{L}_s u \geq 0, & (s, y, \theta) \in \mathbb{R} \times (y_0, +\infty) \times (0, T], \\ u(s, y, 0) = (K - e^s)^+ \geq 0, & (s, y) \in \mathbb{R} \times (y_0, +\infty), \\ u(s, y_0, \theta) \geq (K - e^s)^+ \geq 0, & (s, \theta) \in \mathbb{R} \times (0, T]. \end{cases}$$

Applying strong maximum principle, we obtain

$$u(s, y, \theta) > 0, \quad (s, y, \theta) \in \mathbb{R} \times (y_0, +\infty) \times (0, T].$$

Since y_0 is arbitrary, then we know

$$u(s, y, \theta) > 0, \quad (s, y, \theta) \in \mathbb{R} \times \mathbb{R}^+ \times (0, T].$$

In order to characterize the free boundary, we first define

$$\begin{aligned} \mathcal{C}[u] &:= \{(s, y, \theta) : u(s, y, \theta) = (K - e^s)^+\} \text{(Coincidence set)}, \\ \mathcal{N}[u] &:= \{(s, y, \theta) : u(s, y, \theta) > (K - e^s)^+\} \text{(Noncoincidence set)}. \end{aligned}$$

Thanks to the estimates (16)–(19) of the solution to (5), problem (5) gives rise to a free boundary that can be expressed as a function of (y, θ) . The following three lemmas give the existence and properties of the free boundary.

Proposition 1. *There exists $h(y, \theta) : \mathbb{R}^+ \times (0, T] \rightarrow \mathbb{R}$, such that*

$$\mathcal{C}[u] = \{(s, y, \theta) \in \mathcal{D} : s \leq h(y, \theta), y \in \mathbb{R}^+, \theta \in (0, T]\}. \tag{20}$$

Moreover, for any fixed $y > 0$, $h(y, \theta)$ is monotonic decreasing w.r.t. θ ; for any fixed $\theta \in (0, T]$, $h(y, \theta)$ is monotonic decreasing w.r.t. y .

Proof. Since $(K - e^s)^+ = 0$ when $s \geq \ln K$, in view of Lemma 4, we have

$$\{s \geq \ln K\} \subset \mathcal{N}[u], \quad \mathcal{C}[u] \subset \{s < \ln K\}.$$

Hence problem (5) is equivalent to the following problem

$$\begin{cases} \min \left\{ \partial_\theta u - \mathcal{L}_s u, u - (K - e^s) \right\} = 0, & (s, y, \theta) \in \mathcal{D} := \mathbb{R} \times \mathbb{R}^+ \times (0, T], \\ u(s, y, 0) = (K - e^s)^+, & s \in \mathbb{R}, y \in \mathbb{R}^+, \\ u(s, 0, \theta) = (K - e^s)^+, & s \in \mathbb{R}, \theta \in (0, T]. \end{cases}$$

Together with (18), we can define

$$h(y, \theta) := \max\{s \in \mathbb{R} : u(s, y, \theta) = (K - e^s)\}, \quad (y, \theta) \in \mathbb{R}^+ \times (0, T],$$

by the definition of $h(y, \theta)$, we know (20) is true.

Suppose $h(y, \theta_1) = s_1$, notice that $\partial_\theta u(s, y, \theta) \geq 0$, then for any $\theta_2 \leq \theta_1$,

$$0 \leq u(s_1, y, \theta_2) - (K - e^{s_1}) \leq u(s_1, y, \theta_1) - (K - e^{s_1}) = 0,$$

from which we infer that

$$u(s_1, y, \theta_2) = (K - e^{s_1}), \quad \theta_2 \leq \theta_1.$$

By the definition of $h(y, \theta)$, we know $h(y, \theta_2) \geq s_1 = h(y, \theta_1)$, thus $h(y, \cdot)$ is monotonic decreasing w.r.t. θ .

Similarly, the monotonicity of $h(y, \theta)$ w.r.t. y can be deduced by virtue of $\partial_y u(s, y, \theta) \geq 0$ and the definition of $h(y, \theta)$.

Proposition 2. $h(y, \theta)$ is continuous on $\mathbb{R}^+ \times [0, T]$ with

$$h(y, 0) := \lim_{\theta \rightarrow 0^+} h(y, \theta) = \ln K, \quad y > 0.$$

Proof. We first prove $h(y, \theta)$ is continuous w.r.t. θ . Suppose not. There exists $y_0 > 0$, $\theta_0 > 0$ such that $s_1 := h(y_0, \theta_0 +) < h(y_0, \theta_0) := s_2$. Since $h(y_0, \theta_0 +) = s_1$ (see Fig. 3.), then

$$u(s, y_0, \theta) > K - e^s, \quad s > s_1, \quad \theta > \theta_0.$$

In fact, $u \in W_p^{2,1}$ and the embedding theorem imply that u is uniformly continuous, thus there exists $\delta > 0$, take $\mathcal{S}_0 = (s_1, s_2) \times (y_0 - \delta, y_0)$ such that $U_0 := \mathcal{S}_0 \times (\theta_0, T] \subseteq \mathcal{N}[u]$, then

$$\partial_\theta u - \mathcal{L}_s u = 0, \quad (s, y, \theta) \in U_0.$$

Moreover, in view of $h(y_0, \theta_0) := s_2$, then

$$h(y, \theta_0) \geq s_2, \quad y_0 - \delta < y \leq y_0,$$

hence

$$u(s, y, \theta_0) = K - e^s, \quad s \leq s_2, \quad y_0 - \delta < y \leq y_0.$$

In particular,

$$u(s, y, \theta_0) = K - e^s, \quad (s, y) \in \overline{U_0} \cap \{\theta = \theta_0\},$$

thus

$$\begin{aligned} \partial_\theta u|_{\theta=\theta_0} &= \mathcal{L}_s u|_{\theta=\theta_0} = \frac{1}{2} \sigma^2(y) (-e^s) + \left(r - \frac{1}{2} \sigma^2(y) \right) (-e^s) - r(K - e^s) \\ &= -rK < 0, \end{aligned}$$

which comes to a contradiction with the fact that $\partial_\theta u \geq 0$. Hence $h(y, \theta)$ is continuous w.r.t. θ .

Since $h(y, \theta)$ is monotonic decreasing w.r.t. θ , then we can define $h(y, 0) := \lim_{\theta \rightarrow 0^+} h(y, \theta)$. In the same way we can prove $h(y, 0) = \ln K$.

Now we aim to prove the continuity of $h(y, \theta)$ w.r.t. y . If this is not true, there exists $\theta_0, y_0 > 0$ such that $s_1 := h(y_0 +, \theta_0) < h(y_0, \theta_0) := s_2$. Since $h(y, \theta)$ is continuous w.r.t. θ and $u(s, y, \theta) \in C^{1,1/2}(\mathcal{Q})$, take $U_0 := (\tilde{s}, \bar{s}) \times (y_0, +\infty) \times (\tilde{\theta}, \bar{\theta})$, where $(\tilde{s}, \bar{s}) \times (\tilde{\theta}, \bar{\theta}) \subseteq (h(y_0 +, \theta), h(y_0, \theta))$ (see Fig. 4.), then $u(s, y, \theta)$ satisfies

$$\begin{aligned} \partial_\theta u - \mathcal{L}_s u &= 0, \quad (s, y, \theta) \in U_0, \\ u(s, y_0, \theta) &= K - e^s, \quad (s, \theta) \in \overline{U_0} \cap \{y = y_0\}, \\ \partial_y u(s, y_0, \theta) &= 0, \quad (s, \theta) \in \overline{U_0} \cap \{y = y_0\}. \end{aligned}$$

Thus

$$\partial_\theta u(s, y_0, \theta) = \partial_{y,\theta} u(s, y_0, \theta) = 0, \quad (s, \theta) \in \overline{U_0} \cap \{y = y_0\}.$$

Since $\partial_\theta u \geq 0$ and $\partial_\theta(\partial_\theta u) - \mathcal{L}_s(\partial_\theta u) = 0$ in U_0 , by Hopf lemma we know

$$\partial_{y,\theta} u(s, y_0, \theta) > 0, \quad (s, \theta) \in \overline{U_0} \cap \{y = y_0\},$$

or

$$\partial_\theta u(s, y, \theta) \equiv 0, \quad (s, y, \theta) \in U_0,$$

but both come to contradictions.

Together with the monotonicity of $h(y, \theta)$ w.r.t. y and θ , we conclude that $h(y, \theta)$ is continuous on $\mathbb{R}^+ \times [0, T]$.

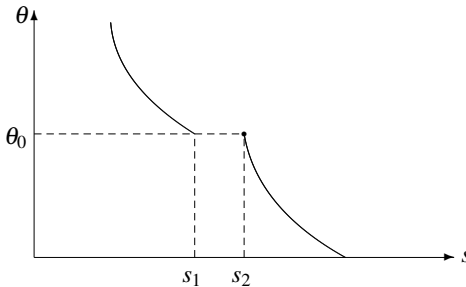


Fig. 3. Discontinuity of $h(y, \theta)$ w.r.t. θ

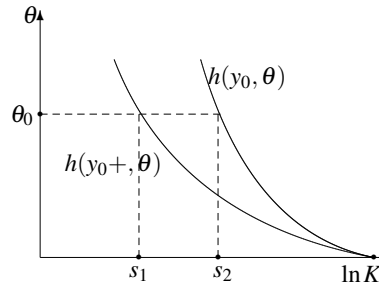


Fig. 4. Discontinuity of $h(y, \theta)$ w.r.t.

Proposition 3. *The free boundary $h(y, \theta)$ satisfies*

$$h_0(y) \leq h(y, \theta) < \ln K, \quad y > 0, \quad \theta \in (0, T],$$

where $h_0(y)$ is the free boundary curve of

$$\min\{-\mathcal{L}_s u_\infty(s, y), u_\infty(s, y) - (K - e^s)^+\} = 0, \quad (s, y) \in \mathbb{R} \times \mathbb{R}^+. \quad (21)$$

Proof. Since $\partial_\theta u_\infty(s, y) = 0$, then we can rewrite (21) as

$$\begin{cases} \min\{\partial_\theta u_\infty - \mathcal{L}_s u_\infty, u_\infty - (K - e^s)^+\} = 0, & (s, y, \theta) \in \mathcal{Q}, \\ u_\infty(s, y)|_{\theta=0} = u_\infty(s, y) \geq (K - e^s)^+ = u(s, y, 0). \end{cases}$$

Applying the monotonicity of solution of variational inequality w.r.t. initial value, we have

$$u_\infty(s, y) \geq u(s, y, \theta), \quad \theta \geq 0.$$

By the definitions of $h_0(y)$ and $h(y, \theta)$, we know

$$h_0(y) \leq h(y, \theta).$$

Now we will prove $h(y, \theta) < \ln K$. Suppose not. There exists $y_0 > 0$, $\theta_0 > 0$ such that $h(y_0, \theta_0) = \ln K$. Then by the monotonicity of $h(y, \theta)$ and the fact that $h(y, \theta) \leq \ln K$, we have

$$u(s, y, \theta) = K - e^s, \quad (s, y, \theta) \in [0, \ln K] \times (0, y_0] \times (0, \theta_0].$$

Thus

$$\partial_\theta u(\ln K, y, \theta) = \partial_{s\theta} u(\ln K, y, \theta) = 0, \quad (y, \theta) \in (0, y_0] \times (0, \theta_0].$$

Since $\partial_\theta u(s, y, \theta) \geq 0$ and $\partial_\theta(\partial_\theta u) - \mathcal{L}_s(\partial_\theta u) = 0, s > \ln K$. By Hopf lemma [7] we know

$$\partial_{s\theta} u(\ln K, y, \theta) > 0, \quad (y, \theta) \in (0, y_0] \times (0, \theta_0],$$

or

$$\partial_\theta u(s, y, \theta) = 0, \quad (s, y, \theta) \in (\ln K, +\infty) \times (0, y_0] \times (0, \theta_0],$$

but both come to contradictions.

Remark 1. The numerical result of $h_0(y)$ under Heston model is given in [20], and by similar methods, under the assumptions (A1)–(A2), we can also obtain the existence of $h_0(y)$.

5 Characterization of the value function

Now, we are ready to present the characterization of the value function of (1) to the variational inequality (2) with boundary condition (4). To proceed, we present the solvability and regularity results on the variational inequality (2) with boundary condition (4) via the counterpart on the transformed problem obtained in the above.

Theorem 1. *Suppose a bounded function $v(x, y, t) \in W_{p,loc}^{2,1}(\mathcal{Q}) \cap C(\bar{\mathcal{Q}})$ satisfies variational inequality (2) with boundary condition (4), the following assertions hold.*

1. $v(x, y, t)$ satisfies the following estimates

$$(K - x)^+ \leq v(x, y, t) \leq K + 1, \tag{22}$$

$$-1 \leq \partial_x v(x, y, t) \leq 0, \tag{23}$$

$$\partial_y v(x, y, t) \geq 0. \tag{24}$$

2. *There exists a continuous function $g(y,t) : \mathbb{R}^+ \times [0,T) \rightarrow \mathbb{R}^+$, such that for any fixed $y > 0$, $g(y,t)$ is monotonic increasing w.r.t. t ; for any fixed $t \in [0,T)$, $g(y,t)$ is monotonic decreasing w.r.t. y with*

$$g(y,t) < g(y,T) = K, \quad y > 0, t \in [0,T),$$

and

$$\begin{cases} -\partial_t v(x,y,t) - \mathcal{L}_x v(x,y,t) = 0, & x > g(y,t), \\ v(x,y,t) = (K-x)^+, & 0 \leq x \leq g(y,t). \end{cases}$$

3. *Especially, $v(x,y,t) \in C^{2,1}$ when $x > g(y,t)$.*

Proof. By the transformations $s = \ln x, \theta = T - t, u(s,y,\theta) = v(x,y,t)$, noting $x\partial_x v(x,y,t) = \partial_s u(s,y,\theta)$ and using estimates (16), (18) and (19), we can obtain (22)–(24).

Let $g(y,t) = \exp\{h(y,\theta)\}$, by proposition 1–2, we can conclude 2.

For any $x_0 > g(y_0, t_0)$, then $v(x_0, y_0, t_0) > (K - x_0)^+$, since $v(x,y,t)$ is uniformly continuous, there exists a disk $B_\delta(x_0, y_0, t_0)$ with center (x_0, y_0, t_0) and radius δ such that

$$v(x,y,t) > (K-x)^+, \quad (x,y,t) \in B_\delta(x_0, y_0, t_0).$$

Applying $C^{2,1}$ interior estimate to

$$\partial_t v(x,y,t) + \mathcal{L}_x v(x,y,t) = 0, \quad (x,y,t) \in B_\delta(x_0, y_0, t_0),$$

to obtain $v(x,y,t) \in C^{2,1}(B_\delta(x_0, y_0, t_0))$, hence $v(x,y,t) \in C^{2,1}$ when $x > g(y,t)$.

Finally, the uniqueness result is given in this below through the arguments of verification theorem.

Theorem 2. *Suppose there exists $v(x,y,t) \in W_{p,loc}^{2,1}(Q)$ to the problem (2) with boundary condition (4), then $v(x,y,t) \geq V(x,y,t)$. If, in addition, there exists the region $\mathcal{N}[v] := \{(x,y,t) \in Q, v(x,y,t) > (K-x)^+\}$ satisfies*

$$(\partial_t v + \mathcal{L}_x v)(X_s, Y_s, s) = 0, \quad s \in [t, \tau^*],$$

for the stopping time $\tau^* := \inf\{s > t : (X_s, Y_s, s) \notin \mathcal{N}[v]\} \wedge T$. Then the variational inequality (2) with boundary condition (4) admits a unique solution in $W_{p,loc}^{2,1}(Q)$ and $v(x,y,t) = V(x,y,t)$.

Proof. Let $\tau_x^\beta := \inf\{s > t : X_s \leq \frac{1}{\beta} \text{ or } X_s \geq \beta\} \wedge T$ be the first hitting time of the process X_s to the upper bound β or the lower bound $\frac{1}{\beta}$ or terminal time T , $\tau_y^\beta := \inf\{s > t : Y_s \leq \frac{1}{\beta} \text{ or } Y_s \geq \beta\} \wedge T$ be the first hitting time of the process Y_s to the upper bound β or the lower bound $\frac{1}{\beta}$ or terminal time T . Let $\tau \in \mathcal{T}_{t, \tau_x^\beta \wedge \tau_y^\beta}$, by the general Itô's formula [12],

$$\begin{aligned}
e^{-r(\tau-t)}v(X_\tau, Y_\tau, \tau) &= v(x, y, t) + \int_t^\tau e^{-r(s-t)}(\partial_t v + \mathcal{L}_x v)(X_s, Y_s, s)ds \\
&\quad + \int_t^\tau e^{-r(s-t)}[\sigma(Y_s)X_s \partial_x v dW_s + b(Y_s) \partial_y v dB_s]. \quad (25)
\end{aligned}$$

Since $v(x, y, t)$ is bounded and the Itô integrals in (25) are local martingales, hence they are martingales. Moreover, by Theorem 1, we know $v(x, y, t)$ satisfies $\partial_t v + \mathcal{L}_x v \leq 0$, $v(X_\tau, Y_\tau, \tau) \geq (K - X_\tau)^+$, hence

$$v(x, y, t) \geq \mathbb{E}_{x, y, t}[e^{-r(\tau-t)}(K - X_\tau)^+], \quad \tau \in \mathcal{T}_{t, \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}.$$

Since X_s is a positive, non-explosive local martingale, then

$$\lim_{\beta \rightarrow \infty} \tilde{\tau}_x^\beta = T, \quad a.s. - \mathbb{P}. \quad (26)$$

Since Y_s is non-negative, non-explosive local martingale, then

$$\lim_{\beta \rightarrow \infty} \tilde{\tau}_y^\beta = v \wedge T, \quad a.s. - \mathbb{P}, \quad (27)$$

where v is the first hitting time of Y_s to the boundary $y = 0$. Hence the arbitrariness of $\tau \in \mathcal{T}_{t, \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}$ and the above two limits imply that

$$v(x, y, t) \geq \sup_{\tau \in \mathcal{T}_{t, v \wedge T}} \mathbb{E}_{x, y, t}[e^{-r(\tau-t)}(K - X_\tau)^+]. \quad (28)$$

In view of (3), when $v < T$,

$$\mathbb{E}_{x, y, t}[e^{-r(v-t)}(K - X_v)^+] \geq \mathbb{E}_{x, y, t}[e^{-r(\tau-t)}(K - X_\tau)^+], \quad \tau \in \mathcal{T}_{v, T}.$$

Together with (28), we have

$$v(x, y, t) \geq \sup_{\tau \in \mathcal{T}_{t, T}} \mathbb{E}_{x, y, t}[e^{-r(\tau-t)}(K - X_\tau)^+] = V(x, y, t).$$

On the other hand, define $\tilde{\tau}_x^\beta := \inf\{s > t : X_s \geq \beta\} \wedge T$ be the first hitting time of the process X_s to the upper bound β or terminal time T , $\tilde{\tau}_y^\beta := \inf\{s > t : Y_s \geq \beta\} \wedge T$ be the first hitting time of the process Y_s to the upper bound β or terminal time T . By (26) and (27) we know

$$\lim_{\beta \rightarrow \infty} \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta = T, \quad a.s. - \mathbb{P}.$$

Together with Monotone Convergence Theorem, we obtain

$$\lim_{\beta \rightarrow \infty} \mathbb{E}_{x,t} \left[X_T I_{\{\tilde{\tau}_x^\beta = T\}} \right] = \mathbb{E}_{x,t} \left[\lim_{\beta \rightarrow \infty} X_T I_{\{\tilde{\tau}_x^\beta = T\}} \right] = \mathbb{E}_{x,t} [X_T], \quad (29)$$

$$\lim_{\beta \rightarrow \infty} \mathbb{E}_{y,t} \left[Y_T I_{\{\tilde{\tau}_y^\beta = T\}} \right] = \mathbb{E}_{y,t} \left[\lim_{\beta \rightarrow \infty} Y_T I_{\{\tilde{\tau}_y^\beta = T\}} \right] = \mathbb{E}_{y,t} [Y_T]. \quad (30)$$

Moreover, by the definitions of $\tilde{\tau}_x^\beta$, $\tilde{\tau}_y^\beta$, we can have

$$\begin{aligned} \mathbb{E}_{x,t} [X_{\tilde{\tau}_x^\beta}] &= \mathbb{E}_{x,t} \left[X_{\tilde{\tau}_x^\beta} I_{\{\tilde{\tau}_x^\beta < T\}} \right] + \mathbb{E}_{x,t} \left[X_T I_{\{\tilde{\tau}_x^\beta = T\}} \right] \\ &= \beta \mathbb{P}\{\tilde{\tau}_x^\beta < T\} + \mathbb{E}_{x,t} \left[X_T I_{\{\tilde{\tau}_x^\beta = T\}} \right]. \end{aligned}$$

Forcing the limit $\beta \rightarrow \infty$, due to (29),

$$\lim_{\beta \rightarrow \infty} \mathbb{E}_{x,t} [X_{\tilde{\tau}_x^\beta}] = \lim_{\beta \rightarrow \infty} \beta \mathbb{P}\{\tilde{\tau}_x^\beta < T\} + \mathbb{E}_{x,t} [X_T].$$

For all $\beta > x$, since $\{X_{\tilde{\tau}_x^\beta \wedge s} : s > t\}$ is a bounded local martingale, hence it is a martingale. So, $\mathbb{E}_{x,t} [X_{\tilde{\tau}_x^\beta}] = x$ for all $\beta > x$. Rearranging the above equality, we have

$$\lim_{\beta \rightarrow \infty} \beta \mathbb{P}\{\tilde{\tau}_x^\beta < T\} = x - \mathbb{E}_{x,t} [X_T] \leq x. \quad (31)$$

Similarly,

$$\lim_{\beta \rightarrow \infty} \beta \mathbb{P}\{\tilde{\tau}_y^\beta < T\} = y - \mathbb{E}_{y,t} [Y_T] \leq y. \quad (32)$$

By Theorem 1 we know $\mathcal{N}[v] = \{(x, y, t) \in \mathcal{Q} : x > g(y, t)\}$, noting that $v(x, y, t) \in C^{2,1}$ and $\partial_t v + \mathcal{L}_x v = 0$ in $\mathcal{N}[v]$, using the classical Itô's formula [15] in $[t, \tau^* \wedge \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta]$, we have

$$\begin{aligned} v(x, y, t) &= \mathbb{E}_{x,y,t} \left[e^{-r(\tau^* \wedge \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta - t)} v(X_{\tau^* \wedge \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, Y_{\tau^* \wedge \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, \tau^* \wedge \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta) \right] \\ &= \mathbb{E}_{x,y,t} \left[e^{-r(\tau^* - t)} v(X_{\tau^*}, Y_{\tau^*}, \tau^*) I_{\{\tau^* \leq \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta\}} \right] \\ &\quad + \mathbb{E}_{x,y,t} \left[e^{-r(\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta - t)} v(X_{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, Y_{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta) I_{\{\tau^* > \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta\}} \right]. \end{aligned}$$

Forcing $\beta \rightarrow +\infty$, since $\lim_{\beta \rightarrow \infty} \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta = T$,

$$\begin{aligned} v(x, y, t) &= \mathbb{E}_{x,y,t} [e^{-r(\tau^* - t)} (K - X_{\tau^*})^+] \\ &\quad + \lim_{\beta \rightarrow \infty} \mathbb{E}_{x,y,t} \left[e^{-r(\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta - t)} v(X_{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, Y_{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta) I_{\{\tau^* > \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta\}} \right]. \end{aligned}$$

In the following we show the limit $\lim_{\beta \rightarrow \infty} \mathbb{E}_{x,y,t} [e^{-r(\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta - t)} v(X_{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, Y_{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta) I_{\{\tau^* > \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta\}}]$ in the above equality is 0. Since $v(x, y, t) \leq K + 1$, then there exists

$g(\beta) = o(\beta)$, $\beta \rightarrow +\infty$ such that $v(X_{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, Y_{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta) \leq g(\beta) = o(\beta)$, together with (31) and (32), we can obtain

$$\begin{aligned}
0 &\leq \lim_{\beta \rightarrow \infty} \mathbb{E}_{x,y,t} \left[e^{-r(\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta - t)} v(X_{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, Y_{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta}, \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta) I_{\{\tau^* > \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta\}} \right] \\
&\leq \lim_{\beta \rightarrow \infty} g(\beta) \mathbb{E}_{x,y,t} \left[e^{-r(\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta - t)} I_{\{\tau^* > \tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta\}} \right] \\
&\leq \lim_{\beta \rightarrow \infty} \frac{g(\beta)}{\beta} \lim_{\beta \rightarrow \infty} \beta \mathbb{P}\{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta < \tau^*\} \\
&\leq \lim_{\beta \rightarrow \infty} \frac{g(\beta)}{\beta} \lim_{\beta \rightarrow \infty} \beta \mathbb{P}\{\tilde{\tau}_x^\beta \wedge \tilde{\tau}_y^\beta < T\} = 0.
\end{aligned}$$

Hence $v(x, y, t) = \mathbb{E}_{x,y,t} [e^{-r(\tau^* - t)} (K - X_{\tau^*})^+]$, therefore $v(x, y, t) = V(x, y, t)$.

6 Conclusion

In this paper, we consider an American put option of stochastic volatility with negative Fichera function on the degenerate boundary $y = 0$, we impose a proper boundary condition from the definition of the option pricing to show that the solution to the associated variational inequality is unique, which is the value of the option, and the free boundary is the optimal exercise boundary of the option. Although the asset-price volatility coefficient may grow faster than linear growth and the domain is unbounded, we are able to show the uniqueness by verification theorem. In this paper we only consider the payoff function $(K - x)^+$, but the method in this paper will be useful for any nonnegative, continuous payoff function $f(x)$ which is of strictly sublinear growth, i.e., $\lim_{x \rightarrow +\infty} \frac{f(x)}{x} = 0$.

The problem under study belongs to a general category of stochastic control problems, see for instance [18, 8]. Due to the nonlinearity, the numerical solution is always an issue in the general setup, see for instance [11] for Markov chain approximation method. In particular, regarding the current formulation of variational inequalities with Fichera functions on the boundary, it is unclear how its associated Markov chain behaves asymptotically as the step size goes to zero, see [19]. Appropriate scaling of step size at the boundary may be a key to make the obtained Markov chain consistent to the variational problem, and it will be pursued in our future work.

Acknowledgements This paper is partially supported by Faculty Research Grant of University of Melbourne, the Startup fund of WPI, the Research Grants Council of Hong Kong CityU (11201518), and CityU SRG of Hong Kong (7004667).

References

1. F. Aitsahlia, M. Goswami, and S. Guha. American option pricing under stochastic volatility: an efficient numerical approach. *Computational Management Science*, 2010.
2. E. Bayraktar, K. Kardaras, and H. Xing. Valuation equations for stochastic volatility models. *SIAM Journal on Financial Mathematics*, 3:351–373, 2012.
3. X. Chen, Q. Song, F. Yi, and G. Yin. Characterization of stochastic control with optimal stopping in a Sobolev space *Automatica*, 49:1654–1662, 2013.
4. C. Chiarella, B. Kang, G. H. Meyer, and A. Ziogas. The evaluation of American option prices under stochastic volatility and jump-diffusion dynamics using the method of lines. *International Journal of Theoretical Applied Finance*, 12: 393–425, 2009.
5. M. Crandall, H. Ishii, and P. Lions. User’s guide to viscosity solutions of second order partial differential equations. *Bull. Amer. Math. Soc. (N.S.)*, 27(1):1–67, 1992.
6. E. Ekstrom, and J. Tysk. The Black-Scholes equation in stochastic volatility models. *Journal of Mathematical Analysis and Applications*, 368:498–507, 2010.
7. A. Friedman. *Partial Differential Equations of Parabolic Type*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1964.
8. W. H. Fleming and H. M. Soner. *Controlled Markov processes and viscosity solutions*, volume 25 of *Stochastic Modelling and Applied Probability*. Springer, New York, second edition, 2006.
9. S. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6: 327–343, 1993.
10. J. C. Hull, and A. White. The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, 42:281–300, 1987.
11. H. J. Kushner and P. Dupuis. *Numerical methods for stochastic control problems in continuous time*, volume 24. Springer-Verlag, New York, second edition, 2001.
12. N. V. Krylov. Controlled diffusion processes, *volume 14 of Applications of Mathematics*. Springer-Verlag, New York, 1980.
13. O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural’ceva. *Linear and Quasilinear Equations of Parabolic Type*. *Translations of Mathematical Monographs, Vol. 23*. American Mathematical Society, Providence, R.I., 1967.
14. G. M. Lieberman. *Second Order Parabolic Differential Equations*. World Scientific Publishing Co. Inc., River Edge, N.J., 1996.
15. B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. 6th ed., Springer-Verlag, Berlin, 2003
16. O. A. Oleinik, E. V. Radkevich. *Second Order Equations With Nonnegative Characteristic Form*. Plenum Press, New York, 1973.
17. K. Tso. On an Aleksandrov-Bakel’man type maximum principle for second-order parabolic equations. *Comm. Partial Differential Equations*, 10(5):543–553, 1985.
18. Jiongmin Yong and Xun Yu Zhou. *Stochastic controls*, volume 43 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1999. Hamiltonian systems and HJB equations.
19. G. Yin and Q. Zhang. *Discrete-time Markov chains: Two-time-scale methods and applications*, volume 55 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 2005. Stochastic Modelling and Applied Probability.
20. S. Zhu, and W. Chen. Should an American option be exercised earlier or later if volatility is not assumed to be a constant? *International Journal of Theoretical and Applied Finance*, 14: 1279–1297, 2011.



Continuous-Time Markov Chain and Regime Switching Approximations with Applications to Options Pricing

Zhenyu Cui, J. Lars Kirkby and Duy Nguyen

Abstract In this chapter, we present recent developments in using the tools of continuous-time Markov chains for the valuation of European and path-dependent financial derivatives. We also survey results on a newly proposed regime switching approximation to stochastic volatility, and stochastic local volatility models. The presented framework is part of an exciting recent stream of literature on numerical option pricing, and offers a new perspective that combines the theory of diffusion processes, Markov chains, and Fourier techniques. It is also elegantly connected to partial differential equation (PDE) approaches.

1 Introduction

Markov processes are ubiquitous in finance, as they provide important building blocks for constructing stochastic models to describe the dynamics of financial assets. A representative Markov process that is widely used is the diffusion process, which is characterized through a stochastic differential equation (SDE). Diffusion processes evolve continuously in time and in state, and there is usually limited analytical tractability except for a few very special cases, thus an efficient and accurate approximation method is needed. In general, there are two possible directions for approximating a diffusion process:

Zhenyu Cui
School of Business, Stevens Institute of Technology, Hoboken, NJ 07030. e-mail:
zcuif@stevens.edu

J. Lars Kirkby
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA
30318 e-mail: jkirkby3@gatech.edu

Duy Nguyen
Corresponding Author. Department of Mathematics, Marist College, Poughkeepsie, NY 12601
e-mail: nducduy@gmail.com

1. *Time discretization*: discretize the time space into a finite discrete grid of time points, while preserving the continuous state space of the diffusion process, and then approximate the evolution of the diffusion process through time-stepping. Representative methods in this category include the Euler discretization as well as higher order time-stepping schemes (see [36] for a comprehensive account of existing methods), and the Ito-Taylor expansion method which is based on iterative applications of the Dynkin formula and fundamental properties of infinitesimal generators of the diffusion process.
2. *State discretization*: discretize the state space into a finite discrete grid of spatial points, while preserving the continuous time dimension of the diffusion process, and then approximate the evolution of the diffusion process through a continuous-time Markov chain (CTMC). A CTMC is a natural approximation tool here as it evolves continuously in time, and its transition density can be completely characterized through the *rate matrix* or *generator matrix*, which is a (discrete-state) analogue to the infinitesimal generator of the diffusion process.

There are pros and cons associated with either of the above two possible approximation methods, which will be discussed in details in subsequent sections. The previous (finance and economics) literature has mainly focused on the first approach, which we briefly summarize below:

- The Euler discretization has been very popular in numerical solutions of SDEs arising in finance, e.g. the Cox-Ingersoll-Ross (CIR) process. The convergence properties of the discretization scheme, and careful handling of the boundary behaviors have been discussed in the literature, see [35]. The Euler method is also the pillar for the “simulated maximum likelihood estimation” (SMLE) popular in financial econometrics, see [22]
- The Ito-Taylor expansion is based on a small-time expansion, and it has been applied in parameter estimation of diffusion process (see [3]), and options pricing (see [46]).

On the other hand, the second approximation approach has a relatively thinner literature and has received much less attention from academics in finance and economics. Thus it is our focus to survey the recent literature on CTMC approximation methods applied to options pricing. A brief summary of the extant literature is as follows:

- The Markov chain approximation method was first developed in the setting of general stochastic control theory, for which it yields tractable solutions for general Markovian control problems, see [44]. Note, however, that the main tool employed there is the discrete-time Markov chain (DTMC). The more specific application to finance (e.g. the Merton optimal investment and consumption problem) has been considered in [55].
- In the realm of options pricing, to the best of authors’ knowledge, the DTMC method was first applied in the GARCH option pricing setting, see [19, 20].

- The above previous literature concerns the DTMC method, and some of the more recent literature considers applying the CTMC method to both path-independent and path-dependent options pricing, see some of the recent developments in [8, 14, 54, 71, 72, 73]. Rigorous convergence analysis for the CTMC approximation method has been established in [47, 70] for the case of path-independent options, and in [53, 61] for the case of a class of path-dependent options (e.g. arithmetic Asian option and step option, which is based on the occupation time.).

Regime switching models are popular in financial applications, such as time series modeling (see [29]), interest rate/foreign exchange rate movements (see [4]), credit rating transitions (see [6]), economic booms and recessions (see [39]), stock trading (see [74]) etc. It has also been popular in the options pricing and portfolio choice literature, see for example [75, 76]. Note that most of the previous literature mentioned above concerns the regime switching model itself. Regime switching models are closely related to the continuous-time Markov chain. Intuitively, we can think of a CTMC as a stochastic process making transitions among a finite number of “regimes”. Regime switching models also reflect the idea of “random volatility”, since we can understand the different regime levels as corresponding to different volatility levels for the financial asset of interest. Motivated by these two insights, there is recent development in the literature utilizing the regime switching model as an approximation tool for continuous stochastic volatility models. The method reduces a multi-factor stochastic volatility model to a one-dimensional diffusion model subject to regime switching, and handy analytical expressions have been developed, see [12, 13, 14, 15, 43].

There are two major components in this chapter: first we shall describe the main ideas behind utilizing a CTMC in approximating a diffusion process, and then discuss the applications and survey the recent relevant literature; second, we depict the main ideas on regime switching approximation to continuous stochastic volatility and stochastic local volatility models.

The chapter is organized as follows: Section 2 recalls the basic theory underlying the use of a continuous time Markov chain to approximate a general diffusion process, and then presents the main method for approximating time-changed Markov processes. Section 3 presents the method for approximating general stochastic volatility models by a Markov modulated diffusion process, and furthermore by a Markov modulated CTMC for the case of stochastic local volatility models. Section 4 concludes the chapter.

2 Univariate Markov Chain Approximations

Early research in Markov chain based option pricing [9] dealt with approximating the univariate diffusion dynamics for an underlying risky asset. Our treatment starts with this case, and builds gradually to more complex dynamics, including general continuous stochastic local volatility models.

2.1 Markov Processes, Diffusion Models, and Option Pricing

Assume that we are equipped with a complete filtered probability space $(\Omega, \mathcal{F}, \mathbf{F}, \mathbb{P})$, where $\mathbf{F} = \{\mathcal{F}_t\}_{t \geq 0}$ denotes the standard filtration, and here \mathbb{P} is the risk-neutral measure under which we price options. Consider a real-valued (time-homogeneous) diffusion process $\{S_t\}_{t \geq 0}$, which satisfies the following stochastic differential equation:

$$dS_t = \mu(S_t)dt + \sigma(S_t)dW_t, \quad 0 \leq t \leq T, \quad (1)$$

where W_t is a standard Brownian motion, and $\mu, \sigma : \mathbb{R} \rightarrow \mathbb{R}$ are respectively drift and diffusion functions satisfying appropriate regularity conditions so that (1) has a unique solution¹. The random process $S = \{S_t\}_{t \geq 0}$ belongs to Markov process class, and is often used to model the price evolution of a risky asset, for example the stock price or the commodity price. For a rigorous and more in depth treatment of Markov processes, the reader is invited to refer to the monograph [24]. The diffusion characterized by (1) nests some important models in finance as special cases, such as the geometric Brownian motion (Black-Scholes model), the Cox-Ingersoll-Ross (CIR) process, etc. Assume that the state space for S is given by $\mathcal{S} = [0, \infty)$, and this is intuitive because most financial assets are positive valued. In general, we are interested in computing the following quantity:

$$\mathbb{E}[H(S_T)|S_0], \quad (2)$$

which is a conditional expectation for some payoff function H under the risk-neutral probability measure \mathbb{P} . For example, when $H(s) = \max(s - K, 0) = (s - K)^+$, it represents the payoff of a European call option with expiry T , and a strike price $K > 0$. This is a representative example for path-independent payoffs. As for path-dependent derivatives, [54] consider the expectation of the following form

$$\mathbb{E}[g(S_T)\mathbf{I}_{\{\tau_A > T\}} + H(S_{\tau_A})\mathbf{I}_{\{\tau_A \leq T\}}|S_0] \quad (3)$$

with $\tau_A = \inf\{t \geq 0 : S_t \in A\}$ denoting the first time that S enters the set A , which represents knock-in or knock-out events depending on contract specifications. Assuming that A represents knock-out events, then (3) concerns an option that consists of a payment $g(S_T)$ in the case the contract has not been knocked out by time T , and a rebate $H(S_{\tau_A})$ if it has. This type of (path-dependent) payoff is commonly encountered in the options market. Other variants of barrier options include the down-and-out, up-and-out, and double knock-out options. In particular, the expectation in (2) is just a special case of (3) when $A = \emptyset$.

For some special cases in which the probability density function of S is known, it is possible to obtain exact analytical expressions for $\mathbb{E}[H(S_T)|S_0]$. However, we

¹ Depending on particular applications, it can be either a strong or weak solution. Usually Lipschitz-type conditions are required for there to exist a unique strong solution (c.f. [33]). As for a unique-in-law weak solution to exist, the Engelbert-Schmidt condition (c.f. [38]) may be imposed. Since we are mainly interested in applications to options pricing, the existence of a unique-in-law weak solution is sufficient for our discussions.

note that, in general, it is difficult to compute $\mathbb{E}[H(S_T)|S_0]$ exactly for a general diffusion model. As a result, various numerical methods are considered. Some representative methods are numerical PDE methods (through the link provided by the Feynman-Kac theorem), Monte Carlo simulation methods, Fast Fourier Transform (FFT) methods (applicable only when the characteristic function of S is known), to name just a few. In this chapter, we consider an alternative yet very general approach through the use of continuous-time Markov chain approximations, which has been recently proposed in [54, 14, 49, 15] and has received appreciable attention. Next, for a bounded Borel function H , define

$$P_t H(x) := \mathbb{E}_x[H(S_t)] := \mathbb{E}[H(S_t)|S_0 = x]. \quad (4)$$

Recall that S satisfies the Markov property:

$$\mathbb{E}[H(S_{t+r})|\mathcal{F}_t] = P_r H(S_t). \quad (5)$$

From (5), by taking the expectation on both sides, it is easy to see that the family of (pricing) operators $(P_t)_{t \geq 0}$ forms a semigroup:

$$P_{t+r} H = P_t(P_r H), \quad \forall r, t \geq 0, \quad \text{and} \quad P_0 H = H. \quad (6)$$

Let $C_0(\mathcal{S})$ denote the set of continuous functions on the state space \mathcal{S} that vanish at infinity. To guarantee the existence of a *version* of S with càdlàg paths satisfying the (strong) Markov process, we assume the following Feller's properties:

Assumption 1 $S = \{S_t\}_{t \geq 0}$ is a Feller process on \mathcal{S} . That is, for any $H \in C_0(\mathcal{S})$, the family of operators $(P_t)_{t \geq 0}$ satisfies

- $P_t H \in C_0(\mathcal{S})$ for any $t \geq 0$;
- $\lim_{t \rightarrow 0} P_t H(x) = H(x)$ for any $x \in \mathcal{S}$.

The family $(P_t)_{t \geq 0}$ is determined by its infinitesimal generator \mathcal{L} , where

$$\mathcal{L}H(x) := \lim_{t \rightarrow 0^+} \frac{P_t H(x) - H(x)}{t}, \quad \forall H \in C_0(\mathcal{S}). \quad (7)$$

For the diffusion given in (1), we have

$$\mathcal{L}H(x) = \frac{1}{2} \sigma^2(x) \frac{\partial^2 H}{\partial x^2} + \mu(x) \frac{\partial H}{\partial x}. \quad (8)$$

For example, the standard Black-Scholes-Merton (BSM) model is described by $\mu(S) = (r - q) \cdot S$ and $\sigma(S) := \sigma \cdot S$, where $r, q \in \mathbb{R}$ represent the continuous rates of interest and dividends, respectively, and by abuse of notation σ is a constant volatility rate.

2.2 Markov Chain Approximation

With the basic setup in previous section, we shall discuss the construction of approximating CTMC for a particular diffusion process. In the literature, there have been various methods in the constructions, and they mainly differ in the allocation schemes of grid points to “fill up” the state space, see [49]. In this section, we shall introduce a particular method to construct the approximating CTMC, and the issue of optimal design of grids is discussed in Section 2.3. This work considers two main directions in the CTMC approximation literature. In the first case we will approximate the underlying process S_t directly, and we shall use \bar{n} to denote the number of states in the CTMC approximating the underlying asset process. In the second case we approximate a related (latent) stochastic factor, such as stochastic volatility, and will use \bar{m} to denote the number of states in the approximating CTMC of that stochastic factor. We start with the first approach.

Given the diffusion characterized by (1), the goal is to construct a continuous-time Markov chain $\{S_t^{\bar{n}}\}_{t \geq 0}$ taking values in $\mathbb{S}_{\bar{n}} = \{s_1, s_2, \dots, s_{\bar{n}}\}$ - the finite state-space set, and having its dynamics “close” to those of S_t . Then, $S_t^{\bar{n}}$ will be used in approximating quantities involving the original process S_t , such as expected values of path functionals. For the Markov chain $S_t^{\bar{n}}$, its transitional dynamics are described by the *rate matrix* $\mathbf{Q} = [q_{ij}]_{\bar{n} \times \bar{n}} \in \mathbb{R}^{\bar{n} \times \bar{n}}$, whose elements q_{ij} satisfy the q -property: (i) $q_{ii} \leq 0$, $q_{ij} \geq 0$ for $i \neq j$, and (ii) $\sum_j q_{ij} = 0$, $\forall i = 1, 2, \dots, \bar{n}$. In terms of q_{ij} 's, the transitional probability of the CTMC $S_t^{\bar{n}}$ is given by:

$$\mathbb{P}(S_{t+\Delta t}^{\bar{n}} = s_j | S_t^{\bar{n}} = s_i, S_{t'}^{\bar{n}}, 0 \leq t' \leq t) = \delta_{ij} + q_{ij}\Delta t + o((\Delta t)^2), \quad (9)$$

where in the above expression δ_{ij} denotes the Kronecker delta. In particular, the transitional matrix is represented in the form of a matrix exponential:

$$\mathbf{P}(\Delta t) = \exp(\mathbf{Q}\Delta t) = \sum_{k=0}^{\infty} (\mathbf{Q}\Delta t)^k / (k!), \quad \Delta t > 0. \quad (10)$$

Here the finite set $\mathbb{S}_{\bar{n}}$, which is the state space of the CTMC $\{S_t^{\bar{n}}\}_{t \geq 0}$, is carefully chosen such that the state space of S_t is sufficiently covered. Details on how to choose the grid points $s_1, s_2, \dots, s_{\bar{n}}$ are given in Section 2.3. In addition, the construction must guarantee that $S_t^{\bar{n}}$ weakly converges to its continuous counterpart S_t under appropriate technical conditions. This is particularly helpful since it guarantees that the desired expected values of well behaved path functionals converge to the true values as the grid points are made denser in the space of S_t .

To this end, for each $i \in \{1, 2, \dots, \bar{n} - 1\}$ define $k_i := v_{i+1} - v_i$, and let $\mu^+(\mu^-)$ denote respectively the positive (negative) part of the function μ . A non-uniform finite discretization of $\mathcal{L}H(x)$ in (8) is given by:

$$\begin{aligned}
 \mu(s_i) & \left(\frac{-k_i}{k_{i-1}(k_{i-1}+k_i)} H(s_{i-1}) + \frac{k_i-k_{i-1}}{k_i k_{i-1}} H(s_i) + \frac{k_{i-1}}{k_i(k_{i-1}+k_i)} H(s_{i+1}) \right) \\
 & + \frac{\sigma^2(s_i)}{2} \left(\frac{2}{k_{i-1}(k_{i-1}+k_i)} H(s_{i-1}) - \frac{2}{k_{i-1}k_i} H(s_i) + \frac{2}{k_i(k_{i-1}+k_i)} H(s_{i+1}) \right) \\
 & = q_{i,i-1} H(s_{i-1}) + q_{i,i} H(s_i) + q_{i,i+1} H(s_{i+1}) =: \mathcal{L}^n H(s). \tag{11}
 \end{aligned}$$

where $q_{i,j}$'s are chosen as in [49], which is recalled here

$$q_{ij} = \begin{cases} \frac{\mu^-(s_i)}{k_{i-1}} + \frac{\sigma^2(s_i) - (k_{i-1}\mu^-(s_i) + k_i\mu^+(s_i))}{k_{i-1}(k_{i-1}+k_i)}, & \text{if } j = i-1, \\ \frac{\mu^+(s_i)}{k_i} + \frac{\sigma^2(s_i) - (k_{i-1}\mu^-(s_i) + k_i\mu^+(s_i))}{k_i(k_{i-1}+k_i)}, & \text{if } j = i+1, \\ -q_{i,i-1} - q_{i,i+1}, & \text{if } j = i, \\ 0, & \text{if } j \neq i-1, i, i+1. \end{cases} \tag{12}$$

Here $\mathbf{k} := \{k_1, k_2, \dots, k_{\bar{n}-1}\}$ is chosen such that

$$0 < \max_{1 \leq i \leq \bar{n}-1} \{k_i\} \leq \min_{1 \leq i \leq \bar{n}} \left\{ \frac{\sigma^2(s_i)}{|\mu(s_i)|} \right\}.$$

With this choice of k_i 's, $\mathbf{Q} = [q_{ij}]_{\bar{n} \times \bar{n}}$ is a tridiagonal matrix. Moreover, we have

$$\begin{aligned}
 \sigma^2(s_i) & \geq \max_{1 \leq i \leq \bar{n}-1} \{k_i\} \cdot |\mu(s_i)| \geq \max_{1 \leq i \leq \bar{n}-1} \{k_i\} \cdot (\mu^+(s_i) + \mu^-(s_i)) \\
 & \geq k_{i-1} \mu^-(s_i) + k_i \mu^+(s_i). \tag{13}
 \end{aligned}$$

As a result, the q -property is satisfied: $q_{ij} \geq 0, \forall 1 \leq i \neq j \leq \bar{n}$, and $\sum_{j=1}^{\bar{n}} q_{ij} = 0, i = 1, \dots, \bar{n}$. In addition, we have the following property regarding the diagonalizability of the generator matrix $\mathbf{Q} = [q_{ij}]_{\bar{n} \times \bar{n}}$.

Theorem 1. (Diagonalization [15]) *The tridiagonal matrix \mathbf{Q} defined in (12) is diagonalizable. In addition, \mathbf{Q} has exactly \bar{n} simple real eigenvalues satisfying $0 \geq \lambda_1 > \lambda_2 > \dots > \lambda_{\bar{n}}$. Hence, the transitional matrix $\mathbf{P}(t)$ has the following decomposition:*

$$\mathbf{P}(\Delta t) = \mathbf{\Gamma} e^{\mathbf{D}_0 \Delta t} \mathbf{\Gamma}^{-1} \quad \text{with} \quad \mathbf{Q} = \mathbf{\Gamma} \mathbf{D}_0 \mathbf{\Gamma}^{-1}, \tag{14}$$

where $\mathbf{D}_0 := \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{\bar{n}})$ is a diagonal matrix of the eigenvalues of \mathbf{Q} , $\mathbf{\Gamma} = (\gamma_{ij})_{i,j=1,\dots,\bar{n}}$ is a matrix whose columns are the corresponding eigenvectors, and we write $\mathbf{\Gamma}^{-1} = (\tilde{\gamma}_{ij})_{i,j=1,\dots,\bar{n}}$.

Furthermore, under some appropriate conditions, it can be shown that $S_t^{\bar{n}}$ converges weakly to S_t as $\bar{n} \rightarrow \infty$. More specifically, there is the following result.

Theorem 2. (Weak convergence [54]) *Let S be a Feller process whose infinitesimal generator \mathcal{L} does not vanish at zero and infinity. Let $S_t^{\bar{n}}$ be the continuous time Markov chain with the generator given in (11). Assume that $\max_{s \in \mathbb{S}_{\bar{n}}} |\mathcal{L}H(s) -$*

$\mathcal{L}^{\bar{n}}H(s) \rightarrow 0$ as $\bar{n} \rightarrow \infty$ for all functions H in the core of \mathcal{L} and $\lim_{s \rightarrow 0^+} \mathcal{L}H(s) = 0$, then $S_t^{\bar{n}}$ converges weakly to S_t as $\bar{n} \rightarrow \infty$. That is, $\mathbb{E}[H(S_T^{\bar{n}})|S_0] \rightarrow \mathbb{E}[H(S_T)|S_0]$ for all bounded continuous functions H .

As an illustration, consider the value of an European option $e^{-rT} \mathbb{E}[H(S_T)|S_0]$ with the payoff function $H(x) = (x - K)^+$ for a call option and $H(x) = (K - x)^+$ for a put option. Assume that $S_0 = s_i \in \mathbb{S}_{\bar{n}}$, i.e., the initial value of the stock price belongs to the state space of the CTMC, for $1 \leq i \leq \bar{n}$ let \mathbf{e}_i denote the column vector of size \bar{n} having the value 1 on the i -th entry and 0 elsewhere; \mathbf{e}'_i denotes the transpose of \mathbf{e}_i . We have the following two results for the applications respectively to path-independent and path-dependent options; more details can be found in [15].

Theorem 3. (European Option) *The value of a European option written on S_T can be approximated by*

$$\mathbb{E}[e^{-rT} H(S_T)|S_0 = s_i] \approx e^{-rT} \mathbf{e}'_i \exp(\mathbf{Q}T) \mathbf{H}(S_T^{\bar{n}}),$$

where $\mathbf{H}(S_T^{\bar{n}})$ is an $\bar{n} \times 1$ vector whose j th entry is given by $H(s_j)$.

Theorem 4. (Bermudan Option) *Let $\Delta = T/M$, where M is the number of monitoring dates, and assume that $S_0 = s_i$. The approximate value of a Bermudan option with monitoring dates $t_0 < t_1 < \dots < t_M$ is evaluated recursively by*

$$\begin{cases} B_M = \mathbf{H}(S_T^{\bar{n}}), \\ B_m = \max\{e^{-r\Delta} \mathbf{e}'_i \exp(\mathbf{Q}\Delta) B_{m+1}, \mathbf{H}(S_T^{\bar{n}})\}, m = M-1, M-2, \dots, 0. \end{cases} \quad (15)$$

2.3 Grid and Boundary Design

In this section, we shall describe the detailed construction of the grids, and hence the state space of the approximating CTMC. As previously mentioned, there are a few ways to generate the grid points, for example, a uniform grid can be constructed from two pre-chosen left and right boundary values s_1 and $s_{\bar{n}}$, and then inserting equally-spaced grid points in between. However, intuitively the uniform grid should not perform very well, and the reason is that it may not be equally likely for the stochastic process to visit each point in its state space. For example, consider the CIR process, which is mean-reverting, and by its mean-reverting property it tends to revert to its mean level either from above or below in equilibrium. Thus it is more likely for the CIR process to visit its long-term mean level rather than the two boundary points. This indicates that we shall insert more grid points around places in the state space that are more often visited, i.e., there are dense clusters of grid points in the state space, and in general this leads to a non-uniform grid.

In the following, we construct a non-uniform grid by carefully choosing the terminal values s_1 and $s_{\bar{n}}$ so that the state space of S_t is sufficiently covered and we manage to place more points around the important values (for example, around S_0). The choice of s_1 and $s_{\bar{n}}$ depends on the boundary condition of the diffusion process.

Assume that the state space of the diffusion is given by $\mathcal{S} = (e_1, e_2)$, then we usually take $s_1 = e_1$ and $s_{\bar{n}} = e_2$. For example, in the CIR model, $\mathcal{S} = [0, \infty)$, and we take $s_1 = 0$ and $s_{\bar{n}} = L$, where L is chosen sufficiently large. Note that the detailed classification of the exact properties of the two boundaries (e.g. as inaccessible, exit or regular) does not impact our choices of s_1 and $s_{\bar{n}}$. One advantage of choosing s_1 and $s_{\bar{n}}$ according to the boundaries of S_t is that we can guarantee that the approximating CTMC has the same boundary for its state space. This indicates one clear advantage of the CTMC approximation method over time-discretization methods such as the Euler time-stepping method. It is well-known in the literature (see [52]) that the Euler time-stepping may yield a boundary bias that is hard to quantify. The reason behind this is that the approximating process from the Euler time-stepping may no longer have similar boundary behaviors as the original process.

After we have fixed the left and right boundaries of the grid, what remains is to determine the spacing of the grid points. To this end, define two constants² $\gamma > 0$ and $\bar{s}^\epsilon > 0$, then we fix $t = T/2$, and center the grid about the mean of the process S_t by: $s_1 := \max\{\bar{s}^\epsilon, \bar{\mu}(t) - \gamma\bar{\sigma}(t)\}$ if the domain of S_t is positive; otherwise $s_1 := \bar{\mu}(t) - \gamma\bar{\sigma}(t)$. We next choose $s_{\bar{n}} := \bar{\mu}(t) + \gamma\bar{\sigma}(t)$, and here we have defined $\bar{\mu}(t) := \mathbb{E}[S_t|S_0]$ and $\bar{\sigma}(t)$ as the standard deviation conditional³ on S_0 . Finally, we generate $s_2, s_3, \dots, s_{\bar{n}-1}$ using the following procedure:

$$s_i = S_0 + \bar{\alpha} \sinh\left(c_2 \frac{i}{\bar{n}} + c_1 \left(1 - \frac{i}{\bar{n}}\right)\right), \quad i = 2, 3, \dots, \bar{n} - 1,$$

where

$$c_1 = \operatorname{arcsinh}\left(\frac{s_1 - S_0}{\bar{\alpha}}\right), \quad c_2 = \operatorname{arcsinh}\left(\frac{s_{\bar{n}} - S_0}{\bar{\alpha}}\right)$$

for $\bar{\alpha} < (s_{\bar{n}} - s_1)$. This transformation concentrates more grid points near the critical point S_0 , where the magnitude of non-uniformity of the grid is controlled by the parameter $\bar{\alpha}$. More specifically, a smaller $\bar{\alpha}$ results in a more nonuniform grid. For numerical computations later, we choose $\bar{\alpha} = (s_{\bar{n}} - s_1)/5$. Since S_0 is not likely a member of the variance grid, we can find the bracketing index j_0 such that $s_{j_0} \leq S_0 < s_{j_0+1}$. Holding the points s_1, s_2 constant⁴, we then shift the remaining points $s_j, j \geq 2$ by $S_0 - s_{j_0}$ so that $s_{j_0} = S_0$ is now a member of the adjusted grid.

For an illustration, in Figure 1 we consider the case $S \in [s_1, s_{60}] = [0, 25]$ and $S_0 = 12.5$. A non-uniform grid of size $\bar{n} = 60$ is formed using the procedure described above. Recall that the non-uniformity of the grid is controlled by the parameter $\bar{\alpha}$: the smaller the value of $\bar{\alpha}$, the more points are placed densely around S_0 , which is evident from the plot of Figure 1. It is noted that non-uniform grid has been used extensively in the literature, for example, it has been utilized in forming the PDE grid

² We can increase γ to make it large enough to sufficiently cover the domain of v_t . From numerical experimentation, we find that $\gamma = 4.5$ and $\bar{s}^\epsilon = 0.00001$ are sufficient for the models considered in this work.

³ If moments of the variance process are unknown, the grids can be fixed using $s_1 = \beta_1 S_0$ and $s_{\bar{n}} = \beta_2 S_0$. For example, we can take $\beta_1 = 10^{-3}$ and $\beta_2 = 4$.

⁴ This keeps an ‘‘anchor’’ at the boundary in the case where $S_0 \approx 0$.

(see [63]). Non-uniform grids have also been used in options pricing, for example in [42] the authors show that non-uniform grid is more favorable as compared to the uniform grid and helps to improve the rate of convergence. A 2-dimensional plot is considered in Figure 2.

Rigorous convergence and error analyses for the CTMC approximation method to option pricing, and the optimal grid design are provided respectively in [47, 70], to which we refer the reader for more details. Regarding the prices and greeks (delta and gamma) of continuously-monitored barrier options, we briefly summarize the main findings in [70]:

1. If there is no grid coinciding with the barrier level, then the convergence can only be of first order.
2. If the barrier level is part of the grid points, then the convergence is of second order for call/put type payoffs. For digital type payoffs, it is in general of first order unless the strike is exactly at the middle two grid points, in which case there is second order convergence.
3. To summarize, there are the following two conditions necessary and sufficient for achieving second order convergence for both prices and greeks:
 - A grid point falls exactly at the barrier level;
 - The strike price is exactly at the middle of two grid points.

Non-uniform grids that satisfy the two conditions in the third item above can be easily constructed. In particular, the authors of [70] propose a class of piecewise uniform grids fulfilling these two conditions that further remove convergence oscillations. Hence, Richardson extrapolation can be applied to accelerate convergence to the third order.

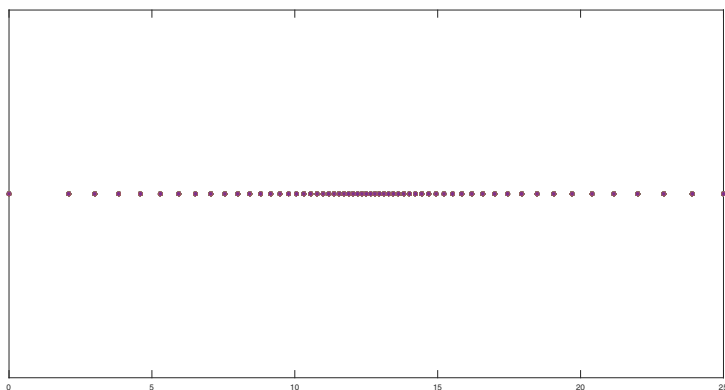


Fig. 1 Nonuniform grid plot with $S \in [0, 25]$ and $\bar{n} = 60, S_0 = 12.5, \bar{\alpha} = 25/15$.

2.4 Relation to PDE

The Markov chain approximation corresponds to a state discretization, and it is of interest to see its connection to discretization methods of PDEs. In [70, 47], the connection is established between the Markov chain approximation techniques and the numerical solution to classic PDEs (see also [54]). Consider the time-homogeneous diffusion as in (1), and with natural boundaries⁵ $-\infty \leq e_1 < e_2 \leq \infty$. Given a well-behaved payoff $H(\cdot)$ (for example, continuous on (e_1, e_2)), the expected value

$$u(t, x) = \mathbb{E}_x[H(S_t)]$$

satisfies the following partial differential equation (PDE)

$$\begin{aligned} \partial_t u(t, s) &= \mu(s)\partial_s u(t, s) + \frac{1}{2}\sigma^2(s)\partial_{ss}u(t, s), \quad t > 0, \quad s \in (e_1, e_2), \\ u(0, s) &= H(s), \quad s \in (e_1, e_2). \end{aligned} \quad (16)$$

Recall that the continuous process S_t is approximated by the CTMC $S_t^{\bar{n}}$ with state space $\mathbb{S}_{\bar{n}} := \{s_i\}_{i=1}^{\bar{n}}$. For ease of exposition, we assume (without loss of generality) a uniform step size $k \equiv k_i = s_i - s_{i-1}$. A semi-discrete approximation is made using a central difference discretization in the space of y :

$$\begin{aligned} &\mu(s)\partial_s u(t, s) + \frac{1}{2}\sigma^2(s)\partial_{ss}u(t, s) \\ &\approx \mu(s_i)\frac{u(t, s_{i+1}) - u(t, s_{i-1})}{2k} + \frac{1}{2}\sigma^2(s_i)\frac{u(t, s_{i+1}) - 2u(t, s_i) + u(t, s_{i-1}))}{k^2}, \end{aligned} \quad (17)$$

with appropriate boundary conditions (e.g. reflecting, killing, or absorbing).

Let $u_k(t)$ be the approximate option price based on the discretization in (17). Then from the work of [54, 47] we have that $u_k(t)$ satisfies the following (matrix-valued) ordinary differential equation (ODE):

$$\frac{d}{dt}u_k(t) = \mathbf{Q}u_k(t), \quad u_k(0) = \mathbf{H}_k, \quad (18)$$

where u_k and $\mathbf{H}_k = [H(s_1), \dots, H(s_{\bar{n}})]^\top$ are $\mathbb{R}^{\bar{n}}$ column vectors, and \mathbf{Q} is the $\bar{n} \times \bar{n}$ tridiagonal generator matrix given (12) with a constant step size k . We can solve the ODE and represent the solution as matrix exponential:

$$u_k(t) = e^{\mathbf{Q}t}\mathbf{H}_k. \quad (19)$$

Define $\pi_k g(\cdot) = (g(y_1), g(y_2), \dots, g(y_{\bar{n}}))^T$, and $\|A\|_\infty = \max_{i,j} |A_{i,j}|$, and consider the option written on the underlying process S_t . Let $u(\cdot)$ and $u_k(\cdot)$ denote respec-

⁵ Note that we make the same assumptions (e.g. Assumption 2.1 to 2.3) as in [47], to which we refer the reader for more details.

tively the true value and the approximate option value written on S_t and $S_t^{\bar{n}}$, then we have the following result:

Theorem 5. ([47]) *Suppose that H is piece-wise twice continuously differentiable (i.e., there are only a finite number of points in (e_1, e_2) where this is not true) and that at any non-differentiable point s , there exists some $\delta_s > 0$ such that H is Lipschitz continuous in $(s - \delta_s, s + \delta_s)$. Consider $k \in (0, \epsilon)$, where ϵ is sufficiently small such that $\epsilon \leq \delta_s$ for all non-differentiable points s . For any $t > 0$, there is some constant $C_t > 0$ independent of k such that*

$$\| u_k(t) - \pi_k u(t, \cdot) \|_{\infty} \leq C_t k^2. \tag{20}$$

This establishes the second order convergence from the approximate solution to the true solution.

2.5 Additive Functionals and Exotic Options

One of the benefits of the CTMC framework is the availability of closed-form pricing formulas given the relative simplicity of a finite state Markov process. Recall the transitional rate matrix \mathbf{Q} and the probability transitional matrix \mathbf{P} given in Section 2.2. For a function $h : \mathbb{R} \rightarrow \mathbb{R}$, define a diagonal matrix $\mathbf{D} := \text{diag}(d_{jj})_{\bar{n} \times \bar{n}}$ with $d_{jj} = h(s_j)$, $j = 1, \dots, \bar{n}$. The following Proposition 1 is concerned with the Laplace transforms of discrete and continuous additive functionals defined therein. In the following, we use M to denote number of observation points.

Proposition 1. ([14]) *Define the additive functionals $B_M^{\bar{n}}$ and $A_t^{\bar{n}}$ for the CTMC $S_t^{\bar{n}}$ by:*

$$B_M^{\bar{n}} := \sum_{m=0}^M h(S_{t_m}^{\bar{n}}), \quad k \geq 0, \quad A_t^{\bar{n}} := \int_0^t h(S_u^{\bar{n}}) du, \quad t \geq 0. \tag{21}$$

- (i) *Discrete case:* $g_d(M; x) := \mathbb{E}_x \left[e^{-\theta B_M^{\bar{n}}} \right] = (e^{-\theta \mathbf{D} \mathbf{P}(\Delta)})^M e^{-\theta \mathbf{D} \mathbf{1}}$, where $\Delta := t/M$.
- (ii) *Continuous case:* $g_c(t; x) := \mathbb{E}_x \left[e^{-\theta A_t^{\bar{n}}} \right] = e^{(\mathbf{G} - \theta \mathbf{D})t} \mathbf{1}$.

Consider the the following functions which are related to Asian options:

$$v_d(M, K, x) = \mathbb{E}_x[(B_M^{\bar{n}} - K)^+], \quad v_c(t, K, x) = \mathbb{E}_x[(A_t^{\bar{n}} - K)^+],$$

where $B_M^{\bar{n}}$ and $A_t^{\bar{n}}$ are defined in (21).

Theorem 6. (Laplace transform of Asian option [14])

- (i) *Discrete case:* Let $l_d(M, \theta, x) := \int_0^{\infty} e^{-\theta k} v_d(M, k, x) dk$. Then for any complex number θ satisfying $\text{Re}(\theta) > 0$, we have

$$l_d(M, \theta, x) = \frac{1}{\theta^2} \left(e^{-\theta D} \exp(\mathbf{Q}\Delta) \right)^M e^{-\theta \mathbf{D}} \mathbf{1} - \frac{1}{\theta^2} \mathbf{1} + \frac{x}{\theta} \frac{1 - e^{(M+1)r\Delta}}{1 - e^{r\Delta}},$$

where $\mathbf{1}$ is the $\bar{n} \times 1$ vector with all entries equal to 1.

(ii) *Continuous case:* Let $l_c(t, \theta, x) = \int_0^\infty e^{-\theta k} v_c(t, k, x) dk$. Then for nay complex number θ satisfying $Re(\theta) > 0$, we have

$$l_c(t, \theta, x) = \frac{1}{\theta^2} e^{(\mathbf{Q} - \theta \mathbf{D})t} \mathbf{1} - \frac{1}{\theta^2} \mathbf{1} + \frac{x}{r\theta} (e^{rt} - 1).$$

We note that the value of a discretely monitored Asian call option is given by $\frac{e^{-rt}}{M+1} v_d(M, (M+1)K, x)$, and similarly for a continuously monitored Asian call option. The results from Theorem 6 can be combined with numerical inverse Laplace transform techniques (see [1]) to price an arithmetic Asian option numerically.

Remark 1. There is a recent ground-breaking paper ([8]) that obtains the *double* transforms for the valuation of discretely-monitored and continuously-monitored arithmetic Asian options in the case when the underlying follows a CTMC. Later, [18] managed to reduce the double transforms therein to a *single* Laplace transform, which yields improved numerical performance. The topic on valuation of Asian options under different model dynamics has been of interest as reflected in recent literature, see [42, 25, 41, 10, 37, 11].

3 Regime Switching Approximations

For some applications in finance, such as volatility modeling, it is often the case that a multi-factor stochastic model is needed. One representative example is the stochastic volatility model, in which both the stock price process and the stochastic variance process (latent process that is not directly observable) are following diffusion processes. Due to the leverage effect documented in the equity market, there is usually a negative correlation between the stock price diffusion process and the variance diffusion process. It has been a challenge to decouple the non-zero correlation between the stock price and the volatility when designing approximation schemes. For example, it is a challenging task when carrying out Euler discretizations to the system of SDEs in a stochastic volatility model (see [58]).

Previous literature mostly considers the CTMC approximation of a one dimensional diffusion process, and here we shall describe a recent approach, which is developed in a series of papers ([12, 13, 14]), that has expanded the approach from univariate processes to cover multi-factor dynamics. It is based on a regime switching approximation to the stochastic volatility models, and the key insight is to simplify the dynamics in such a way that a regime switching approximation can be applied.

3.1 Markov Modulated Dynamics

Regime switching or *Markov modulated* models are a natural extension of the dynamics in (1), allowing for state dependent drift and volatility coefficients. Here the underlying (modulating) state is governed by a CTMC, $\{\alpha(t)\}_{t \geq 0}$, which takes values in $\mathcal{M} := \{1, 2, \dots, \bar{m}\}$, and is specified by its generator matrix or rate matrix, $\mathbf{A} = [\lambda_{ij}]_{\bar{m} \times \bar{m}}$. We denote the underlying process, which is being modulated, by $S_t^{\bar{m}}$. We model the log return process $X_t^{\bar{m}} := \log(S_t^{\bar{m}}/S_0^{\bar{m}})$, $t \in [0, T]$ by

$$dX_t^{\bar{m}} = \mu_{\alpha(t)} dt + \sigma_{\alpha(t)} dW^*(t), \quad (22)$$

where W_t^* is a standard Brownian motion independent of $\alpha(t)$, $\mu_{\alpha(t)} := r - q - \frac{1}{2}\sigma_{\alpha(t)}^2$. In particular, regime changes coincide with changes in the state of $\alpha(t)$. Between state transitions, the asset price is governed by a standard diffusion process with constant drift and volatility coefficients. This corresponds to the following model for the underlying:

$$dS_t^{\bar{m}} = S_t^{\bar{m}}(r - q)dt + S_t^{\bar{m}}\sigma_{\alpha(t)}dW^*(t). \quad (23)$$

An important property of regime-switching models is that the characteristic function (ChF) of the log-return process is available in closed-form. In particular, define the set of functions

$$\psi_j(\xi) = i\xi\mu_j - \frac{1}{2}\xi^2\sigma_j^2, \quad j = 1, \dots, \bar{m}, \quad (24)$$

which represents the characteristic exponents of $X_t^{\bar{m}}$ when each of the states is fixed.

Lemma 1. ([7]) *For $t > 0$, the characteristic function of $X_t^{\bar{m}}$ is given by the following matrix form*

$$\mathbb{E}[\exp(iX_t^{\bar{m}}\xi) | \alpha(0) = j_0] = \mathbf{e}'_{j_0} \exp(t(\mathbf{A} + \text{diag}(\psi_1(\xi), \dots, \psi_{\bar{m}}(\xi)))) \mathbf{1}, \quad (25)$$

where $\mathbf{1} \in \mathbb{R}^{\bar{m}}$ is a unit (column) vector, and $\mathbf{e}_{j_0} \in \mathbb{R}^{\bar{m}}$ is a vector of zeros, except for the value 1 in the position $\alpha(0) = j_0$.

Our treatment of regime-switching models has been necessarily brief. There are several excellent further references including the following: [34, 50, 51, 56, 57, 65, 69]. For a comprehensive treatment of regime switching models, and in particular regime switching diffusion processes and their applications, please refer to monographs [66, 67, 68].

3.2 Stochastic Volatility

Consider the stochastic volatility model whose dynamics are of the following form:

$$\begin{cases} \frac{dS_t}{S_t} = \gamma(v_t)dt + \varkappa(v_t)dW_t^{(1)}, \\ dv_t = \mu_v(v_t)dt + \sigma_v(v_t)dW_t^{(2)}, \end{cases} \tag{26}$$

where $\mathbb{E}[dW_t^{(1)}dW_t^{(2)}] = \rho dt$ with $\rho \in (-1, 1)$ denoting the correlation level between asset and volatility. For the model considered in (26), we assume that there exists a constant $C > 0$ such that for all v_1, v_2 in the state space of v_t

$$|\mu_v(v_1) - \mu_v(v_2)| + |\sigma_v(v_1) - \sigma_v(v_2)| \leq C|v_1 - v_2|, \quad (\mu_v(v_1))^2 + (\sigma_v(v_1))^2 \leq C(1 + v_1^2).$$

The above conditions guarantee that there exists a unique solution v_t possessing the strong Markov property (see [26]). Moreover, we assume that $\sigma_v(\cdot)$ and $\varkappa(\cdot)$ are continuously differentiable, with $\sigma_v(\cdot) > 0$ on the domain of v_t .

We sometimes call this model the “linear” stochastic volatility model since the stock price dynamic is linear in the stock price state variable S . The model (26) is very general, and encompasses many well-known SV models in the literature. A representative list of common SV models can be found in Table 1.

Heston (31)	$dS_t = rS_t dt + \sqrt{V_t}S_t dW_t^{(1)}$ $dV_t = \eta(\theta - V_t)dt + \alpha\sqrt{V_t}dW_t^{(2)}$	$r \in \mathbb{R}$ $\eta, \theta, \alpha, v_0 > 0$
3/2 (45)	$dS_t = rS_t dt + \sqrt{V_t}S_t dW_t^{(1)}$ $dV_t = V_t[\eta(\theta - V_t)dt + \alpha\sqrt{V_t}dW_t^{(2)}]$	$r \in \mathbb{R}$ $\eta, \theta, \alpha, v_0 > 0$
4/2 (27)	$dS_t = rS_t dt + S_t[\alpha\sqrt{V_t} + b/\sqrt{V_t}]dW_t^{(1)}$ $dV_t = \eta(\theta - V_t)dt + \alpha\sqrt{V_t}dW_t^{(2)}$	$r \in \mathbb{R}$ $a, b, \eta, \theta, \alpha, v_0 > 0$
Hull-White (32)	$dS_t = rS_t dt + \sqrt{V_t}S_t dW_t^{(1)}$ $dV_t = \alpha V_t dt + \beta V_t dW_t^{(2)}$	$r \in \mathbb{R}$ $\beta, v_0 > 0$
Stein-Stein (62)	$dS_t = rS_t dt + V_t S_t dW_t^{(1)}$ $dV_t = \eta(\theta - V_t)dt + \beta V_t dW_t^{(2)}$	$r \in \mathbb{R}$ $\beta, v_0 > 0$
α -Hypergeometric (23)	$dS_t = rS_t dt + e^{V_t} S_t dW_t^{(1)}$ $dV_t = (\eta - \theta e^{\alpha V_t})dt + \beta V_t dW_t^{(2)}$	$r \in \mathbb{R}$ $\beta, v_0 > 0$
Jacobi (2)	$dS_t = (r - V_t/2)dt + \sqrt{V_t - \rho^2 Q(V_t)}dW_t^{(1)}$ $dV_t = (\eta - \theta e^{\alpha V_t})dt + \beta\sqrt{Q(V_t)}dW_t^{(2)}$	$r \in \mathbb{R}$ $\beta, v_0 > 0$

Table 1 Some stochastic volatility models. For Jacobi model, we have $Q(v) := (v - v_{\min})(v_{\max} - v)/(\sqrt{v_{\max}} - \sqrt{v_{\min}})^2$.

In the next subsection, we seek to single out the correlation ρ and decouple the SDE system in (26).

3.2.1 Decoupled Dynamics

Options pricing in a general stochastic volatility model is notoriously difficult due to the general correlation structure between the two driving Brownian motions $W_t^{(1)}$ and $W_t^{(2)}$. In this section, we will provide a general procedure to decouple the correlation between the two Brownian motions. From the Ito’s lemma, we have

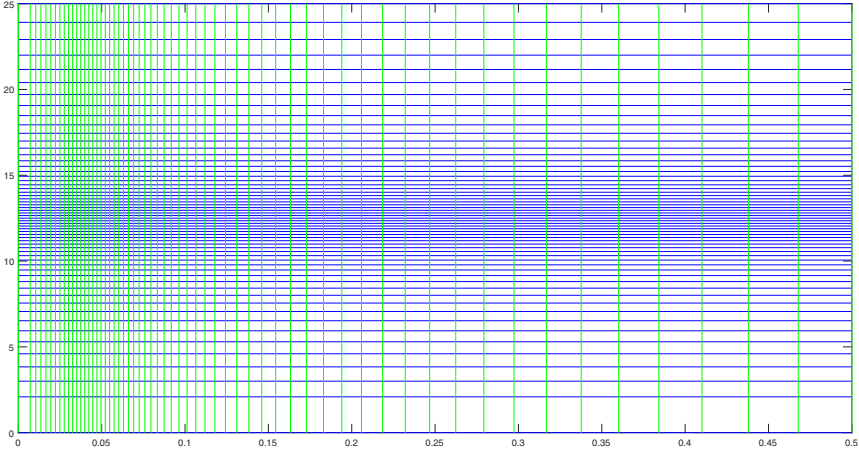


Fig. 2 Nonuniform grid plot with $(S, v) \in [0, 25] \times [0, 0.5]$, $\bar{n} = 60, \bar{m} = 60, S_0 = 12.5, v_0 = 0.05, \bar{\alpha}_S = 25/15, \bar{\alpha}_v = 0.5/15$.

$$d \log(S_t) = \left(\gamma(v_t) - \frac{1}{2} \varkappa^2(v_t) \right) dt + \varkappa(v_t) dW_t^{(1)}. \quad (27)$$

Next, define $\hat{f}(x) := \int_c^x \frac{\varkappa(u)}{\sigma_v(u)} du$ with c being a constant, and

$$h(x) := \mathcal{L}(\hat{f}(x)) = \mu_v(x) \hat{f}'(x) + \frac{1}{2} \sigma_v^2(x) \hat{f}''(x).$$

Denote $f(v_t, v_0) := \rho(\hat{f}(v_t) - \hat{f}(v_0))$, then we have

$$df(v_t, v_0) = \rho d\hat{f}(v_t) = \rho h(v_t) dt + \rho \varkappa(v_t) dW_t^{(2)}. \quad (28)$$

Finally, define $W_t^* := \frac{W_t^{(1)} - \rho W_t^{(2)}}{\sqrt{1 - \rho^2}}$, then one can easily verify that W_t^* is a standard Brownian motion and $\mathbb{E}[dW_t^{(1)*} dW_t^{(2)}] = 0$, i.e., the two Brownian motions W_t^* and $W_t^{(2)}$ are independent. Next, we plug (28) into (27), and obtain

$$\begin{aligned} d \log(S_t) &= \left(\gamma(v_t) - \frac{1}{2} \varkappa^2(v_t) \right) dt + \varkappa(v_t) (\rho dW_t^{(2)} + \sqrt{1 - \rho^2} dW_t^*) \\ &= \left(\gamma(v_t) - \frac{1}{2} \varkappa^2(v_t) \right) dt + df(v_t, v_0) - \rho h(v_t) dt + \sqrt{1 - \rho^2} \varkappa(v_t) dW_t^*. \end{aligned}$$

Denote $\tilde{X}_t := \log(S_t/S_0) - f(v_t, v_0)$, then we can rewrite (26) as

$$\begin{cases} d\tilde{X}_t = (\gamma(v_t) - \frac{1}{2}\varkappa^2(v_t) - \rho h(v_t)) dt + \sqrt{1 - \rho^2}\varkappa(v_t)dW_t^*, \\ dv_t = \mu_v(v_t)dt + \sigma_v(v_t)dW_t^{(2)}. \end{cases} \quad (29)$$

Observe that the two Brownian motions in (29) are independent, and we have successfully decoupled the correlation structure. We will refer to the process \tilde{X}_t which provides the decoupling as the *auxiliary process*.⁶

3.2.2 Regime Switching Approximation: Affine Case

Using the continuous time Markov chain approximation in Section 2, we will convert (29) into a regime switching model. More specifically, for $\bar{m} \in \mathbb{N}^+$ we will approximate the variance process v_t by another independent finite state Markov chain α_t taking values in the state space $\mathcal{M} := \{1, 2, \dots, \bar{m}\}$ with the generator $\mathbf{A} = [\lambda_{ij}]_{\bar{m} \times \bar{m}}$ obtained as in (12) using the dynamics of v_t given in (29). Then the model in (29) is reduced to

$$\begin{aligned} d\tilde{X}_t^{\bar{m}} &= \left(\gamma(v_{\alpha_t}) - \frac{1}{2}\varkappa^2(v_{\alpha_t}) - \rho h(v_{\alpha_t}) \right) dt + \sqrt{1 - \rho^2}\varkappa(v_{\alpha_t})dW_t^*, \\ &=: \mu_X(v_{\alpha_t})dt + \sigma_X(v_{\alpha_t})dW_t^*, \end{aligned} \quad (30)$$

and note that notation-wise $v_t^{\bar{m}} = v_{\alpha_t}$, where v_{α_t} takes values in $\mathbb{S}_v := \{v_1, \dots, v_{\bar{m}}\}$. In particular, the diffusion coefficients depend only on v_{α_t} . As with the regime-switching models discussed in Section 3.1, once \mathbf{A} is determined, $\tilde{X}_t^{\bar{m}}$ can be described by its generator in each state, or equivalently by its set of characteristic exponents

$$\tilde{\Psi}_j(\xi) = i\xi\mu_X(v_j) - \frac{1}{2}\xi^2\sigma_X(v_j)^2, \quad j = 1, \dots, \bar{m}. \quad (31)$$

From Lemma 1, the ChF of $\tilde{X}_t^{\bar{m}}$, $\mathcal{E}(\xi) = [\mathcal{E}_{j,k}]$, is given by

$$\mathcal{E}(\xi) := \mathbb{E}[\exp(i\xi\tilde{X}_t^{\bar{m}}) | \alpha(0) = j] = \mathbf{e}'_j \exp(t(\mathbf{A} + \text{diag}(\tilde{\Psi}_1(\xi), \dots, \tilde{\Psi}_{\bar{m}}(\xi)))) \mathbf{1}.$$

Moreover, we can recover the ChF of the log-return approximation as

$$\begin{aligned} \tilde{\mathcal{E}}_{j,k}(\xi) &:= E[\exp(i\xi \cdot \log(S_t^{\bar{m}}/S_0)) | \alpha(0) = j, \alpha(t) = k] \\ &= \mathcal{E}_{j,k}(\xi) \cdot \exp(i\xi \cdot f(v_k, v_j)), \end{aligned} \quad (32)$$

which follows from the original representation $\tilde{X}_t := \log(S_t/S_0) - f(v_t, v_0)$. The availability of a closed form ChF is a key advantage of the approximation framework, as it enables the use of highly efficient Fourier transform based approaches, which we demonstrate in Section 3.2.3 for barrier options.

⁶ We note that the extension of this procedure to processes with jumps is straightforward.

3.2.3 Recursive Option Pricing under Stochastic Volatility

Provided the decoupled regime-switching dynamics in (30), Lemma 1 provides a closed-form characteristic function, which enables option pricing via Fourier techniques. As an example, consider a barrier option with terminal payoff $G(X_T) = H(S_0 \exp(X_T)) = H(S_T)$, where $X_t = \ln(S_t/S_0)$, and fix a set of monitoring dates $t_m = m\Delta$, $m = 0, \dots, M$, where $\Delta = T/M$. Let \mathcal{C} denote the continuation region, and \mathcal{C}^c the knock-out region for X_t , so the option expires worthless if it is observed within \mathcal{C}^c at any time t_m , and pays $G(X_T)$ otherwise. For a double barrier option with knockout barriers L and U in the space of S_t , $\mathcal{C} = [l_x, u_x]$ where $l_x := \ln(L/S_0)$ and $u_x := \ln(U/S_0)$ in log space.

Barrier options can be priced for the SV model defined in (26) using the Markov chain approximation, which yields

$$\log(S_t^{\bar{m}}/S_0) = \tilde{X}_t^{\bar{m}} + f(v_t^{\bar{m}}, v_0) := X_t^{\bar{m}},$$

from which $S_t^{\bar{m}} = S_0 \exp(X_t^{\bar{m}})$. The barrier option price is calculated through the following recursive procedure, starting from the known terminal values and working backwards:

$$\begin{cases} \mathcal{V}_M(X_M^{\bar{m}}, \alpha_M) = H(X_M^{\bar{m}}) \mathbb{1}_{\{X_M^{\bar{m}} \in \mathcal{C}\}} \\ \mathcal{V}_m(X_m^{\bar{m}}, \alpha_m) = e^{-r\Delta} \mathbb{E} \left[\mathcal{V}_{m+1}(X_{m+1}^{\bar{m}}, \alpha_{m+1}) \mathbb{1}_{\{X_m^{\bar{m}} \in \mathcal{C}\}} | X_m^{\bar{m}}, \alpha_m \right] \quad m = M-1, \dots, 0, \end{cases} \quad (33)$$

where $X_m^{\bar{m}} := X_{t_m}^{\bar{m}}$ and $\alpha_m = \alpha(t_m)$. By definition, $\mathcal{V}_{m+1}(X_{m+1}^{\bar{m}}, \alpha_{m+1}) = 0$ for $X_{m+1}^{\bar{m}} \in \mathcal{C}^c = [l_x, u_x]^c$.

Next define the transition probability matrix $\mathbf{P}(\Delta)$ as in (10), with elements

$$P_{jk}^A = \mathbb{P}[\alpha(t+\Delta) = k | \alpha(t) = j], \quad j, k = 1, \dots, \bar{m},$$

which captures transitions of the volatility state. Then with $\alpha_m = j$ and $X_m^{\bar{m}} = x \in \mathcal{C}$, we have for $m = M-1, \dots, 0$,

$$\begin{aligned} \mathcal{V}_m(x, j) &= e^{-r\Delta} \mathbb{E} [\mathcal{V}_{m+1}(X_{m+1}^{\bar{m}}, \alpha_{m+1}) | X_m^{\bar{m}} = x, \alpha_m = j] \\ &= e^{-r\Delta} \sum_{k=1, \dots, \bar{m}} P_{j,k}^A \mathbb{E} [\mathcal{V}_{m+1}(X_{m+1}^{\bar{m}}, k) | X_m^{\bar{m}} = x, \alpha_m = j, \alpha_{m+1} = k] \\ &= e^{-r\Delta} \sum_{k=1, \dots, \bar{m}} P_{j,k}^A \int_{\mathcal{C}} \mathcal{V}_{m+1}(y, k) p_{j,k}(y|x) dy, \end{aligned}$$

where we have defined the set of transition densities for the log return process for $j, k = 1, \dots, \bar{m}$

$$p_{j,k}(y|x) = \mathbb{P}[X^{\bar{m}}(\Delta) \in y + dy | X^{\bar{m}}(0) = x, \alpha(0) = j, \alpha(\Delta) = k].$$

⁷ A European option can be priced recursively by setting $\mathcal{C} = (-\infty, \infty)$.

As demonstrated in [43], the transition densities $p_{j,k}(y|x)$ can be approximated with high efficiency using the ChF of log returns of $X_{\Delta}^{\bar{m}}, \tilde{\mathcal{E}}_{j,k}(\xi)$, by combing the closed form expression in (32) with the Fourier method of [40].

3.2.4 Example: 4/2 model CTMC Approximation

Many prominent examples fall within the framework of dynamics (26), including those of Heston [31], Hull-White [32], Stein-Stein [62], α -Hypergeometric [23], Jacobi [2], 3/2 [45] and the 4/2 model for which we are going to illustrate in detail. We illustrate the transform required to obtain a de-correlated representation for the 4/2 model. The reader is invited to refer to Table 1 for a list of additional models that can also be similarly considered.

The 4/2 stochastic volatility model (without jumps) was recently proposed in [27], with the important property that the instantaneous volatility can be uniformly bounded away from zero (unlike Heston’s model, for example). It contains the Heston model (can be thought of as a “1/2” model) and the 3/2 model as special cases, and thus earns itself the name of a “4/2” model. Extension of the 4/2 model by adding the jump component in the underlying process can be founded in [42]. The dynamics of the 4/2 model are given by

$$\begin{cases} \frac{dS_t}{S_t} = (r - q - \lambda \kappa)dt + \left[a\sqrt{v_t} + \frac{b}{\sqrt{v_t}} \right] dW_t^{(1)}, \\ dv_t = \eta(\theta - v_t)dt + \sigma_v \sqrt{v_t} dW_t^{(2)}. \end{cases} \quad (34)$$

For this model, it is assumed that the Feller’s condition $2\eta\theta > \sigma_v^2$ is satisfied, and for $a, b > 0$, the volatility component is uniformly bounded away from zero, which follows from applying Cauchy’s inequality to $\left[a\sqrt{v_t} + \frac{b}{\sqrt{v_t}} \right] \geq 2\sqrt{a\sqrt{v_t}\frac{b}{\sqrt{v_t}}} = 2\sqrt{ab} > 0$ for $a, b > 0$. The change of variable, which will help us remove the correlation between the two stochastic processes $W_t^{(1)}, W_t^{(2)}$ in (34), is given by

$$\tilde{X}_t = \log\left(\frac{S_t}{S_0}\right) - \frac{\rho}{\sigma_v} (a(v_t - v_0) + b(\log(v_t) - \log(v_0))). \quad (35)$$

Therefore, if we denote

$$\mu_X(v_t) = \left(\frac{a\rho\eta}{\sigma_v} - \frac{a^2}{2} \right) v_t + \left(\frac{\rho b\sigma_v - b^2}{2} - \frac{b\rho\eta\theta}{\sigma_v} \right) \frac{1}{v_t} + \frac{\rho\eta}{\sigma_v} (b - a\theta) + r - q - \lambda\kappa - ab,$$

then the dynamics in (34) can be written as

$$\begin{cases} d\tilde{X}_t = \mu_X(v_t)dt + \left[a\sqrt{v_t} + \frac{b}{\sqrt{v_t}} \right] \sqrt{(1 - \rho^2)} dW_t^*, \\ dv_t = \eta(\theta - v_t)dt + \sigma_v \sqrt{v_t} dW_t^{(2)}. \end{cases} \quad (36)$$

After approximating the variance process v_t by a \bar{m} -state Markov chain, and substituting it into (36), we have that the dynamics in (30) are reduced to

$$d\tilde{X}_t^{\bar{m}} = \mu_X(v_{\alpha_t})dt + \left[a\sqrt{v_{\alpha_t}} + \frac{b}{\sqrt{v_{\alpha_t}}} \right] \sqrt{(1-\rho^2)}dW_t^*, \tag{37}$$

where v_{α_t} takes values in $\mathbb{S}_v = \{v_1, v_2, \dots, v_{\bar{m}}\}$.

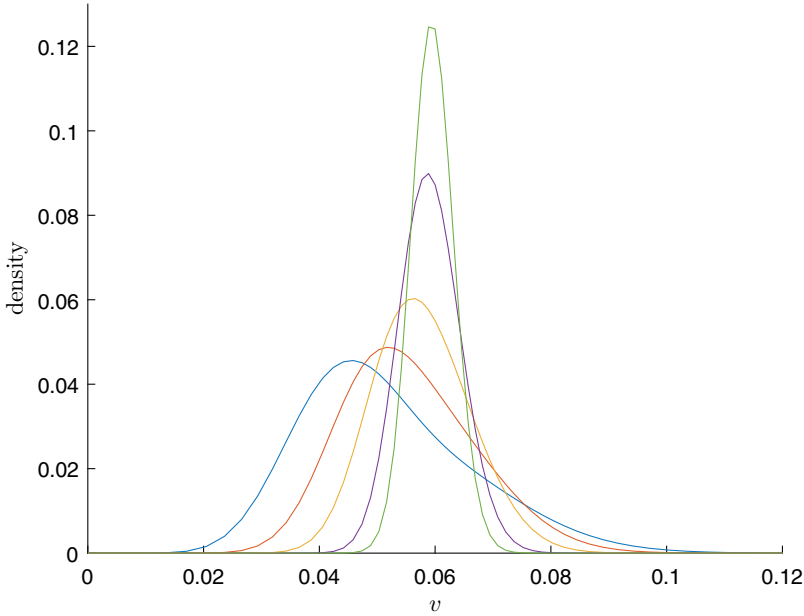


Fig. 3 Conditional transition densities of $\bar{m} = 40$ state CTMC approximation to (CIR) volatility process v_t under 4/2 (and Heston) stochastic volatility, for several values of t .

In Figure 3, we illustrate the CTMC approximation of the underlying variance process, which in the 4/2 (and Heston) model is a Cox-Ingersol-Ross (CIR) process. For $t \in \{1/5, 1/10, 1/20, 1/50, 1/100\}$, we plot the transition density of v_{α_t} , conditional on v_0 , for the CIR process with $\eta = 2, \theta = 0.04, v_0 = 0.06, \rho = -0.9, \sigma_v = 0.15$. In this example, the initial variance $v_0 = 0.06$ is higher than its longer term mean, $\theta = 0.04$. When $t = 1/100$, the density is centered about v_0 , and the diffusive term dominates the transition probabilities, leading to a roughly symmetric (approximately normal) transition density. As time increases up to $t = 1/5$, the densities spread out, with more mass clustering near the long term mean θ . With just $\bar{m} = 40$ points, the densities of the CTMC are a faithful representation of the underlying continuous density of v_t .

3.2.5 Example: Heston and 4/2 model option pricing

We now illustrate the application of the CTMC approximation for option pricing under the 4/2 model discussed in Section 3.2.4, starting with the special case of Heston’s model, which is obtained by setting $a = 1$ and $b = 0$. The recursive pricing strategy outlined in Section 3.2.3 can be used to price European options, and in Heston’s model reference prices can be obtained to machine precision using Fourier techniques (e.g. [40]). The model parameters are set to be

$$\eta = 1, \quad \theta = 0.025, \quad v_0 = 0.025, \quad \rho = -0.7, \quad \sigma_v = 0.18.$$

The state space for the CTMC approximation of v_t is determined as described in Section 2.3, with the grid width parameter $\bar{\alpha} = (v_{\bar{m}} - v_1)/\zeta$ parameterized by $\zeta > 0$. For $\zeta \approx 1$, the grid becomes uniform, while for $\zeta \approx 0$, the grid is tightly clustered around the initial variance v_0 .

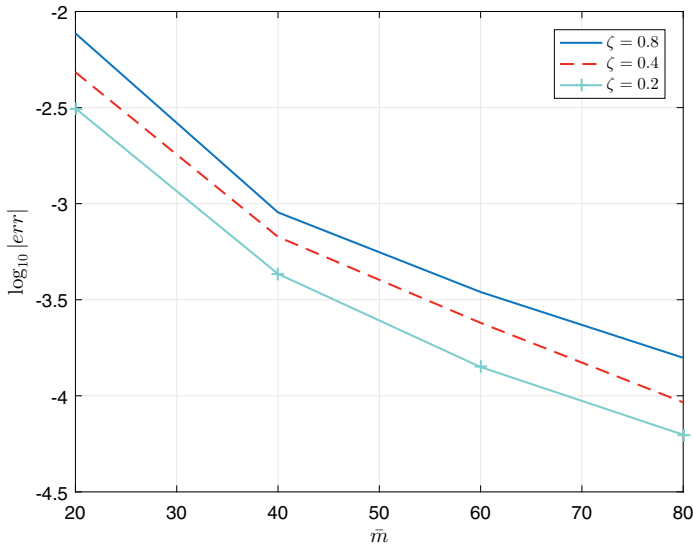


Fig. 4 European option convergence in Heston’s model as a function of grid non-uniformity parameter ζ . $T = 0.5$, $K = S_0 = 100$, $r = 0.05$. Ref price: 5.7574.

Figure 4 illustrates the pricing error for an at-the-money European call option with $\zeta \in \{0.8, 0.4, 0.2\}$. As is typically the case, having a more non-uniform grid is most beneficial when the number of grid points \bar{m} is small. Other factors can also influence this choice in practice, including the long term level of variance and its relation to the initial variance, as well as the time to maturity. For example, a short maturity option with initial variance near its longer term level will benefit the

most from a grid which clusters around v_0 , while as T increases (or equivalently σ_v increases) the benefit diminishes.

In the next set of experiments, we consider the 4/2 model, with base parameters

$$\eta = 1.8, \quad \theta = 0.04, \quad v_0 = 0.04, \quad \rho = -0.7, \quad \sigma_v = 0.1.$$

Table 2 illustrates the convergence in \bar{m} for three 4/2 models, which vary based on the values of a and b . The first case, $a = 1, b = 0$, is simply Heston’s model, while the other two cases have the additional variance term $b/\sqrt{v_t}$ from (36). Reference prices, to which the approximations have converged within four decimal places in the last row ($\bar{m} = 60$), are computed using $\bar{m} = 120$.

\bar{m}	$a = 1, b = 0$		$a = 0.5, b = 0.5v_0$		$a = 0.5, b = 0.25v_0$	
	price	error	price	error	price	error
10	6.9020	1.46e-03	6.9623	5.51e-02	5.5935	5.16e-02
20	6.8999	5.93e-04	6.9066	5.47e-04	5.5414	4.25e-04
40	6.9005	3.68e-05	6.9071	9.98e-05	5.5418	5.36e-05
60	6.9005	2.43e-06	6.9071	3.22e-05	5.5419	1.42e-05

Table 2 European call option prices under 4/2 model. $T = 0.5, K = S_0 = 100, r = 0.05$.

3.3 Regime Switching CTMCs

In Section 3.2, we discussed the use of a CTMC to approximate one dimension of the two-dimensional stochastic volatility model, which resulted in a regime-switching diffusion. Taking this idea one step further, we can consider the case of a Markov modulated CTMC, i.e., a regime switching CTMC. In this case, the \bar{n} -state CTMC $S_t^{\bar{n}}$ is further modulated by a second independent CTMC, $\{\alpha_t\}_{t \geq 0}$, with state space $\mathcal{M} = \{1, \dots, \bar{m}\}$. Then we have a RS-CTMC, denoted as $S_t^{\bar{n}, \bar{m}}$, of the following form:

$$dS_t^{\bar{n}, \bar{m}} = S_t^{\bar{n}, \bar{m}}(r - q)dt + S_t^{\bar{n}, \bar{m}} \sigma_{\alpha(t)} dW^*(t). \tag{38}$$

In particular, conditioned on $\alpha_t = l$, the instantaneous transitions of $S_t^{\bar{n}, \bar{m}}$ can be described by the following generator $\mathcal{G}_l, l \in \mathcal{M}$:

$$\mathcal{G}_l h(x) := \lim_{\delta \downarrow 0} \frac{\mathbb{E} \left[h(S_{t+\delta}^{\bar{n}, \bar{m}}) | \alpha(t) = l, S_t^{\bar{n}, \bar{m}} = x \right] - h(x)}{\delta}, \tag{39}$$

which corresponds to a rate matrix $\mathbf{G}_l = (g_{kj}^l)_{\bar{n} \times \bar{n}}$. An explicit example of \mathbf{G}_l will be given in Section 3.4.2.

In Section 3.4, we consider the applications of regime switching Markov chain approximation to several options pricing problems. In the following, we shall dis-

cuss the details of the regime switching approximation in two types of models in increasing order of generality: the stochastic volatility(SV) model, and the stochastic local volatility(SLV) model. In the SV model, we utilize one approximating CTMC, and for the case of the SLV model, there are two independent approximating CTMCs introduced.

3.4 Stochastic Local Volatility

The proposed valuation framework is also applicable to general stochastic local volatility models whose dynamics are given by:

$$\begin{cases} dS_t = \omega(S_t, v_t)dt + \varkappa(v_t)\Gamma(S_t)dW_t^{(1)}, \\ dv_t = \mu_v(v_t)dt + \sigma_v(v_t)dW_t^{(2)}, \end{cases} \quad (40)$$

where $\mathbb{E}[dW_t^{(1)}dW_t^{(2)}] = \rho dt$ with $\rho \in (-1, 1)$. Here we assume that $\omega(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, $\varkappa(\cdot) : \mathbb{R} \rightarrow \mathbb{R}_+$ and $\Gamma(\cdot) : \mathbb{R} \rightarrow \mathbb{R}_+$. Some representative local volatility models are listed in Table 3. We make the following assumption about the Feller property of (S_t, v_t) .

Assumption 2 For any $\Phi \in C_0(\mathbb{S} \times \mathbb{V})$, define the pricing operator $P_t\Phi(S, v) := \mathbb{E}[\Phi(S_t, v_t)|S_0 = S, v_0 = v]$, and assume that (S_t, v_t) is a Feller process, i.e.,

- $P_t\Phi \in C_0(\mathbb{S} \times \mathbb{V})$ for any $t \geq 0$;
- $\lim_{t \rightarrow 0} P_t\Phi(S, v) = \Phi(S, v)$ for any $(S, v) \in \mathbb{S} \times \mathbb{V}$.

The Feller property guarantees that there exists a *version* of the process (S_t, v_t) with càdlàg paths satisfying the strong Markov property. Similar to the scalar case, the family (P_t) is determined by its infinitesimal generator \mathcal{L}^S , where

$$\mathcal{L}^S\Phi(S, v) := \lim_{t \rightarrow 0^+} \frac{(P_t\Phi - \Phi)(S, v)}{t}, \quad (41)$$

for any $\Phi \in C_0(\mathbb{S} \times \mathbb{V})$ for which the right-hand side of (41) converges in the strong sense.⁸ From (40), we can calculate

$$\begin{aligned} \mathcal{L}^S\Phi &= \frac{[\varkappa(v)\Gamma(S)]^2}{2} \frac{\partial^2\Phi}{\partial S^2} + \rho\varkappa(v)\Gamma(S)\sigma(v) \frac{\partial^2\Phi}{\partial v\partial S} + \\ &+ \frac{\sigma_v^2(v)}{2} \frac{\partial^2\Phi}{\partial v^2} + \omega(S, v) \frac{\partial\Phi}{\partial S} + \mu_v(v) \frac{\partial\Phi}{\partial v}. \end{aligned} \quad (42)$$

For example, for the classical SABR model, see Table 3, the generator is given by

⁸ Convergence is with respect to the norm $\|\Phi\| := \sup_{(s,v) \in \mathbb{S} \times \mathbb{V}} |\Phi(s, v)|$ on the Banach space $(C_0(\mathbb{S} \times \mathbb{V}), \|\cdot\|)$. The domain of \mathcal{L}^S is dense in $C_0(\mathbb{S} \times \mathbb{V})$.

$$\mathcal{L}^S \Phi = \frac{1}{2} v^2 S^{2\beta} \frac{\partial^2 \Phi}{\partial S^2} + \rho \alpha v^2 S^\beta \frac{\partial^2 \Phi}{\partial v \partial S} + \frac{1}{2} (\alpha v)^2 \frac{\partial^2 \Phi}{\partial v^2}.$$

3.4.1 Decoupled Dynamics

Define the functions $g(x) := \int^x \frac{1}{\Gamma(u)} du$ and $\hat{f}(x) := \int^x \frac{\varkappa(u)}{\sigma_v(u)} du$, and let $\tilde{X}_t := g(S_t) - \rho \hat{f}(v_t)$. Then similarly to the stochastic volatility case, the dynamics in (40) can be rewritten as

$$\begin{cases} d\tilde{X}_t = \left(\frac{\omega(S_t, v_t)}{\Gamma(S_t)} - \frac{\Gamma'(S_t)}{2} \varkappa^2(v_t) - \rho h(v_t) \right) dt + \sqrt{1 - \rho^2} \varkappa(v_t) dW_t^*, \\ dv_t = \mu_v(v_t) dt + \sigma_v(v_t) dW_t^{(2)}, \end{cases} \quad (43)$$

where

$$\begin{aligned} h(x) &:= \mathcal{L}^v f(x) = \mu_v(x) \hat{f}'(x) + \frac{1}{2} \sigma_v^2(x) \hat{f}''(x) \\ &= \mu_v(x) \frac{\varkappa(x)}{\sigma_v(x)} + \frac{1}{2} (\sigma_v(x) \varkappa'(x) - \sigma_v'(x) \varkappa(x)). \end{aligned} \quad (44)$$

We shall carry out the approximation procedure in *two layers*: one for the stochastic variance process, and one for the asset price process. The first layer approximation is obtained by replacing v_t with $v_t^{\tilde{m}} = v_{\alpha(t)}$, and we obtain

$$\tilde{X}_t^{\tilde{m}} := g(S_t^{\tilde{m}}) - \rho \hat{f}(v_t^{\tilde{m}}),$$

where $S_t^{\tilde{m}}$ is used to denote the dependence of S_t on $v_t^{\tilde{m}}$. Next, let

$$\zeta_0(\tilde{X}_t^{\tilde{m}}, v_t^{\tilde{m}}) := g^{-1}(\tilde{X}_t^{\tilde{m}} + \rho \hat{f}(v_t^{\tilde{m}})),$$

and for any fixed state $v_l \in \mathbb{S}_v$, define

$$\tilde{\omega}(\cdot, v_l) := \omega(\zeta_0(\cdot, v_l), v_l), \quad \tilde{\Gamma}(\cdot, v_l) := \Gamma(\zeta_0(\cdot, v_l)).$$

We further define:

$$\mu_X(x, v_l) := \left(\frac{\tilde{\omega}(x, v_l)}{\tilde{\Gamma}(x, v_l)} - \frac{\tilde{\Gamma}'(x, v_l)}{2} \varkappa^2(v_l) - \rho h(v_l) \right), \quad (45)$$

where $\tilde{\Gamma}'(\cdot, v_l) = \Gamma'(\zeta_0(\cdot, v_l))$. We manage to obtain the following dynamics for $\tilde{X}_t^{\tilde{m}}$ conditional on the value of $v_t^{\tilde{m}}$:

$$d\tilde{X}_t^{\tilde{m}} = \mu_X(\tilde{X}_t^{\tilde{m}}, v_t^{\tilde{m}}) dt + \sqrt{1 - \rho^2} \varkappa(v_t^{\tilde{m}}) dW_t^*. \quad (46)$$

3.4.2 Regime Switching Approximation: Linear and Nonlinear Case

The insight of [14] is to combine the RS-CTMC representation provided in Proposition 2 with a Markov chain approximation for the decoupled dynamics of the SLV model in (40). From the single layer approximation in (46), for each variance state $l \in \mathcal{M}$, the generator satisfies

$$\mathcal{L}_l^{\bar{m}} \xi(x) = \mu_X(x, v_l) \xi'(x) + \frac{(1 - \rho^2) \varkappa^2(v_l)}{2} \xi''(x). \quad (47)$$

We then make a second layer approximation, similarly as before. In particular, for each $l \in \mathcal{M}$, the rate matrix is given by $\mathbf{G}_l = (g_{kj}^l)$, where

$$g_{kj}^l = \begin{cases} \frac{\mu_X^-(x_k, v_l)}{\delta_{k-1}^x} + \frac{\bar{\sigma}^2(v_l) - [\delta_{k-1}^x \mu_X^-(x_k, v_l) + \delta_k^x \mu_X^+(x_k, v_l)]}{\delta_{k-1}^x (\delta_{k-1}^x + \delta_k^x)}, & j = k - 1, \\ \frac{\mu_X^+(x_k, v_l)}{\delta_k^x} + \frac{\bar{\sigma}^2(v_l) - [\delta_{k-1}^x \mu_X^-(x_k, v_l) + \delta_k^x \mu_X^+(x_k, v_l)]}{\delta_k^x (\delta_{k-1}^x + \delta_k^x)}, & j = k + 1, \\ -q_{k,k-1}^l - q_{k,k+1}^l, & j = k, \\ 0, & |j - k| > 1, \end{cases} \quad (48)$$

where $\bar{\sigma}(v_l) = \sqrt{1 - \rho^2} \varkappa(v_l)$ and $\delta_k^x = x_k - x_{k-1}$ for $k = 1, 2, \dots$

The generator is approximated for $l \in \mathcal{M}, k \in \mathcal{N}$ by

$$\mathcal{L}_l^{\bar{n}, \bar{m}} \xi(x_k) = \sum_{j=1}^{\bar{n}} g_{kj}^l \xi(x_j) = \sum_{j=1}^{\bar{n}} g_{kj}^l (\xi(x_j) - \xi(x_k)). \quad (49)$$

The key insight of the paper [60] is that we can represent the RS-CTMC $S_t^{\bar{n}, \bar{m}}$ as a one-dimensional process, by embedding it into a Markov chain, called Y_t , with an enlarged state space \mathbb{S}_Y , which is defined in the following result. The state space of the two-dimensional process $S_t^{\bar{n}, \bar{m}}$ is mapped bijectively to that of Y_t, \mathbb{S}_Y , by the function $\phi(\cdot)$ defined below. The space \mathbb{S}_Y can be interpreted as indexing \bar{m} consecutive copies of \mathbb{S}_X , one for each of the modulating states $l \in \mathcal{M}$. Thus

Proposition 2. ([60]) *Suppose that $\{S_t^{\bar{n}, \bar{m}}, t \geq 0\}$ is a discrete state regime-switching CTMC, and consider another one-dimensional CTMC $\{Y_t, t \geq 0\}$ with state space $\mathbb{S}_Y := \{1, 2, \dots, \bar{n} \cdot \bar{m}\}$ and $\bar{n} \cdot \bar{m} \times \bar{n} \cdot \bar{m}$ transition rate matrix*

$$\mathbf{G} = \begin{pmatrix} \lambda_{11} \mathbf{I}_{\bar{n}} + \mathbf{G}_1 & \lambda_{12} \mathbf{I}_{\bar{n}} & \cdots & \lambda_{1\bar{m}} \mathbf{I}_{\bar{n}} \\ \lambda_{21} \mathbf{I}_{\bar{n}} & \lambda_{22} \mathbf{I}_{\bar{n}} + \mathbf{G}_2 & \cdots & \lambda_{2\bar{m}} \mathbf{I}_{\bar{n}} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{\bar{m}1} \mathbf{I}_{\bar{n}} & \lambda_{\bar{m}2} \mathbf{I}_{\bar{n}} & \cdots & \lambda_{\bar{m}\bar{m}} \mathbf{I}_{\bar{n}} + \mathbf{G}_{\bar{m}} \end{pmatrix}, \quad (50)$$

where $\mathbf{I}_{\bar{n}}$ is the $\bar{n} \times \bar{n}$ identity matrix, $\mathbf{G}_l = (g_{kj}^l)_{\bar{n} \times \bar{n}}$, and $\mathbf{A} = (\lambda_{k,j})_{\bar{m} \times \bar{m}}$. Define the mapping $\phi : \mathbb{S}_X \times \mathcal{M} \rightarrow \mathbb{S}_Y$ by $\phi(x_k, l) = (l - 1)j + k$, and its inverse $\phi^{-1} : \mathbb{S}_Y \rightarrow \mathbb{S}_X \times \mathcal{M}$ by $\phi^{-1}(j) = (x_k, l)$ for $j \in \mathbb{S}_Y$, where k is the unique integer satisfying

$j = (l-1)\bar{n} + k$ for some $l \in \{1, 2, \dots, \bar{m}\}$. Then we have

$$\mathbb{E} \left[\Psi(S^{\bar{n}, \bar{m}}, \alpha) \mid \alpha(0) = i, S_0^{\bar{n}, \bar{m}} = x_k \right] = \mathbb{E}[\Psi \circ \phi^{-1}(Y) \mid Y_0 = (i-1)\bar{n} + k], \quad (51)$$

for any path-dependent payoff function Ψ such that the expectation on the left hand side is finite. Here we have defined $S^{\bar{n}, \bar{m}} := (S_t^{\bar{n}, \bar{m}})_{0 \leq t \leq T}$, $\alpha := (\alpha(t))_{0 \leq t \leq T}$, and $Y := (Y_t)_{0 \leq t \leq T}$.

From this representation, [14] are able to derive closed-form pricing formulas for European, barrier, occupation time, and Asian options. Under appropriate conditions, it can be showed that $(S_t^{\bar{n}, \bar{m}}, v_t^{\bar{m}})$ converges weakly to (S_t, v_t) as $\bar{n}, \bar{m} \rightarrow \infty$. The reader is invited to refer to [14] for more details. An extension of theoretical results and applications to time-changed Markov processes is given in [15].

3.5 European options pricing

Vanilla option prices for the underlying S_T can now be approximated with respect to

$$S_T^{\bar{n}, \bar{m}} := g^{-1}(\tilde{X}_T^{\bar{n}, \bar{m}} + \rho f(v_{\alpha_T})), \quad (52)$$

which is the discrete-space asset process corresponding to $\tilde{X}_T^{\bar{n}, \bar{m}}$:

$$\mathbb{E} \left[e^{-rT} (S_T - K)^+ \mid v_0, S_0 \right] \approx \mathbb{E} \left[e^{-rT} \left(S_T^{\bar{n}, \bar{m}} - K \right)^+ \mid \alpha(0) = i, \tilde{X}_0^{\bar{n}, \bar{m}} = x_k \right],$$

where we assume⁹ that $v_{\alpha(0)} = v_i = v_0$ is a member of the grid for some $i \in \mathcal{M}$, and $\tilde{X}_0^{\bar{n}, \bar{m}} = x_k \in \mathcal{S}_X$. From the standard CTMC theory, an explicit representation can be obtained for a European option on $S_T^{\bar{n}, \bar{m}}$, in terms of the characteristics of the one-dimensional process Y_t .

Theorem 7. ([14]) *Given that $\alpha(0) = i, \tilde{X}_0^{\bar{n}, \bar{m}} = x_k$, for maturity T and strike $K > 0$, the approximate European option price at time 0 is given by*

$$\begin{aligned} \mathbb{E} \left[e^{-rT} \left(S_T^{\bar{n}, \bar{m}} - K \right)^+ \mid \alpha(0) = i, \tilde{X}_0^{\bar{n}, \bar{m}} = x_k \right] &= \mathbf{e}_{i, x_k} \cdot \exp((\mathbf{G} - r\mathbf{I})T) \cdot \mathbf{H}^{(1)} \\ &= e^{-rT} \cdot \mathbf{e}_{i, x_k} \cdot \exp(\mathbf{G}T) \cdot \mathbf{H}^{(1)}, \end{aligned} \quad (53)$$

where \mathbf{e}_{i, x_k} is a $1 \times \bar{n}\bar{m}$ vector with all entries equal to 0 except that the $(i-1)\bar{n} + k$ entry is equal to 1, and $\mathbf{H}^{(1)}$ is an $\bar{n}\bar{m} \times 1$ vector with

$$H_{(l-1)\bar{n}+j}^{(1)} = \begin{cases} (g^{-1}(x_j + \rho f(v_l)) - K)^+ & \text{for a call,} \\ (K - g^{-1}(x_j + \rho f(v_l)))^+ & \text{for a put.} \end{cases} \quad (54)$$

⁹ These assumptions are without loss of generality in the sense that interpolation can be readily applied otherwise. To simplify the discussion, we assume that these points are members of the grid in what follows.

During the calibration process, prices are required for many strikes at each maturity. An advantage of the proposed methodology is that once the shared key component, the matrix exponential $\exp(\mathbf{GT})$, is (pre)computed and cached, which dominates the computational cost, a spectrum of contracts with different strikes may be priced for essentially the same cost as a single contract.

SABR (28)	$dS_t = v_t S_t^\beta dW_t^{(1)}$ $dv_t = \alpha v_t dW_t^{(2)}$	$\beta \in [0, 1)$ $\alpha, v_0 > 0$
λ -SABR (30)	$dS_t = v_t S_t^\beta dW_t^{(1)}$ $dv_t = \lambda(\theta - v_t)dt + \alpha v_t dW_t^{(2)}$	$\beta \in [0, 1)$ $\lambda, \theta, \alpha, v_0 > 0$
Shifted SABR (5)	$dS_t = v_t (S_t + s)^\beta dW_t^{(1)}$ $dv_t = \alpha v_t dW_t^{(2)}$	$\beta \in [0, 1)$ $s, \alpha, v_0 > 0$
Heston-SABR (14)	$dS_t = rS_t dt + \sqrt{v_t} S_t^\beta dW_t^{(1)}$ $dv_t = \eta(\theta - v_t)dt + \alpha \sqrt{v_t} dW_t^{(2)}$	$r \in \mathbb{R}, \beta \in [0, 1)$ $\eta, \theta, \alpha, v_0 > 0$
Quadratic SLV (48)	$dS_t = rS_t dt + \sqrt{v_t} (aS_t^2 + bS_t + c) dW_t^{(1)}$ $dv_t = \eta(\theta - v_t)dt + \alpha \sqrt{v_t} dW_t^{(2)}$	$r \in \mathbb{R}, \beta \in [0, 1)$ $a, \eta, \theta, \alpha, v_0 > 0, 4ac > b^2$
Exponential SLV (14)	$dS_t = rS_t dt + m(v_t)(v_L + \theta \exp(-\lambda S_t)) dW_t^{(1)}$ $dv_t = \mu(v_t)dt + \sigma(v_t) dW_t^{(2)}$	$r \in \mathbb{R}, \lambda, v_L \geq 0$ $v_L + \theta \geq 0$
Root-Quadratic SLV (14)	$dS_t = rS_t dt + m(v_t) \sqrt{aS_t^2 + bS_t + c} dW_t^{(1)}$ $dv_t = \mu(v_t)dt + \sigma(v_t) dW_t^{(2)}$	$r \in \mathbb{R}$ $a > 0, c \geq 0$
Tan-Hyp SLV (14)	$dS_t = rS_t dt + m(v_t) \tanh(\beta S_t) dW_t^{(1)}$ $dv_t = \mu(v_t)dt + \sigma(v_t) dW_t^{(2)}$	$r \in \mathbb{R}$ $\beta \geq 0$
Mean-reverting-SABR (14)	$dS_t = \kappa(\zeta - S_t)dt + m(v_t) S_t^\beta dW_t^{(1)}$ $dv_t = \mu(v_t)dt + \sigma(v_t) dW_t^{(2)}$	$r \in \mathbb{R}, \beta \in [0, 1)$ $\kappa, \zeta, v_0 > 0$
4/2-SABR (14)	$dS_t = rS_t dt + S_t^\beta [a\sqrt{v_t} + b/\sqrt{v_t}] dW_t^{(1)}$ $dv_t = \eta(\theta - v_t)dt + \alpha \sqrt{v_t} dW_t^{(2)}$	$r \in \mathbb{R}, \beta \in [0, 1)$ $a, b, \eta, \theta, \alpha, v_0 > 0$

Table 3 Some stochastic local volatility models

3.5.1 Example: SABR model

A now classic SLV example which has seen tremendous application in practice is the SABR model of [30], which is specified as

$$\begin{cases} dS_t = v_t S_t^\beta dW_t^{(1)}, \\ dv_t = \alpha v_t dW_t^{(2)}, \end{cases} \quad (55)$$

In particular, the variance process is governed by a geometric Brownian motion. Given the practical nature of the SABR model, several approximation frameworks have been introduced to efficiently estimate implied volatiles, such as the original approach of Hagan et. al. [28], as well as the improved approximation introduced in Antonov et. al. [5]. Traditional Monte Carlo is also widely used for this model, especially for exotic options for which no known closed-form pricing formulas exist.

Figure 5 compares the CTMC approach introduced in [14] with each of these methods, using the market standard implied volatilities of European options for illustration. We see close agreement between the method of Antonov et. al. and CTMC, while the other two methods under-perform at the wings, as is well docu-

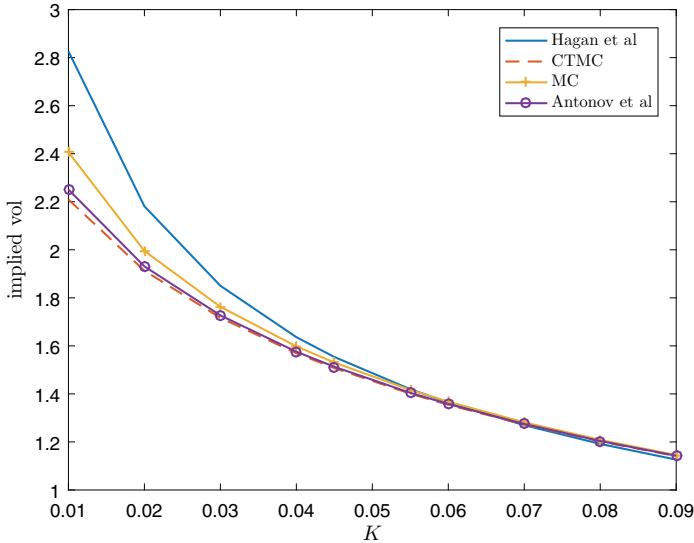


Fig. 5 SABR implied volatilities. $\alpha = 0.2, \beta = 0.1, \rho = 0, v_0 = 0.1, T = 1, S_0 = 0.05, r = 0.0$.

mented. In addition to European options, the CTMC method can be used to price American, Barrier, Asian, and occupation time derivatives in the SABR and other SLV models. Additional SLV model specifications are listed in Table 3. In particular, the λ -SABR model of [30] and the Heston-SABR model studied in [14] offer more realistic models for the variance process, as they permit mean-reversion. A further extension of the method to the shifted SABR model has been considered in [16].

4 Conclusions

This chapter reviews and consolidates recent research activity in the literature on applying continuous-time Markov chains to approximate stochastic processes arising in finance. We discuss the construction, theoretical properties and numerical performance of the CTMC approximations. We also discuss an effective regime-switching approach to approximate the dynamics of stochastic volatility models, which enables us to reduce the valuation problem to one that is concerned with a relatively simple Markov-modulated processes. In particular, explicit valuation formulas are obtained in terms of simple matrix expressions.

Since the CTMC approximation can be thought of as a state-space discretization, as compared to time-discretization schemes (e.g. the Euler scheme), a promising future research direction is to utilize this method in the efficient Monte Carlo sim-

ulation of asset prices. A first step in this direction has obtained promising results which are reported in [17]. We believe that the CTMC approximation method will find applications in various areas including the valuation, estimation and calibration of stochastic models arising in financial engineering and operations research.

References

1. Abate, Joseph, and Ward Whitt. "The Fourier-series method for inverting transforms of probability distributions." *Queueing Systems* 10.1-2 (1992): 5-87.
2. Ackerer, Damien, Damir Filipovic, and Sergio Pulido. "The Jacobi stochastic volatility model." *Finance and Stochastics* (2017): 1-34.
3. Ait-Sahalia, Yacine. "Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach." *Econometrica* 70, no. 1 (2002): 223-262.
4. Ang, Andrew, and Geert Bekaert. "Regime switches in interest rates." *Journal of Business & Economic Statistics* 20, no. 2 (2002): 163-182.
5. Antonov, Alexandre, Michael Konikov, and Michael Spector. "The free boundary SABR: natural extension to negative rates." Preprint, ssn 2557046 (2015).
6. Bangia, Anil, Francis X. Diebold, Andr Kronimus, Christian Schagen, and Til Schuermann. "Ratings migration and the business cycle, with application to credit portfolio stress testing." *Journal of Banking and Finance* 26, no. 2-3 (2002): 445-474.
7. Buffington, John, and Robert J. Elliott. "American options with regime switching." *International Journal of Theoretical and Applied Finance* 5, no. 05 (2002): 497-514.
8. Cai, Ning, Yingda Song, and Steven Kou. "A general framework for pricing Asian options under Markov processes." *Operations Research* 63, no. 3 (2015): 540-554.
9. Chourdakis, Kyriakos, "Continuous Time Regime Switching Models and Applications in Estimating Processes with Stochastic Volatility and Jumps (November 2002)". U of London Queen Mary Economics Working Paper No. 464. Available at SSRN: <https://ssrn.com/abstract=358244> or <http://dx.doi.org/10.2139/ssrn.358244>
10. Chatterjee, Rupak, Zhenyu Cui, Jiacheng Fan, and Mingzhe Liu. "An efficient and stable method for short maturity Asian options." *Journal of Futures Markets* 38 (12) (2018): 1470-1486.
11. Corsaro, Stefania, Ioannis Kyriakou, Daniele Marazzina, and Zeldia Marino. "A general framework for pricing Asian options under stochastic volatility on parallel architectures." *European Journal of Operational Research*, 272(3) (2019): 1082-1095.
12. Cui, Z., J. Lars Kirkby, and Duy Nguyen. "Equity-linked annuity pricing with cliquet-style guarantees in regime-switching and stochastic volatility models with jumps." *Insurance: Mathematics and Economics* 74 (2017): 46-62.
13. Cui, Z., J. Lars Kirkby, and Duy Nguyen. "A general framework for discretely sampled realized variance derivatives in stochastic volatility models with jumps." *European Journal of Operational Research* 262(1) (2017): 381-400.
14. Cui, Z., J. Lars Kirkby, and Nguyen, Duy. "A general valuation framework for SABR and stochastic local volatility models." *SIAM Journal on Financial Mathematics* 9(2) (2018): 520-563.
15. Cui, Z., J. Lars Kirkby and Nguyen, Duy. "A general framework time-changed Markov processes and applications." *European Journal of Operational Research*, 273(2) (2018):785-800.
16. Cui, Z., J. Lars Kirkby and Nguyen, Duy. "Full-fledged SABR through Markov Chains." *Working paper* (2017).
17. Cui, Z., J. Lars Kirkby and Nguyen, Duy. "Efficient simulation of stochastic differential equations based on Markov Chain approximations with applications." *Working paper* (2018).
18. Cui, Z., C. Lee, and Y. Liu. "Single-transform formulas for pricing Asian options in a general approximation framework under Markov processes." *European Journal of Operational Research* 266, no. 3 (2018): 1134-1139.

19. Duan, Jin-Chuan, and Jean-Guy Simonato. "American option pricing under GARCH by a Markov chain approximation." *Journal of Economic Dynamics and Control* 25, no. 11 (2001): 1689-1718.
20. Duan, Jin-Chuan, Evan Dudley, Genevive Gauthier, and J. Simonato. "Pricing discretely monitored barrier options by a Markov chain." *Journal of Derivatives* 10 (2003).
21. Duffie, Darrell, Jun Pan, and Kenneth Singleton. "Transform analysis and asset pricing for affine jump diffusions." *Econometrica* 68, no. 6 (2000): 1343-1376.
22. Durham, Garland B., and A. Ronald Gallant. "Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes." *Journal of Business & Economic Statistics* 20, no. 3 (2002): 297-338.
23. Da Fonseca, Jose, and Claude Martini. "The α -hypergeometric stochastic volatility model." *Stochastic Processes and their Applications* 126.5 (2016): 1472-1502.
24. Ethier, Stewart N., and Thomas G. Kurtz. *Markov processes: characterization and convergence*. Vol. 282. John Wiley & Sons, (2009).
25. Fusai, Gianluca, and Ioannis Kyriakou. "General optimized lower and upper bounds for discrete and continuous arithmetic Asian options." *Mathematics of Operations Research* 41, no. 2 (2016): 531-559.
26. Gihman, Iosif Il'ich, and Anatoli Vladimirovich Skorohod. "Stochastic differential equations." *The Theory of Stochastic Processes III*. Springer, New York, NY, 1979. 113-219.
27. Grasselli, Martino. "The 4/2 stochastic volatility model: A unified approach for the Heston and the 3/2 model." *Mathematical Finance* 27.4 (2017): 1013-1034.
28. Hagan, Patrick S., et al. "Managing smile risk." *The Best of Wilmott* 1 (2002): 249-296.
29. Hamilton, James D. "Analysis of time series subject to changes in regime." *Journal of Econometrics* 45, no. 1-2 (1990): 39-70.
30. Henry-Labordere, Pierre, "A General Asymptotic Implied Volatility for Stochastic Volatility Models (April 2005)". Available at SSRN: <https://ssrn.com/abstract=698601> or <http://dx.doi.org/10.2139/ssrn.698601>.
31. Heston, Steven L. "A closed-form solution for options with stochastic volatility with applications to bond and currency options". *The Review of Financial Studies* 6.2 (1993): 327-343.
32. Hull, John, and Alan White. "The pricing of options on assets with stochastic volatilities." *The Journal of Finance* 42.2 (1987): 281-300.
33. Ikeda, Nobuyuki, and Shinzo Watanabe. "Stochastic differential equations and diffusion processes". Vol. 24. Elsevier, (2014).
34. Jiang, Jiuxin, R. H. Liu, and D. Nguyen. "A recombining tree method for option pricing with state-dependent switching rates." *International Journal of Theoretical and Applied Finance* 19.02 (2016): 1650012.
35. Higham, Desmond J., Xuerong Mao, and Andrew M. Stuart. "Strong convergence of Euler-type methods for nonlinear stochastic differential equations." *SIAM Journal on Numerical Analysis* 40, no. 3 (2002): 1041-1063.
36. Jacod, Jean, and Philip Protter. "Discretization of processes". Vol. 67. *Springer Science & Business Media*, 2011.
37. Kahale, Nabil. "General multilevel Monte Carlo methods for pricing discretely monitored Asian options." arXiv preprint arXiv:1805.09427 (2018).
38. Karatzas, Ioannis, and Steven Shreve. "Brownian motion and stochastic calculus". Vol. 113. *Springer Science & Business Media*, (2012).
39. Kim, Chang-Jin, and Charles R. Nelson. "Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching." *Review of Economics and Statistics* 80, no. 2 (1998): 188-201.
40. Kirkby, J. Lars. "Efficient Option Pricing by Frame Duality with the Fast Fourier Transform". *SIAM J. Financial Mathematics* Vol. 6, no.1 (2015): 713-747.
41. Kirkby, J. Lars. "An Efficient Transform Method for Asian Option Pricing". *SIAM J. Financial Mathematics* Vol. 7, no.1 (2016): 845-892.
42. Kirkby, J. L., and D. Nguyen. "Efficient Asian option pricing under regime switching jump diffusions and stochastic volatility models". *Working paper*, (2016).

43. Kirkby, J. Lars, Duy Nguyen, and Zhenyu Cui. "A unified approach to Bermudan and barrier options under stochastic volatility models with jumps." *Journal of Economic Dynamics and Control* 80 (2017): 75-100.
44. Kushner, Harold, and Paul G. Dupuis. "Numerical methods for stochastic control problems in continuous time". Vol. 24. Springer Science & Business Media, (2013).
45. Lewis, Alan L. "Option Valuation Under Stochastic Volatility II". Finance Press, Newport Beach, CA, 2009.
46. Li, Chenxu, and Xiaocheng Li. "A closed-form expansion approach for pricing discretely monitored variance swaps." *Operations Research Letters* 43, no. 4 (2015): 450-455.
47. Li, Lingfei, and Gongqiu Zhang. "Error analysis of finite difference and Markov chain approximations for option pricing." *Mathematical Finance* 28.3 (2018): 877-919.
48. Lipton, A. (2002). The volatility smile problem. *Risk Magazine*. 15(2), 61-65.
49. Lo, Chia Chun, and Konstantinos Skindilias. "An improved Markov chain approximation methodology: Derivatives pricing and model calibration." *International Journal of Theoretical and Applied Finance* 17.07 (2014): 1450047.
50. Liu, R. H. "Regime-switching recombining tree for option pricing." *International Journal of Theoretical and Applied Finance* 13.03 (2010): 479-499.
51. Liu, R. H. "A new tree method for pricing financial derivatives in a regime-switching mean-reverting model." *Nonlinear Analysis: Real World Applications* 13.6 (2012): 2609-2621.
52. Lord, Roger, Remmert Koekoek, and Dick Van Dijk. "A comparison of biased simulation schemes for stochastic volatility models." *Quantitative Finance* 10, no. 2 (2010): 177-194.
53. Ma, J, W. Yang and Z. Cui. "Convergence rate analysis for the continuous-time Markov chain approximation of occupation time derivatives and Asian option Greeks." *Working paper* (2018).
54. Mijatovic, Aleksandar, and Martijn Pistorius. "Continuously monitored barrier options under Markov processes." *Mathematical Finance* 23 (1),1-38 (2013).
55. Munk, Claus. "The Markov chain approximation approach for numerical solution of stochastic control problems: experiences from Merton's problem." *Applied Mathematics and Computation* 136, no. 1 (2003): 47-77.
56. Nguyen, Duy. "A hybrid Markov chain-tree valuation framework for stochastic volatility jump diffusion models." *International Journal of Financial Engineering* Vol. 05, No. 04, 1850039 (2018).
57. Ramponi, Alessandro. "Fourier transform methods for regime-switching jump-diffusions and the pricing of forward starting options." *International Journal of Theoretical and Applied Finance* 15.05 (2012): 1250037.
58. Schoebel, Rainer, and Jianwei Zhu. "Stochastic volatility with an Ornstein-Uhlenbeck process: an extension." *Review of Finance* 3, no. 1 (1999): 23-46.
59. Scott, Louis O. "Option pricing when the variance changes randomly: Theory, estimation, and an application." *Journal of Financial and Quantitative analysis* 22.4 (1987): 419-438.
60. Song, Yingda, Ning Cai, and Steven Kou. "A Unified Framework for Options Pricing Under Regime Switching Models." *Working paper* (2016).
61. Song, Yingda, Ning Cai, and Steven Kou. "Computable Error Bounds of Laplace Inversion for Pricing Asian Options." *INFORMS Journal on Computing* 30.4 (2018): 634-645..
62. Stein, Elias M., and Jeremy C. Stein. "Stock price distributions with stochastic volatility: an analytic approach." *The review of financial studies* 4.4 (1991): 727-752.
63. Tavella, Domingo, and Curt Randall. *Pricing Financial Instruments: The Finite Difference Method* (Wiley Series in Financial Engineering). New York: Wiley, 2000.
64. Van der Stoep, Anthonie W., Lech A. Grzelak, and Cornelis W. Oosterlee. "The Heston stochastic-local volatility model: Efficient Monte Carlo simulation." *International Journal of Theoretical and Applied Finance* 17.07 (2014): 1450045.
65. Yao, David D., Qing Zhang, and Xun Yu Zhou. "A regime-switching model for European options." *Stochastic processes, optimization, and control theory: applications in financial engineering, queueing networks, and manufacturing systems*. Springer, Boston, MA, 2006. 281-300.

66. Yin, G. George, and Qing Zhang. Continuous-time Markov chains and applications: A two-time-scale approach. Vol. 37. Springer Science & Business Media, 2012.
67. Yin, G. George, and Qing Zhang. Discrete-time Markov chains: two-time-scale methods and applications. Vol. 55. Springer Science & Business Media, 2006.
68. Yin, George, and Chao Zhu. Hybrid switching diffusions: properties and applications. Vol. 63. New York: Springer, 2010.
69. Yuen, Fei Lung, and Hailiang Yang. "Option pricing with regime switching by trinomial tree method." *Journal of Computational and Applied Mathematics* 233.8 (2010): 1821-1833.
70. Zhang Gongqiu, and Lingfei Li. "Analysis of Markov Chain Approximation for Option Pricing and Hedging: Grid Design and Convergence Behavior." *Operations Research*. Forthcoming (2018).
71. Zhang Gongqiu, and Lingfei Li. "A general method for the valuation of drawdown risk under Markovian models." *Working paper* (2018).
72. Zhang Gongqiu, and Lingfei Li. "A unified approach for the analysis of Parisian stopping times and its applications in finance and insurance." *Working paper* (2018).
73. Zhang Gongqiu, and Lingfei Li. "A general approach for the analysis of occupation times and its applications in finance." *Working paper* (2018).
74. Zhang, Qing. "Stock trading: An optimal selling rule." *SIAM Journal on Control and Optimization* 40.1 (2001): 64-87.
75. Zhang, Qing, and Xin Guo. "Closed-form solutions for perpetual American put options with regime switching." *SIAM Journal on Applied Mathematics* 64, no. 6 (2004): 2034-2049.
76. Zhou, Xun Yu, and George Yin. "Markowitz's mean-variance portfolio selection with regime switching: A continuous-time model." *SIAM Journal on Control and Optimization* 42, no. 4 (2003): 1466-1482.



Numerical Approximations for Discounted Continuous Time Markov Decision Processes

François Dufour and Tomás Prieto-Rumeau

Abstract This paper deals with a continuous-time Markov decision process \mathcal{M} , with Borel state and action spaces, under the total expected discounted cost optimality criterion. By suitably approximating an underlying probability measure with a measure with finite support and by discretizing the action sets of the control model, we can construct a finite state and action space Markov decision process that approximates \mathcal{M} and that can be solved explicitly. We can derive bounds on the approximation error of the optimal discounted cost function; such bounds are written in terms of Wasserstein and Hausdorff distances. We show a numerical application to a queueing problem.

1 Introduction

This paper deals with the numerical approximation of a continuous-time Markov decision process under the total expected discounted cost optimality criterion. A typical sample path of the process under consideration consists of a piecewise constant function. We will assume that the Markov decision process has Borel state and action spaces, and its transition rates as well as the cost rate are assumed to be bounded. We are interested in approximating numerically the corresponding optimal discounted cost function.

Continuous-time Markov decision processes have been widely studied, at least from a theoretical point of view. Among the most popular approaches to deal with such problems we can cite dynamic programming and linear programming. The

François Dufour

Institut Polytechnique de Bordeaux; INRIA Bordeaux Sud Ouest, Team CQFD, and Institut de Mathématiques de Bordeaux, Université de Bordeaux, France e-mail: francois.dufour@math.u-bordeaux.fr

Tomás Prieto-Rumeau

Department of Statistics and OR, UNED, Madrid, Spain e-mail: tprieto@ccia.uned.es

first technique establishes an optimality equation, usually referred to as the Bellman or dynamic programming equation, from which the optimal value of the problem and optimal policies can be derived. Under the linear programming approach, the optimization problem is reduced to a linear problem on a space of so-called occupation measures. These techniques enable to address various theoretical problems and, in particular, they provide characterizations of the optimal value function and the existence of optimal control policies. The above mentioned dynamic programming equation and linear problem, however, cannot be explicitly solved in general. There exist some particular problems for which the corresponding solutions can be obtained explicitly (such as, for instance, linear quadratic control problems) but this is not possible in general.

Hence, for practical purposes there is indeed a need of some numerical technique to approximate the solutions of such Markov decision processes. For discrete-time Markov decision processes, there exist several techniques to address the corresponding numerical approximations. A first group of such techniques deals with a model with discrete (say, countable or finite but large) state and action spaces. These approaches rely on stochastic approximation methods, namely, reinforcement learning, neuro-dynamic programming, approximate dynamic programs, and simulation-based methods; the interested reader can consult [3, 5, 15, 19, 20]. The second family of approximation techniques deals with Markov decision processes with general (i.e., Borel) state and action spaces. The approach consists then in approximating the control problem with a discretized Markov decision process with finite state and action spaces. Its optimal solution is used as an approximation of the solution of the original problem. Such methods are studied in, e.g., [6, 7, 8, 9, 18].

The continuous-time counterpart of the above described approximation techniques is however less developed. In [10, 16, 17], for instance, continuous-time control models with countable state space and unbounded transition and cost rates are approximated by means of a sequence of finite models. It is then shown that the optimal value functions and the optimal policies of this sequence of finite models converge to the corresponding optimal solutions of the original control model. Average cost Markov decision processes under an approach similar to the one in the present paper are studied in [2].

Our goal in this paper is to propose a method to approximate the optimal discounted cost of a continuous-time Markov decision process \mathcal{M} with Borel state space \mathbf{X} and Borel action space \mathbf{A} . The original control model \mathcal{M} is approximated by a discretized control model $\mathcal{M}_{k,\eta}$, for $k \geq 1$ and $\eta > 0$, where:

- (i) The state space of $\mathcal{M}_{k,\eta}$ is a finite subset Γ_k of \mathbf{X} . This finite set is obtained as follows. Under the assumption that the positive part of the transition rates of the model \mathcal{M} is absolutely continuous with respect to some probability measure μ on \mathbf{X} , we approximate μ , in the Wasserstein metric, with a measure μ_k with finite support. The set Γ_k is precisely the support of μ_k . This measure is chosen such that, as k grows, the corresponding Wasserstein distance satisfies $W(\mu, \mu_k) \rightarrow 0$.
- (ii) The action sets $\mathbf{A}(x)$, for $x \in \mathbf{X}$, are approximated with finite sets $\mathbf{A}_\eta(x) \subseteq \mathbf{A}(x)$. The accuracy of the approximation is measured in terms of the Hausdorff dis-

tance and we will assume that $\rho_{\mathbf{A}}(\mathbf{A}(x), \mathbf{A}_\eta(x)) \leq \eta$. Hence, the approximating action sets become closer to the original ones as $\eta \rightarrow 0$.

Under suitable hypotheses which will include, among others, Lipschitz continuity of the elements of the control model \mathcal{M} , we will be able to show that the difference between the optimal discounted cost function V^* of \mathcal{M} and the optimal discounted cost $V_{k,\eta}^*$ of the approximating models $\mathcal{M}_{k,\eta}$ is bounded, in the supremum norm, by a combination of the approximation errors in the state and action spaces, respectively. Namely, we will show that

$$\|V^* - V_{k,\eta}^*\| \leq \mathbf{H}_1 \cdot \eta + \mathbf{H}_2 \cdot W(\mu, \mu_k)$$

for some constants $\mathbf{H}_1, \mathbf{H}_2 > 0$ and every $k \geq 1$ and $\eta > 0$. One of the interesting features of the above inequality is that it is a non asymptotic bound. Depending on the nature of the approximation μ_k , we will be able to derive either explicit deterministic error bounds or probabilistic bounds decreasing exponentially in probability as k grows.

The rest of the paper is organized as follows. After introducing some notation, the continuous-time control model \mathcal{M} is constructed in Section 2. Our assumptions and the basic results on discount optimality for \mathcal{M} are studied in Section 3. Our main results in this paper regarding the approximation of \mathcal{M} are given in Section 4. Finally, we show a numerical application to a queueing model in Section 5.

Notation.

The set of nonnegative integers is \mathbb{N} and the real numbers set is \mathbb{R} . The notations $x \vee y$ and $x \wedge y$ stand for the maximum and the minimum of $x, y \in \mathbb{R}$, respectively.

Given a Borel space \mathbf{Y} with metric $d_{\mathbf{Y}}$, its Borel σ -algebra will be denoted by $\mathcal{B}(\mathbf{Y})$. In this paper, measurability is always referred to the Borel σ -algebra. For product spaces, we will consider the taxicab metric and we will as well consider the product σ -algebra. For a bounded function $h : \mathbf{Y} \rightarrow \mathbb{R}$ we will write $\|h\| = \sup_{y \in \mathbf{Y}} |h(y)|$ for its supremum norm.

We say that a function $v : \mathbf{Y} \rightarrow \mathbf{Z}$, where \mathbf{Y} and \mathbf{Z} are Borel spaces, is Lipschitz continuous if there exists $L \geq 0$ with

$$d_{\mathbf{Z}}(v(x), v(y)) \leq L \cdot d_{\mathbf{Y}}(x, y) \quad \text{for all } x, y \in \mathbf{Y}.$$

In this case, we will say that v is L -Lipschitz continuous.

Let $\mathbb{B}(\mathbf{Y})$, $\mathbb{C}(\mathbf{Y})$, and $\mathbb{L}(\mathbf{Y})$ denote the families of real-valued functions on \mathbf{Y} which are bounded and measurable, bounded and continuous, and bounded and Lipschitz continuous, respectively, with, obviously, $\mathbb{L}(\mathbf{Y}) \subseteq \mathbb{C}(\mathbf{Y}) \subseteq \mathbb{B}(\mathbf{Y})$.

We say that $T : \mathcal{B}(\mathbf{Y}) \times \mathbf{Z} \rightarrow \mathbb{R}$ is a transition measure or kernel on the Borel space \mathbf{Y} given the Borel space \mathbf{Z} if $B \mapsto T(B|z)$ is a (signed) measure on $(\mathbf{Y}, \mathcal{B}(\mathbf{Y}))$ for all $z \in \mathbf{Z}$ and $z \mapsto T(B|z)$ is measurable for every $B \in \mathcal{B}(\mathbf{Y})$. For measurable $v : \mathbf{Y} \rightarrow \mathbb{R}$, we will denote by Tv the function on \mathbf{Z} defined as

$$Tv(z) = \int_{\mathbf{Y}} v(y)T(dy|z) \quad \text{for } z \in \mathbf{Z},$$

whenever the integral is well defined. We say that T is a stochastic kernel when $T(\cdot|z)$ is a probability measure on \mathbf{Y} for all $z \in \mathbf{Z}$. We say that T is L_T -Lipschitz continuous for some $L_T \geq 0$ when, for any L_v -Lipschitz continuous function $v : \mathbf{Y} \rightarrow \mathbb{R}$, the mapping $z \mapsto Tv(z)$ is $(L_T \cdot L_v)$ -Lipschitz continuous on \mathbf{Z} .

The Hausdorff metric, on the class of nonempty closed sets of a Borel space \mathbf{Z} , is defined as

$$\rho_{\mathbf{Z}}(C_1, C_2) = \sup_{z_1 \in C_1} \inf_{z_2 \in C_2} \{d_{\mathbf{Z}}(z_1, z_2)\} \vee \sup_{z_2 \in C_2} \inf_{z_1 \in C_1} \{d_{\mathbf{Z}}(z_1, z_2)\}.$$

A multifunction Ψ from \mathbf{Y} to \mathbf{Z} is a function that associates to each $y \in \mathbf{Y}$ a nonempty subset $\Psi(y)$ of \mathbf{Z} . It is said to be closed-valued when $\Psi(y)$ is a closed subset of \mathbf{Z} for any $y \in \mathbf{Y}$. A closed-valued multifunction Ψ is Lipschitz continuous when $\rho_{\mathbf{Z}}(\Psi(x), \Psi(y)) \leq L_{\Psi} \cdot d_{\mathbf{Y}}(x, y)$ for some constant $L_{\Psi} \geq 0$ and all $x, y \in \mathbf{Y}$.

The family of probability measures on $(\mathbf{Y}, \mathcal{B}(\mathbf{Y}))$ is denoted by $\mathcal{P}(\mathbf{Y})$. Given $y \in \mathbf{Y}$, the Dirac probability measure concentrated at y will be denoted by δ_y , that is, $\delta_y(B) = \mathbf{1}_B(y)$, where $\mathbf{1}$ denotes the indicator function of the set $B \in \mathcal{B}(\mathbf{Y})$. The class of probability measures $\mu \in \mathcal{P}(\mathbf{Y})$ with finite first moment (meaning that $\int_{\mathbf{Y}} d_{\mathbf{Y}}(y, y_0)\mu(dy)$ is finite for some $y_0 \in \mathbf{Y}$) is denoted by $\mathcal{P}_1(\mathbf{Y})$. Finally, we say that $\mu \in \mathcal{P}(\mathbf{Y})$ has a finite exponential moment if there is some $\gamma > 0$ with

$$\int_{\mathbf{Y}} \exp\{\gamma d_{\mathbf{Y}}(y, y_0)\}\mu(dy) < \infty$$

for some $y_0 \in \mathbf{Y}$. The class of all such measures is denoted by $\mathcal{P}_{\text{exp}}(\mathbf{Y})$. The following inclusions hold: $\mathcal{P}_{\text{exp}}(\mathbf{Y}) \subseteq \mathcal{P}_1(\mathbf{Y}) \subseteq \mathcal{P}(\mathbf{Y})$.

The Wasserstein distance between μ and ν in $\mathcal{P}_1(\mathbf{Y})$ is defined as

$$W(\mu, \nu) = \sup_f \left\{ \int_{\mathbf{Y}} f d\mu - \int_{\mathbf{Y}} f d\nu \right\}$$

where the sup is taken over the set of 1-Lipschitz continuous functions $f : \mathbf{Y} \rightarrow \mathbb{R}$. In particular, if $f : \mathbf{Y} \rightarrow \mathbb{R}$ is L_f -Lipschitz continuous, then

$$\left| \int_{\mathbf{Y}} f d\mu - \int_{\mathbf{Y}} f d\nu \right| \leq L_f W(\mu, \nu).$$

2 Definition of the Control Model

In this section we will define the control model \mathcal{M} we will be dealing with, and we will provide a formal construction of the corresponding controlled process.

Elements of the Control Model

The control model \mathcal{M} consists of the following elements.

- The state space \mathbf{X} and the action space \mathbf{A} are Borel spaces with metrics $d_{\mathbf{X}}$ and $d_{\mathbf{A}}$.
- The family $\{\mathbf{A}(x)\}_{x \in \mathbf{X}}$ of nonempty measurable subsets of \mathbf{A} stands for the actions available at each state $x \in \mathbf{X}$. We assume that

$$\mathbf{K} = \{(x, a) \in \mathbf{X} \times \mathbf{A} : a \in \mathbf{A}(x)\}$$

is in $\mathcal{B}(\mathbf{X} \times \mathbf{A})$ and that it contains the graph of some measurable $f : \mathbf{X} \rightarrow \mathbf{A}$. Let Ψ be the multifunction from \mathbf{X} to \mathbf{A} defined by $x \mapsto \mathbf{A}(x)$.

- The transition rates kernel is $q(B|x, a)$, for $B \in \mathcal{B}(\mathbf{X})$ and $(x, a) \in \mathbf{K}$. For each fixed $(x, a) \in \mathbf{K}$, we have that $B \mapsto q(B|x, a)$ is a signed measure on \mathbf{X} satisfying

$$q(B|x, a) \geq 0 \text{ when } x \notin B, \text{ and } q(\mathbf{X}|x, a) = 0.$$

In this paper, we will further assume that the transition rates are bounded, meaning that the $-q(\{x\}|x, a)$ are bounded when $(x, a) \in \mathbf{K}$. In particular, we can choose a constant $\hat{q} > 0$ such that

$$\sup_{(x, a) \in \mathbf{K}} \{-q(\{x\}|x, a)\} < \hat{q}. \quad (1)$$

- A bounded and measurable cost rate function $c : \mathbf{K} \rightarrow \mathbb{R}$.

The control model \mathcal{M} is therefore defined by the tuple

$$\mathcal{M} = (\mathbf{X}, \mathbf{A}, \{\mathbf{A}(x) : x \in \mathbf{X}\}, q, c).$$

The positive part of the transition rates kernel q^+ is defined as

$$q^+(B|x, a) = q(B|x, a) - q(\{x\}|x, a)\mathbf{1}_B(x) \quad \text{for } B \in \mathcal{B}(\mathbf{X}) \text{ and } (x, a) \in \mathbf{K}. \quad (2)$$

It indeed satisfies $q^+(B|x, a) \geq 0$ for all $B \in \mathcal{B}(\mathbf{X})$ and $(x, a) \in \mathbf{K}$. The condition (1) that the transition rates are bounded can be equivalently formulated as $\sup_{(x, a) \in \mathbf{K}} q^+(\mathbf{X}|x, a) < \infty$.

Construction of the Process

In what follows, we address the construction of the controlled process. We augment the state space with an isolated point: $\mathbf{X}_\infty = \mathbf{X} \cup \{x_\infty\}$. For each $n \in \mathbb{N}$, let

$$\Omega_n = \mathbf{X} \times ((0, \infty) \times \mathbf{X})^n \times (\{\infty\} \times \{x_\infty\})^\infty.$$

We define the canonical space Ω by means of

$$\Omega = (\mathbf{X} \times ((0, \infty) \times \mathbf{X})^\infty) \cup \bigcup_{n \in \mathbb{N}} \Omega_n,$$

and we will consider the corresponding product σ -algebra \mathcal{F} . An element of the canonical space Ω of the form

$$\omega = (x_0, \dots, x_n, \theta_{n+1}, x_{n+1}, \dots) \tag{3}$$

is interpreted in the following way. For any $n \in \mathbb{N}$, if $x_n \in \mathbf{X}$ then θ_{n+1} is the sojourn time of the process at x_n ; then, either

- $\theta_{n+1} < \infty$ and $x_{n+1} \in \mathbf{X}$ is the post-jump location of the process, or
- $\theta_{n+1} = \infty$ and no further jumps occur. In this case we let $x_m = x_\infty$ and $\theta_m = \infty$ for all $m > n$. Note that such an ω is in Ω_n .

For each $n \in \mathbb{N}$ we define the function $X_n : \Omega \rightarrow \mathbf{X}_\infty$ which associates to each $\omega \in \Omega$ as in (3) the state $x_n \in \mathbf{X}_\infty$. Similarly, the function $\Theta_n : \Omega \rightarrow [0, \infty]$ is given by $\Theta_n(\omega) = \theta_n$, where we make the convention that Θ_0 is a constant function $\Theta_0 \equiv 0$.

We also define the functions T_n on Ω taking values in $[0, \infty]$ as $T_n = \Theta_0 + \dots + \Theta_n$. The function $T_\infty = \lim_n T_n = \sum \Theta_n$ is called the explosion time of the process, and $\omega \in \Omega$ is called explosive when $T_\infty(\omega)$ is finite.

Finally, for each $n \in \mathbb{N}$ we let $H_n = (X_0, \Theta_1, \dots, \Theta_n, X_n)$, which assigns to each $\omega \in \Omega$ its path up to step n , namely,

$$H_n(\omega) = (x_0, \theta_1, x_1, \theta_2, x_2, \dots, \theta_n, x_n).$$

The set of all such paths is denoted by \mathbf{H}_n .

We define a random point measure ν on $(0, \infty) \times \mathbf{X}$. For any $\omega \in \Omega$, the measure $\nu(\omega, \cdot)$ places a mass equal to one on the pairs $(T_n(\omega), X_n(\omega))$, for each $n \geq 1$, provided that $T_n(\omega) < \infty$. In this way, knowledge of the point measure $\nu(\omega, \cdot)$, together with the initial state $x_0 \in \mathbf{X}$, gives full knowledge of the sample path $\omega \in \Omega$. Formally, we write

$$\nu(\omega, dt, dx) = \sum_{n \geq 1} I_{\{T_n(\omega) < \infty\}} \delta_{(T_n(\omega), X_n(\omega))}(dt, dx).$$

For any $t \geq 0$, we define $\mathcal{F}_t \subseteq \mathcal{F}$ as the minimal σ -algebra on Ω for which the mappings

$$\omega \mapsto H_0(\omega) \quad \text{and} \quad \omega \mapsto \nu(\omega, (0, s] \times B), \quad \text{for } 0 < s \leq t \text{ and } B \in \mathcal{B}(\mathbf{X}),$$

are measurable. To conclude, we can define the continuous-time process $\{\xi_t\}_{t \geq 0}$ taking values in \mathbf{X}_∞ as

$$\xi_t(\omega) = \begin{cases} X_n(\omega), & \text{if } T_n(\omega) \leq t < T_{n+1}(\omega) \text{ for some } n \in \mathbb{N}, \\ x_\infty, & \text{if } t \geq T_\infty(\omega). \end{cases}$$

Admissible Policies

The action space is also augmented with an isolated point, i.e., $\mathbf{A}_\infty = \mathbf{A} \cup \{a_\infty\}$. The isolated action is the only action available at state x_∞ ; hence, we define $\mathbf{A}(x_\infty) = \{a_\infty\}$. The transition rates q are extended to a signed kernel on \mathbf{X}_∞ given $\mathbf{K} \cup \{(x_\infty, a_\infty)\}$ by defining $q(\{x_\infty\}|x, a) = 0$ for every $(x, a) \in \mathbf{K}$ and $q(\cdot|x_\infty, a_\infty) \equiv 0$. The definition of q^+ in (2) is extended to the so-defined signed kernel.

An admissible control policy u is given by a sequence $u = (\pi_n)_{n \in \mathbb{N}}$, where each π_n is a stochastic kernel on \mathbf{A}_∞ given $\mathbf{H}_n \times (0, \infty)$ that satisfies, in addition,

$$\pi_n(A(x_n)|h_n, t) = 1 \quad \text{for all } h_n = (x_0, \theta_1, \dots, \theta_n, x_n) \in \mathbf{H}_n \text{ and } t > 0.$$

We will denote by \mathcal{U} the family of all admissible control policies.

Given $u \in \mathcal{U}$, we define a random process $\pi_t(da, \omega)$, for $t > 0$, taking values in $\mathcal{P}(\mathbf{A}_\infty)$ as follows:

$$\pi_t(da, \omega) = \sum_{n \in \mathbb{N}} I_{\{T_n(\omega) < t \leq T_{n+1}(\omega)\}} \pi_n(da|H_n(\omega), t - T_n(\omega)) + \mathbf{1}_{\{t \geq T_\infty(\omega)\}} \delta_{a_\infty}(da).$$

Observe that $\{\pi_t\}$ is an $\{\mathcal{F}_t\}$ -predictable random process.

By hypothesis, the set \mathbb{F} of measurable functions $f: \mathbf{X} \rightarrow \mathbf{A}$ such that $f(x) \in A(x)$ for all $x \in \mathbf{X}$ is nonempty. Any $f \in \mathbb{F}$ can be extended to a function from \mathbf{X}_∞ to \mathbf{A}_∞ by letting $f(x_\infty) = a_\infty$. We can associate to any such $f \in \mathbb{F}$ the control policy $u = (\pi_n)_{n \in \mathbb{N}} \in \mathcal{U}$ such that $\pi_n(B|h_n, t) = \delta_{f(x_n)}(B)$ for all $n \in \mathbb{N}$, $h_n = (x_0, \theta_1, \dots, \theta_n, x_n) \in \mathbf{H}_n$, $t > 0$, and $B \in \mathcal{B}(\mathbf{A}_\infty)$. We will identify $f \in \mathbb{F}$ with the above defined policy and hence we have $\mathbb{F} \subseteq \mathcal{U}$. We will refer to $f \in \mathbb{F}$ as to a deterministic stationary policy.

The Controlled Stochastic Process

Let $u = (\pi_n)_{n \in \mathbb{N}} \in \mathcal{U}$ be an admissible control policy. Given $\Gamma \in \mathcal{B}(\mathbf{X}_\infty)$, $n \in \mathbb{N}$, $h_n = (x_0, \theta_1, x_1, \dots, \theta_n, x_n) \in \mathbf{H}_n$, and $t > 0$, we define the intensity of the jumps

$$\lambda_n(\Gamma, h_n, t) = \int_{\mathbf{A}_\infty} q^+(\Gamma|x_n, a) \pi_n(da|h_n, t),$$

and the jump rate as

$$\Lambda_n(\Gamma, h_n, t) = \int_0^t \lambda_n(\Gamma, h_n, s) ds.$$

Now, given $\Gamma \in \mathcal{B}((0, \infty] \times \mathbf{X}_\infty)$, $n \in \mathbb{N}$, and $h_n = (x_0, \theta_1, x_1, \dots, \theta_n, x_n) \in \mathbf{H}_n$, we consider the stochastic kernel G_n on $(0, \infty] \times \mathbf{X}_\infty$ given \mathbf{H}_n is defined by

$$\begin{aligned} G_n(\Gamma|h_n) &= \mathbf{1}_{\Gamma}(\infty, x_\infty) \left[\mathbf{1}_{\{x_\infty\}}(x_n) + \mathbf{1}_{\mathbf{X}}(x_n) e^{-\Lambda_n(\mathbf{X}, h_n, \infty)} \right] \\ &\quad + \mathbf{1}_{\mathbf{X}}(x_n) \int_{\Gamma \cap ((0, \infty) \times \mathbf{X})} \lambda_n(dx, h_n, t) e^{-\Lambda_n(\mathbf{X}, h_n, t)} dt. \end{aligned}$$

The interpretation is the following: given the path of the process $h_n \in \mathbf{H}_n$ up to step n , the kernel G_n gives the (conditional) distribution of the sojourn time θ_{n+1} and the post-jump location x_{n+1} .

We can now use Remark 3.43 in [12] to establish that, given an initial state $x \in \mathbf{X}$ and an admissible control policy $u \in \mathcal{U}$, there exists a probability measure $\mathbb{P}^{x,u}$ on (Ω, \mathcal{F}) with

$$\mathbb{P}^{x,u}\{X_0 = x\} = 1$$

and such that, in addition, for any $\Gamma \in \mathcal{B}((0, \infty] \times \mathbf{X}_\infty)$ and $n \in \mathbb{N}$

$$\mathbb{P}^{x,u}\{(\Theta_{n+1}, X_{n+1}) \in \Gamma \mid H_n\} = G_n(\Gamma \mid H_n)$$

almost surely. This probability measure indeed models the controlled process $\{\xi_t\}$ under the policy u . The corresponding expectation operator is denoted by $\mathbb{E}^{x,u}$.

Remark 1. It is important to mention that under the condition that the transition rates are bounded (see (1)), the sample paths ω of the process are non-explosive with probability one, meaning that for any initial state and any control policy we have $\mathbb{P}^{x,u}\{T_\infty < \infty\} = 0$; see, e.g., [14, Theorem 1].

3 Assumptions and Basic Results

In this section we introduce the total expected discounted cost optimality criterion. We also state our assumptions on \mathcal{M} and prove some important preliminary facts on the dynamic programming equation.

The Discount Optimality Criterion

Let $\alpha > 0$ be a given discount rate. The total expected discounted cost of the admissible control policy $u \in \mathcal{U}$ when the initial state of the system is $x \in \mathbf{X}$ is defined as

$$V(x, u) = \mathbb{E}^{x,u} \left[\int_0^\infty e^{-\alpha s} \int_{\mathbf{A}(\xi_s)} c(\xi_s, a) \pi_s(da) ds \right].$$

It is well defined and finite because the cost rate function c is bounded. In particular, $|V(x, u)| \leq \|c\|/\alpha$ for any $x \in \mathbf{X}$ and $u \in \mathcal{U}$.

The optimal total expected discounted cost function is

$$V^*(x) = \inf_{u \in \mathcal{U}} V(x, u) \quad \text{for } x \in \mathbf{X}.$$

We say that an admissible control policy $u^* \in \mathcal{U}$ is optimal when $V^*(x) = V(x, u^*)$ for every $x \in \mathbf{X}$. We also have $\|V^*\| \leq \|c\|/\alpha$.

Assumptions

In Assumption 1 below we will use the following notation. We define the stochastic kernel Q on \mathbf{X} given \mathbf{K} as

$$Q(dy|x, a) = \frac{1}{\hat{q}} \cdot q(dy|x, a) + \delta_x(dy) \quad \text{for any } (x, a) \in \mathbf{K}, \quad (4)$$

where $\hat{q} > 0$ is taken from in (1). It is easily seen that for every $B \in \mathcal{B}(\mathbf{X})$ and $(x, a) \in \mathbf{K}$ we have $Q(B|x, a) \geq 0$, and also that $Q(\mathbf{X}|x, a) = 1$.

- Assumption 1.** (i) The cost function c is in $\mathbb{L}(\mathbf{K})$, with Lipschitz constant L_c .
(ii) The multifunction Ψ is compact-valued and L_Ψ -Lipschitz continuous.
(iii) If $v \in \mathbb{C}(\mathbf{X})$ then $qv \in \mathbb{C}(\mathbf{K})$ or, equivalently, $Qv \in \mathbb{C}(\mathbf{K})$.
(iv) The stochastic kernel Q is L_Q -Lipschitz continuous.
(v) The Lipschitz constants above and the discount rate satisfy

$$\alpha > \hat{q}(L_Q(1 + L_\Psi) - 1).$$

We say that a function $v \in \mathbb{B}(\mathbf{X})$ is a solution of the dynamic programming equation if it satisfies

$$\alpha v(x) = \inf_{a \in A(x)} \left\{ c(x, a) + \int_{\mathbf{X}} v(y) q(dy|x, a) \right\} \quad \text{for each } x \in \mathbf{X}.$$

Moreover, we say that $f \in \mathbb{F}$ attains the minimum in the dynamic programming equation when

$$\alpha v(x) = c(x, f(x)) + \int_{\mathbf{X}} v(y) q(dy|x, f(x)) \quad \text{for each } x \in \mathbf{X}.$$

Our next result characterizes the optimal discounted cost function V^* as the solution of the dynamic programming equation and it explores further properties. The proof of this result follows standard arguments and it will be omitted; see, for instance, Theorem 4 in [14] or Lemma 2.1 in [2].

Theorem 1. *Suppose that the control model \mathcal{M} satisfies Assumption 1.*

- (i) *The optimal discounted cost function V^* is the unique solution in $\mathbb{B}(\mathbf{X})$ of the dynamic programming equation, i.e.,*

$$\alpha V^*(x) = \min_{a \in A(x)} \left\{ c(x, a) + \int_{\mathbf{X}} V^*(y) q(dy|x, a) \right\} \quad \text{for each } x \in \mathbf{X}.$$

- (ii) *Any $f \in \mathbb{F}$ attaining the minimum in the dynamic programming equation, and such f indeed exist, is an optimal deterministic stationary policy (that is why we write min instead of sup in (i) above).*
(iii) *The optimal discounted cost function V^* is in $\mathbb{L}(\mathbf{X})$ with $\|V^*\| \leq \|c\|/\alpha$ and*

$$L_{V^*} = \frac{L_c(1 + L_{\Psi})}{\alpha + \hat{q}(1 - L_Q(1 + L_{\Psi}))}.$$

Note that the dynamic programming equation can be equivalently written as

$$V^*(x) = \inf_{a \in A(x)} \left\{ \frac{c(x, a)}{\alpha + \hat{q}} + \frac{\hat{q}}{\alpha + \hat{q}} \int_{\mathbf{X}} V^*(y) Q(dy|x, a) \right\} \quad \text{for each } x \in \mathbf{X},$$

and so it can be characterized as a fixed point $TV^* = V^*$ of the operator T defined, for $v \in \mathbb{B}(\mathbf{X})$, as

$$Tv(x) = \inf_{a \in A(x)} \left\{ \frac{c(x, a)}{\alpha + \hat{q}} + \frac{\hat{q}}{\alpha + \hat{q}} \int_{\mathbf{X}} v(y) Q(dy|x, a) \right\} \quad \text{for each } x \in \mathbf{X}. \quad (5)$$

It is worth noting that T is a contraction operator with $\|Tv - Tu\| \leq \frac{\hat{q}}{\alpha + \hat{q}} \|v - u\|$.

4 Approximation Results

In this section we are going to approximate the control model \mathcal{M} , which we will subsequently call the original control model, by means of discretized control models $\mathcal{M}_{k, \eta}$, indexed by an integer $k \geq 1$ and $\eta > 0$. We will call $\mathcal{M}_{k, \eta}$ the approximating control model.

At this point, it is useful to recall the definition of the positive part q^+ of the transition rates kernel q , given in (2):

$$q^+(B|x, a) = q(B|x, a) - q(\{x\}|x, a) \mathbf{1}_B(x) \quad \text{for } B \in \mathcal{B}(\mathbf{X}) \text{ and } (x, a) \in \mathbf{K}.$$

The above definition implies that for any $(x, a) \in \mathbf{K}$ we have $q^+(\{x\}|x, a) = 0$ and so $q^+(\mathbf{X}|x, a) = -q(\{x\}|x, a)$. In particular, given $v \in \mathbb{B}(\mathbf{X})$ we have

$$\begin{aligned} \int_{\mathbf{X}} v(y) q(dy|x, a) &= \int_{\mathbf{X}} v(y) q^+(dy|x, a) - v(x) q^+(\mathbf{X}|x, a) \\ &= \int_{\mathbf{X}} (v(y) - v(x)) q^+(dy|x, a) \end{aligned} \quad (6)$$

for each $(x, a) \in \mathbf{K}$.

Assumptions 2(i)–(ii) below impose that the kernel q^+ is absolutely continuous with respect to some probability measure μ for a sufficiently regular density function; this will allow us to discretize the state space. The condition in Assumption 2(iii) will be needed to discretize the action sets.

Assumption 2. There exist a probability measure $\mu \in \mathcal{P}_1(\mathbf{X})$ and a nonnegative function $p \in \mathbb{B}(\mathbf{X} \times \mathbf{K})$ such that:

(i) For every $B \in \mathcal{B}(\mathbf{X})$ and $(x, a) \in \mathbf{K}$

$$q^+(B|x, a) = \int_B p(y|x, a)\mu(dy).$$

- (ii) There exists some $L_p > 0$ such that the function $p(\cdot|x, \cdot)$ is L_p -Lipschitz continuous on $\mathbf{X} \times \mathbf{A}(x)$ for each $x \in \mathbf{X}$.
- (iii) For every $\eta > 0$ and $x \in \mathbf{X}$, there exists a finite set $\mathbf{A}_\eta(x) \subseteq \mathbf{A}(x)$ such that the multifunction defined on \mathbf{X} by $x \mapsto \mathbf{A}_\eta(x)$ is Borel-measurable and

$$\rho_{\mathbf{A}}(\mathbf{A}(x), \mathbf{A}_\eta(x)) \leq \eta.$$

Given $\eta > 0$, we will denote by \mathbf{K}_η the graph of the multifunction $x \mapsto \mathbf{A}_\eta(x)$ which, as a consequence of Proposition D.4 in [11], is a measurable subset of $\mathbf{X} \times \mathbf{A}$. Moreover, the sets $\mathbf{A}_\eta(x)$ being compact (they are finite) the set \mathbb{F}_η of functions $f : \mathbf{X} \rightarrow \mathbf{A}$ such that $f(x) \in \mathbf{A}_\eta(x)$ for all $x \in \mathbf{X}$ is not empty; see [13].

Under this assumption, the transition rates of the control model \mathcal{M} can be written

$$q(B|x, a) = \int_B p(y|x, a)\mu(dy) - \mathbf{1}_B(x) \int_{\mathbf{X}} p(y|x, a)\mu(dy) \tag{7}$$

for $B \in \mathcal{B}(\mathbf{X})$ and $(x, a) \in \mathbf{K}$.

Construction of the Approximating Model $\mathcal{M}_{k,\eta}$

For each $k \geq 1$, let $\mu_k \in \mathcal{P}(\mathbf{X})$ be a probability measure with finite support $\Gamma_k \subseteq \mathbf{X}$. For interpretation purposes, one may think of μ_k as a probability measure supported on k points in \mathbf{X} , and such that $W(\mu, \mu_k) \rightarrow 0$ as $k \rightarrow \infty$. Our next definition uses the sets $\mathbf{A}_\eta(x)$ given in Assumption 2.

Definition 1. Given $k \geq 1$ and $\eta > 0$, define the continuous-time control model

$$\mathcal{M}_{k,\eta} = (\mathbf{X}, \mathbf{A}, \{\mathbf{A}_\eta(x) : x \in \mathbf{X}\}, q_k, c)$$

where

$$q_k(B|x, a) = \int_B p(y|x, a)\mu_k(dy) - \mathbf{1}_B(x) \cdot \int_{\mathbf{X}} p(y|x, a)\mu_k(dy), \tag{8}$$

for $B \in \mathcal{B}(\mathbf{X})$ and $(x, a) \in \mathbf{K}_\eta$.

Observe that q_k in (8) is the analogous of (7) where we have just replaced μ with μ_k . Note that q_k is indeed a transition rate kernel on \mathbf{X} given \mathbf{K}_η because, for $B \in \mathcal{B}(\mathbf{X})$ and $(x, a) \in \mathbf{K}_\eta$, we have $q_k(\mathbf{X}|x, a) = 0$ and $q_k(B|x, a) \geq 0$ whenever $x \notin B$. Note that for any $(x, a) \in \mathbf{K}_\eta$

$$\begin{aligned} -q_k(\{x\}|x, a) &\leq \int_{\mathbf{X}} p(y|x, a)\mu_k(dy) \\ &\leq \int_{\mathbf{X}} p(y|x, a)\mu(dy) + L_p W(\mu, \mu_k) \end{aligned} \tag{9}$$

$$= -q(\{x\}|x, a) + L_p W(\mu, \mu_k), \tag{10}$$

where in (9) we make use of Lipschitz continuity of $p(\cdot|x,a)$, and so the control models $\mathcal{M}_{k,\eta}$ have bounded transition rates as well.

There is a slight abuse of notation in the definition of q_k in (8) because, although the expression given in (8) is valid for any $(x,a) \in \mathbf{K}$, in fact we will only consider it for $(x,a) \in \mathbf{K}_\eta$. We prefer, however, to keep the notation q_k instead of the more cumbersome $q_{k,\eta}$. Finally, we maintain the notation c for the cost rate function of $\mathcal{M}_{k,\eta}$ which is defined on \mathbf{K}_η .

The construction of the controlled stochastic process carried out in Section 2 can be done for the control models $\mathcal{M}_{k,\eta}$. In particular, we can construct the family of admissible control policies $\mathcal{U}_\eta \supseteq \mathbb{F}_\eta$, and the existence of the corresponding probability measure $\mathbb{P}_{k,\eta}^{x,u}$ on the canonical space, for a given initial state $x \in \mathbf{X}$ and a control policy $u \in \mathcal{U}_\eta$, follows as well.

Remark 2. In view of (6) and the definition of q_k we can write, for $v \in \mathbb{B}(\mathbf{X})$,

$$\int_{\mathbf{X}} v(y)q(dy|x,a) = \int_{\mathbf{X}} (v(y) - v(x))p(y|x,a)\mu(dy)$$

for any $(x,a) \in \mathbf{K}$, while for every $(x,a) \in \mathbf{K}_\eta$ we have

$$\int_{\mathbf{X}} v(y)q_k(dy|x,a) = \int_{\mathbf{X}} (v(y) - v(x))p(y|x,a)\mu_k(dy).$$

Regarding the control model $\mathcal{M}_{k,\eta}$, it is worth mentioning that the first jump of the process (if it ever occurs) necessarily takes the process to the support $\bar{\Gamma}_k$ of μ_k , and the process will remain thereafter in $\bar{\Gamma}_k$.

Similarly, given the discount rate $\alpha > 0$, we can define the corresponding optimal discounted cost function on \mathbf{X} , which we will denote by $V_{k,\eta}^*$. Our next result is proved using standard arguments.

Proposition 1. *Suppose that Assumptions 1 and 2 hold. Then for every $k \geq 1$ and $\eta > 0$, the control model $\mathcal{M}_{k,\eta}$ satisfies the following properties.*

(i) *The optimal discounted cost function $V_{k,\eta}^*$ is the unique solution in $\mathbb{B}(\mathbf{X})$ of the dynamic programming equation*

$$\alpha V_{k,\eta}^*(x) = \min_{a \in \mathbf{A}_\eta(x)} \left\{ c(x,a) + \int_{\mathbf{X}} V_{k,\eta}^*(y)q_k(dy|x,a) \right\} \quad \text{for each } x \in \mathbf{X}.$$

(ii) *Any $f \in \mathbb{F}_\eta$ attaining the minimum in the dynamic programming equation (and such f indeed exist) is an optimal deterministic stationary policy for $\mathcal{M}_{k,\eta}$.*

We define the constant (recall (1))

$$\vartheta = \frac{1}{L_p} \cdot \left(\hat{q} - \sup_{(x,a) \in \mathbf{K}} \{ -q(\{x\}|x,a) \} \right) > 0 \quad (11)$$

and suppose now that the probability measure μ_k is such that $W(\mu, \mu_k) \leq \vartheta$. In particular, by (10), this implies that

$$\sup_{(x,a) \in \mathbf{K}_\eta} \{ -q_k(\{x\}|x,a) \} < \hat{q}$$

which is analogous to (1). Then we can define a stochastic kernel on \mathbf{X} given \mathbf{K}_η by means of (cf. (4))

$$Q_k(dy|x,a) = \frac{1}{\hat{q}} \cdot q_k(dy|x,a) + \delta_x(dy).$$

We indeed have, for any $(x,a) \in \mathbf{K}_\eta$, that $Q_k(B|x,a) \geq 0$ for all $B \in \mathcal{B}(\mathbf{X})$ and also that $Q_k(\mathbf{X}|x,a) = 1$. The dynamic programming equation for the control model $\mathcal{M}_{k,\eta}$ can thus be written as a fixed point $V_{k,\eta}^* = T_{k,\eta} V_{k,\eta}^*$ of the operator

$$T_{k,\eta} v(x) = \min_{a \in A_\eta(x)} \left\{ \frac{c(x,a)}{\alpha + \hat{q}} + \frac{\hat{q}}{\alpha + \hat{q}} \int_{\mathbf{X}} v(y) Q_k(dy|x,a) \right\} \quad \text{for each } x \in \mathbf{X}. \quad (12)$$

As for the operator T defined in Section 3, we have that $T_{k,\eta}$ is a contraction operator on $\mathbb{B}(\mathbf{X})$ with modulus $\frac{\hat{q}}{\alpha + \hat{q}}$.

Our next result compares the operators T and $T_{k,\eta}$ when they are applied to a L_v -Lipschitz continuous function $v \in \mathbb{L}(\mathbf{X})$.

Lemma 1. *Suppose that Assumptions 1 and 2 are satisfied. Consider the control model $\mathcal{M}_{k,\eta}$, where we assume that $W(\mu, \mu_k) \leq \mathfrak{d}$. Under these conditions, for any $v \in \mathbb{L}(\mathbf{X})$ we have*

$$\|Tv - T_{k,\eta}v\| \leq \frac{L_c + 2\|v\|L_p}{\alpha + \hat{q}} \cdot \eta + \frac{2\|v\|L_p + \|p\|L_v}{\alpha + \hat{q}} \cdot W(\mu, \mu_k).$$

Proof. Fix $x \in \mathbf{X}$. There exists some $a \in \mathbf{A}_\eta(x)$ attaining the minimum in the definition of $T_{k,\eta}v(x)$ (see (12)). Hence, since $a \in \mathbf{A}(x)$ as well, we have

$$\begin{aligned} Tv(x) - T_{k,\eta}v(x) &\leq \frac{\hat{q}}{\alpha + \hat{q}} \left[\int_{\mathbf{X}} v(y) Q(dy|x,a) - \int_{\mathbf{X}} v(y) Q_k(dy|x,a) \right] \\ &= \frac{1}{\alpha + \hat{q}} \left[\int_{\mathbf{X}} v(y) q(dy|x,a) - \int_{\mathbf{X}} v(y) q_k(dy|x,a) \right] \\ &= \frac{1}{\alpha + \hat{q}} \left[\int_{\mathbf{X}} (v(y) - v(x)) p(y|x,a) \mu(dy) \right. \\ &\quad \left. - \int_{\mathbf{X}} (v(y) - v(x)) p(y|x,a) \mu_k(dy) \right], \end{aligned}$$

where the last equation is derived from Remark 2. The functions $y \mapsto v(y) - v(x)$ and $y \mapsto p(y|x,a)$ are both bounded and Lipschitz continuous; hence, their product is $(2\|v\|L_p + \|p\|L_v)$ -Lipschitz continuous. We conclude that

$$Tv(x) - T_{k,\eta}v(x) \leq \frac{1}{\alpha + \hat{q}} \cdot (2\|v\|L_p + \|p\|L_v) W(\mu, \mu_k). \quad (13)$$

Conversely, suppose that $a \in \mathbf{A}(x)$ attains the minimum in the definition of $Tv(x)$ in (5). Let $a' \in \mathbf{A}_\eta(x)$ be the closest point in $\mathbf{A}_\eta(x)$ to $a \in \mathbf{A}(x)$ with, therefore, $d_{\mathbf{A}}(a, a') \leq \eta$. We have that $T_{k,\eta}v(x) - Tv(x)$ is less than or equal to

$$\frac{c(x, a') - c(x, a)}{\alpha + \hat{q}} + \frac{\hat{q}}{\alpha + \hat{q}} \left[\int_{\mathbf{X}} v(y) Q_k(dy|x, a') - \int_{\mathbf{X}} v(y) Q(dy|x, a) \right].$$

Regarding the first term we have $|c(x, a') - c(x, a)| \leq L_c \eta$. The second term equals

$$\frac{1}{\alpha + \hat{q}} \left[\int_{\mathbf{X}} (v(y) - v(x)) p(y|x, a') \mu_k(dy) - \int_{\mathbf{X}} (v(y) - v(x)) p(y|x, a) \mu(dy) \right].$$

Adding and subtracting $\int (v(y) - v(x)) p(y|x, a') \mu(dy)$ we obtain

$$\begin{aligned} & \int_{\mathbf{X}} (v(y) - v(x)) p(y|x, a') \mu_k(dy) - \int_{\mathbf{X}} (v(y) - v(x)) p(y|x, a') \mu(dy) \\ & \leq (2\|v\|_{L_p} + \|p\|_{L_p}) W(\mu, \mu_k) \end{aligned}$$

arguing as previously, and we also obtain

$$\int_{\mathbf{X}} (v(y) - v(x)) (p(y|x, a') - p(y|x, a)) \mu(dy) \leq 2\|v\|_{L_p} \eta.$$

Summarizing, we have shown that

$$T_{k,\eta}v(x) - Tv(x) \leq \frac{L_c + 2\|v\|_{L_p}}{\alpha + \hat{q}} \cdot \eta + \frac{2\|v\|_{L_p} + \|p\|_{L_p}}{\alpha + \hat{q}} \cdot W(\mu, \mu_k).$$

Combined with (13), we obtain the desired result. \square

We define the constant \mathbf{H}_1 and \mathbf{H}_2 as follows:

$$\begin{aligned} \mathbf{H}_1 &= \frac{L_c}{\alpha} + \frac{2\|c\|_{L_p}}{\alpha^2} \\ \mathbf{H}_2 &= \frac{2\|c\|_{L_p}}{\alpha^2} + \frac{\|p\|_{L_c}(1 + L_\Psi)}{\alpha(\alpha + \hat{q}(1 - L_Q(1 + L_\Psi)))}. \end{aligned}$$

We are now ready to state our main result in the paper.

Theorem 2. *Suppose that Assumptions 1 and 2 hold and consider the control model $\mathcal{M}_{k,\eta}$. For every $\eta > 0$ and any $k \geq 1$ with $W(\mu, \mu_k) \leq \mathfrak{d}$ we have*

$$\|V^* - V_{k,\eta}^*\| \leq \mathbf{H}_1 \cdot \eta + \mathbf{H}_2 \cdot W(\mu, \mu_k).$$

Proof. Observe that, the optimal discounted cost functions V^* and $V_{k,\eta}^*$ being the respective fixed points of the operators T and $T_{k,\eta}$, we obtain

$$\begin{aligned}
\|V^* - V_{k,\eta}^*\| &= \|TV^* - T_{k,\eta}V_{k,\eta}^*\| \\
&\leq \|TV^* - T_{k,\eta}V^*\| + \|T_{k,\eta}V^* - T_{k,\eta}V_{k,\eta}^*\| \\
&\leq \|TV^* - T_{k,\eta}V^*\| + \frac{\hat{q}}{\alpha + \hat{q}}\|V^* - V_{k,\eta}^*\|,
\end{aligned}$$

and so

$$\|V^* - V_{k,\eta}^*\| \leq \frac{\alpha + \hat{q}}{\alpha} \|TV^* - T_{k,\eta}V^*\|.$$

In Lemma 1 we derived bounds on $\|TV^* - T_{k,\eta}V^*\|$, where we know that the function V^* is in $\mathbb{L}(\mathbf{X})$: its supremum norm satisfies $\|V^*\| \leq \|c\|/\alpha$ and its Lipschitz constant L_{V^*} is given in Theorem 1(iii). The result readily follows. \square

The important feature of the constants \mathbf{H}_1 and \mathbf{H}_2 is that they depend on the original data of the control model \mathcal{M} . Hence, for a given precision $\varepsilon > 0$ it is possible to determine explicitly the values of $\eta > 0$ and $W(\mu, \mu_k)$ needed to achieve the accuracy $\|V^* - V_{k,\eta}^*\| \leq \varepsilon$.

Now we address the numerical applicability of the approximation method.

Theorem 3. *Under Assumptions 1 and 2, given $\eta > 0$ and $k \geq 1$ with $W(\mu, \mu_k) \leq \mathfrak{d}$, the optimal discounted cost $V_{k,\eta}^*(x)$ of the control model $\mathcal{M}_{k,\eta}$ can be explicitly computed for any $x \in \mathbf{X}$.*

Proof. For the model $\mathcal{M}_{k,\eta}$, starting from arbitrary $x \in \mathbf{X}$ the first jump of the controlled process leads it to the support Γ_k of μ_k , and the process will remain in Γ_k afterwards. Hence, if the initial state of the system is in Γ_k , the control model $\mathcal{M}_{k,\eta}$ indeed behaves as a finite state and action system, with state space Γ_k and action sets $\mathbf{A}_\eta(x)$, for each $x \in \Gamma_k$.

Therefore, the optimal discounted cost function $V_{k,\eta}^*$ can be explicitly computed on Γ_k . To see this, just use the fixed point equation $V_{k,\eta}^* = T_{k,\eta}V_{k,\eta}^*$ (recall (12)), which is, on Γ_k , the dynamic programming equation of a discounted discrete-time Markov decision process with finite state and action spaces. This equation can be solved explicitly by using, for instance, the policy iteration algorithm, which converges in a finite number of steps.

Once we have determined $V_{k,\eta}^*$ on Γ_k , let us show how to compute $V_{k,\eta}^*(x)$ for any $x \in \mathbf{X}$ that is not in Γ_k . A careful inspection of the dynamic programming equation in Proposition 1(i) at x shows that it takes the form

$$0 = \min_{a \in \mathbf{A}_\eta(x)} \{F(a) - V_{k,\eta}^*(x)G(a)\} \quad (14)$$

where

$$F(a) = c(x, a) + \sum_{y \in \Gamma_k} V_{k,\eta}^*(y) p(y|x, a) \mu_k(\{y\})$$

and

$$G(a) = \alpha + \sum_{y \in \Gamma_k} p(y|x, a) \mu_k(\{y\}),$$

that is, F and G depend on a and on the previously computed $V_{k,\eta}^*(y)$ for $y \in \Gamma_k$. The solution of (14) is

$$V_{k,\eta}^*(x) = \min_{a \in \mathbf{A}_\eta(x)} \{F(a)/G(a)\}.$$

Hence, given any $x \in \mathbf{X}$, it is possible to compute explicitly $V_{k,\eta}^*(x)$. □

Now we discuss how to approximate the probability measure $\mu \in \mathcal{P}_1(\mathbf{X})$ in the Wasserstein distance by means of probability measures with finite support.

Deterministic Approximations

Proposition 2. *Given $\mu \in \mathcal{P}_1(\mathbf{X})$ and $\varepsilon > 0$, there exists a probability measure $\nu \in \mathcal{P}(\mathbf{X})$ with finite support such that $W(\mu, \nu) \leq \varepsilon$.*

For the proof of this result we refer to [1, Proposition 1.1]. It is based on covering the Borel space \mathbf{X} with balls of small radius. The construction of ν given in Proposition 2 controls tightly the distance $W(\mu, \nu)$ but there is not an a priori bound on the number of points in the support of ν , which is related to the dimension of \mathbf{X} .

Empirical Approximations

Given $\mu \in \mathcal{P}_1(\mathbf{X})$, consider now the probability space $(\mathbf{X}^\infty, \mathcal{B}(\mathbf{X}^\infty), \mathbb{P}_\mu)$ which is the probability space related to sampling a sequence $\{\zeta_n\}_{n \geq 1}$ of i.i.d. random variable on \mathbf{X} with distribution μ . For each $n \geq 1$, the empirical probability measure obtained from the first n samples is a random probability measure on \mathbf{X} supported on (at most) n points:

$$\mu_n = \frac{1}{n} \sum \delta_{\zeta_i}.$$

Our next result is taken from Corollary 2.5 in [4].

Proposition 3. *Given $\mu \in \mathcal{P}_{\text{exp}}(\mathbf{X})$ and $\varepsilon > 0$ there exist positive constants C_ε and D_ε such that*

$$\mathbb{P}_\mu \{W(\mu, \mu_n) > \varepsilon\} \leq C_\varepsilon \exp\{-D_\varepsilon n\} \quad \text{for all } n \geq 1.$$

With the so-defined empirical approach, we obtain convergence in probability at an exponential speed. Moreover, we control the number of points in the support of the measure μ_n but we do not have a priori knowledge of the constants C_ε and D_ε , which depend on the dimension of \mathbf{X} .

We thus observe a sort of duality between the deterministic and the empirical approaches concerning, on one hand, the accuracy of the approximation and, on the other hand, the number of points in the support of the approximating measure

needed to achieve this accuracy. Neither approach can, of course, avoid the influence of the dimension of \mathbf{X} . Concerning both the deterministic and the empirical approaches we have the following result.

Theorem 4. *Suppose that the control model \mathcal{M} satisfies Assumptions 1 and 2.*

(i) *[The deterministic approach]*

If $\mu \in \mathcal{P}_1(\mathbf{X})$ then for any $\varepsilon > 0$ there exist $\eta > 0$ and $\mu_k \in \mathcal{P}(\mathbf{X})$ with finite support such that $\|V^ - V_{k,\eta}^*\| \leq \varepsilon$.*

(ii) *[The empirical approach]*

If $\mu \in \mathcal{P}_{\text{exp}}(\mathbf{X})$ then for any $\varepsilon > 0$ there exist constants $\eta > 0$, \mathbf{C} , and \mathbf{D} such that

$$\mathbb{P}_\mu \{ \|V^* - V_{k,\eta}^*\| > \varepsilon \} \leq \mathbf{C} \exp\{-\mathbf{D}k\} \quad \text{for all } k \geq 1,$$

where μ_k is the empirical probability measure for a sample of size k .

Proof. (i). To prove this part, just choose $\eta = \frac{\varepsilon}{2\mathbf{H}_1}$ and choose a probability measure μ_k with finite support such that $W(\mu, \mu_k) \leq \delta \wedge \frac{\varepsilon}{2\mathbf{H}_2}$. The result readily follows from Theorem 2.

(ii). Choose again $\eta = \frac{\varepsilon}{2\mathbf{H}_1}$ and, for the constant $\varepsilon' = \delta \wedge \frac{\varepsilon}{2\mathbf{H}_2}$ consider $\mathbf{C} = C_{\varepsilon'}$ and $\mathbf{D} = D_{\varepsilon'}$ taken from Proposition 3. We deduce from Theorem 2 the following inclusion of sets (of \mathbf{X}^∞) for any $k \geq 1$

$$\{ \|V^* - V_{k,\eta}^*\| > \varepsilon \} \subseteq \{ W(\mu, \mu_k) > \varepsilon' \}.$$

The result now follows. □

5 Numerical example

We consider a queueing system \mathcal{M} with finite capacity. For some constant $C > 0$ let $\mathbf{X} = [0, C]$ be the state space. The action space is an interval $\mathbf{A} = [a_m, a_M] \subset \mathbb{R}$, and we put $\mathbf{A}(x) = \mathbf{A}$ for each $x \in \mathbf{X}$. The transition rates of the system are defined as

$$q(B|x, a) = \int_{B \cap (x, C]} 2(y - x)dy + a\mathbf{1}_B(0) - (a + (C - x)^2)\mathbf{1}_B(x)$$

for any $(x, a) \in \mathbf{K}$ and $B \in \mathcal{B}(\mathbf{X})$. The cost rate function c is a function in $\mathbb{L}(\mathbf{K})$, while the discount rate is some $\alpha > 0$.

Proposition 4. *The queueing system \mathcal{M} satisfies Assumptions 1 and 2 provided that*

$$\alpha + a_m > (1 + C)(2C + 1). \tag{15}$$

Proof. In the state space \mathbf{X} we will consider the usual metric on $(0, C]$, that is, $d_{\mathbf{X}}(x, y) = |x - y|$ for $0 < x, y \leq C$, and we define $d_{\mathbf{X}}(0, x) = 1 + x$ for $0 < x \leq C$. This is equivalent to identify the point 0 with -1 . We need to do so to avoid discontinuities at 0. In \mathbf{A} we consider the usual metric.

Note that the transition rates of the system are indeed bounded, and we can choose

$$\hat{q} > \sup_{(x,a) \in \mathbf{K}} \{-q(\{x\}|x,a)\} = a_M + C^2.$$

Assumptions 1(i)–(iii) are easy to verify. In particular, note that the Lipschitz constant of the action sets multifunction is $L_\Psi = 0$. To check Assumption 1(iv), given $v \in \mathbb{L}(\mathbf{X})$ we must determine the Lipschitz constant of Qv . To this end, observe that we can assume that $v(0) = 0$. We thus have

$$Qv(x,a) = \frac{1}{\hat{q}} \int_x^C 2v(y)(y-x)du + v(x) \left(1 - \frac{(C-x)^2 + a}{\hat{q}}\right) \quad \text{for } (x,a) \in \mathbf{K}.$$

Some elementary calculations (for details, consult [2, Section 5]) yield that the stochastic kernel Q is indeed Lipschitz continuous with

$$L_Q = 1 - \frac{a_m}{\hat{q}} + \frac{(1+C)(2C+1)}{\hat{q}}.$$

Finally, Assumption 1(v) follows directly from (15). Hence, we have shown that the queueing system \mathcal{M} satisfies Assumption 1.

Now we turn our attention to Assumption 2. Given $B \in \mathcal{B}(\mathbf{X})$, $0 \leq x \leq C$, and $a \in \mathbf{A}$, the positive part of the transition rates q is given by

$$q^+(B|x,a) = \begin{cases} \int_{B \cap (x,C]} 2(y-x)dy + a\mathbf{1}_B(0) & \text{if } x > 0 \\ \int_{B \cap (0,C]} 2ydy & \text{if } x = 0. \end{cases}$$

We fix an arbitrary $0 < \beta < 1$ and we define the probability measure μ on \mathbf{X} as

$$\mu(dy) = \beta \delta_0(dy) + (1-\beta) \cdot \frac{1}{C} \lambda(dy),$$

where λ is the Lebesgue measure on $(0,C]$. It is straightforward to check that the density function p defined, for $0 < x \leq C$ as

$$p(y|x,a) = \begin{cases} \frac{2C}{1-\beta} (y-x)^+ & \text{if } 0 < y \leq C \text{ and } a \in \mathbf{A} \\ \frac{a}{\beta} & \text{if } y = 0 \text{ and } a \in \mathbf{A}, \end{cases}$$

and for $x = 0$ as

$$p(y|0,a) = \begin{cases} \frac{2C}{1-\beta} y & \text{if } 0 < y \leq C \text{ and } a \in \mathbf{A} \\ 0 & \text{if } y = 0 \text{ and } a \in \mathbf{A}, \end{cases}$$

satisfies the requirements of Assumption 2(i). The so-defined function p is bounded and Lipschitz continuous in $(y,a) \in \mathbf{X} \times \mathbf{A}$ uniformly in $x \in \mathbf{X}$. Note that the partic-

ular metric we have considered in \mathbf{X} avoids the discontinuity of $p(\cdot|x,a)$ at 0 when $x > 0$. We have thus established Assumptions 2(i)–(ii).

Regarding Assumption 2(iii) we will only consider values of η less than one. Hence, for any $0 < \eta < 1$ and $x \in \mathbf{X}$, the set $\mathbf{A}_\eta(x)$ consists of $\lceil 1/\eta \rceil + 1$ equally spaced points in \mathbf{A} , that is,

$$\mathbf{A}_\eta(x) = \{a_m + j(a_M - a_m)/\lceil 1/\eta \rceil\}_{j=0,1,\dots,\lceil 1/\eta \rceil}.$$

In the sequel we will simply write \mathbf{A}_η since the above defined sets do not depend on $x \in \mathbf{X}$. It is easy to check that $\rho_{\mathbf{A}}(\mathbf{A}, \mathbf{A}_\eta) \leq \eta$. This completes the proof of the proposition. \square

Given $k \geq 1$, the probability measure $\mu_k \in \mathcal{P}(\mathbf{X})$ is defined as

$$\mu_k(dy) = \beta \delta_0(dy) + (1 - \beta) \frac{1}{k} \sum_{j=1}^k \delta_{x_j}(dy)$$

for some $x_1, \dots, x_k \in (0, C]$. The measure μ_k should be a good approximation in the Wasserstein metric of the measure μ . So, if we follow the so-called deterministic approach described above (recall Proposition 2) then we will let $x_j = jC/k$ for $1 \leq j \leq k$, which yields

$$W(\mu, \mu_k) = \frac{(1 - \beta)C}{2k}.$$

If, on the other hand, we follow the empirical approach (recall Proposition 3) then the $\{x_j\}_{1 \leq j \leq k}$ will be obtained by sampling k i.i.d. random variables uniformly distributed on $(0, C]$.

Let us now determine the transition rates of the control model $\mathcal{M}_{k,\eta}$. We denote by $\Gamma_k = \{x_0, x_1, \dots, x_k\}$ the support of the measure μ_k , where we let $x_0 = 0$. Starting from some (x_i, a) for $1 \leq i \leq k$ and $a \in \mathbf{A}_\eta$, recalling (8) and the definition of the density function p given above, it can be seen that

$$q_k(\{x_j\}|x_i, a) = 2(x_j - x_i)^+ C/k \quad \text{for } 1 \leq j \leq k \text{ with } i \neq j,$$

while $q_k(\{0\}|x_i, a) = a$ for $1 \leq i \leq k$. Starting from $(0, a)$ for $a \in \mathbf{A}_\eta$ we obtain

$$q_k(\{x_j\}|0, a) = 2x_j C/k \quad \text{for } 1 \leq j \leq k.$$

Finally, starting from $(x, a) \in \mathbf{K}_\eta$ with $x \notin \Gamma_k$ we have

$$q_k(\{x_j\}|x, a) = 2(x_j - x)^+ C/k \quad \text{for } 1 \leq j \leq k, \text{ and } q_k(\{0\}|x, a) = a.$$

The value of $q_k(\{x\}|x, a)$ is obtained just by imposing $q_k(\mathbf{X}|x, a) = 0$.

It is worth noting that the particular value of \hat{q} is irrelevant for the control models \mathcal{M} and $\mathcal{M}_{k,\eta}$. Hence, the constant ϑ in (11) can be arbitrarily large, and so the condition that $W(\mu, \mu_k) \leq \vartheta$ is not needed for this queueing model.

Numerical Experimentation

For the values of the parameters of the queueing system we choose $C = 1$, $\mathbf{A} = [7, 8]$, and $c(x, a) = (1 - x)(10 - a)$. By Proposition 4, we can choose any discount rate $\alpha > 0$. For our numerical calculations we will let $\alpha = 0.1$.

Given $k \geq 1$, for the discretization of the state space we choose $0 = x_0 < x_1 < \dots < x_k$ points as described previously. Regarding the action space, we let $\eta = 1/k$, so that we place $k + 1$ equally spaced points in \mathbf{A} . So, the approximating control model will be simply denoted by \mathcal{M}_k .

As a consequence of Theorem 3, we are now in position to determine $V_k^*(x)$ explicitly for any $x \in \mathbf{X}$.

The Deterministic Approach

Given $k \geq 1$, we consider the finite support Γ_k of μ_k consisting of the points $i \cdot C/k$ for $0 \leq i \leq k$. The actions are discretized as described previously. We can compute explicitly the optimal discounted cost $V_k^*(0)$ of the model \mathcal{M}_k for the initial state 0.

We know that $W(\mu, \mu_k) = \frac{(1-\beta)C}{2k}$. Since we have chosen $\eta = 1/k$, we derive from Theorem 2 that the error when approximating $V^*(0)$ is of order $1/k$, that is

$$|V_k^*(0) - V^*(0)| = O(1/k). \tag{16}$$

We perform the calculations for k taking the values $k = 50, 100, 150, 200, \dots, 1000$. For the 20 corresponding values of k we perform a linear regression analysis of the form

$$V_k^*(0) \sim \gamma_0 + \gamma_1 \frac{1}{k}$$

for some values $\gamma_0, \gamma_1 \in \mathbb{R}$.

Figure 1 shows the results: in red we display the 20 values of $V_k^*(0)$, while the solid blue line shows the regression line (a hyperbole since the axis of the abscissa displays k). This yields the regression coefficients

$$\hat{\gamma}_0 = 18.4668 \quad \text{and} \quad \hat{\gamma}_1 = -2.3694$$

with an almost perfect fit (the coefficient of determination is practically equal to one):

$$\max_{k=50, \dots, 1000} |V_k^*(0) - 18.4668 + 2.3694/k| = 3 \cdot 10^{-5}.$$

The estimator of $V^*(0)$ is thus $\hat{\gamma}_0 = 18.4668$. We conclude empirically that the bound (16) is tight even for relatively small values of k , and we can provide an explicit estimation of the multiplicative constant in the O term, namely

$$|V_k^*(0) - V^*(0)| \simeq \frac{2.3694}{k}.$$

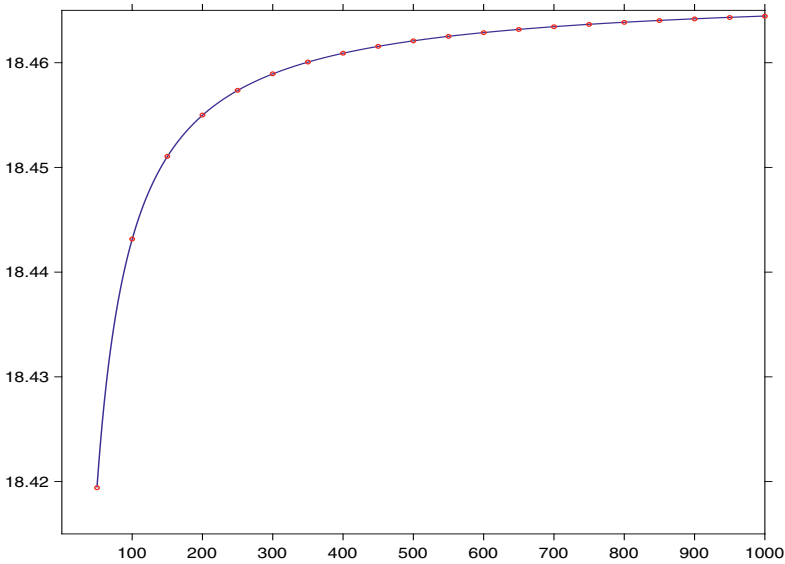


Fig. 1 Fitting $V_k^*(0) \sim \gamma_0 + \gamma_1/k$.

Most importantly, the above approximation is non-asymptotic in the sense that it is valid for relatively small values of k .

The Empirical Approach

Let ζ_1, \dots, ζ_k be i.i.d. random variables uniformly distributed on $(0, C]$. With Γ_k , the support of μ_k , consisting of the point 0 and the k values of samples, we can explicitly compute the optimal discounted cost of the approximating model \mathcal{M}_k at the initial state 0, that is, we can determine $V_k^*(0)$. Note that the so-defined $V_k^*(0)$ is in fact a random variable since it depends on the particular sample that is observed. The random variable $V_k^*(0)$ is thus interpreted as an estimator of the optimal discounted cost $V^*(0)$ of the original control model \mathcal{M} .

We take 10000 independent samples of size k of the random variables ζ_1, \dots, ζ_k , and hence we have at hand 10000 independent realizations of the random variable $V_k^*(0)$. This analysis is carried out for the values of $k = 10, 20, \dots, 120$.

Figure 2 displays the density estimation for $k = 30, 60, 90, 120$. Namely, we fit a density based on normal kernels to the 10000 observations, so as to obtain an estimation of the density of the random variable $V_k^*(0)$. We see that the density functions have approximately the same mode and they become sharper as k grows.

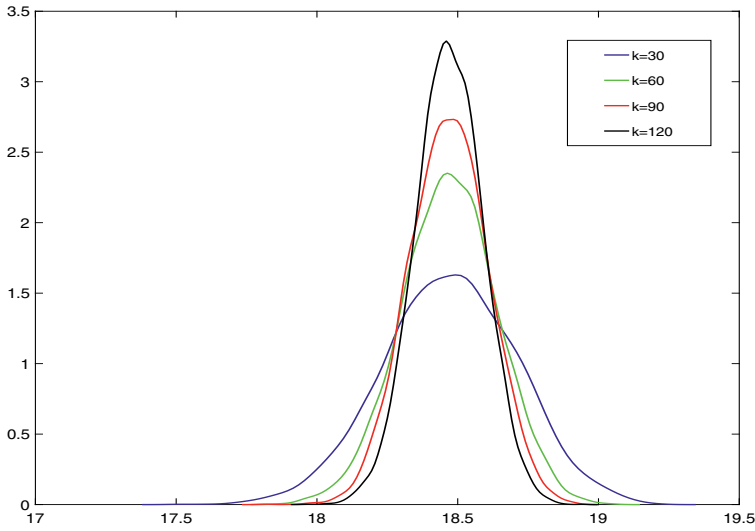


Fig. 2 Estimation of the density function of $V_k^*(0)$.

In Table 1, for several values of k , we give the mean and the standard deviation of the corresponding sample of size 10000 of the random variable $V_k^*(0)$. Note that the means are very stable, while the standard deviation decreases as k grows. The values of the mean are consistent with the approximation $\hat{\gamma}_0$ obtained with the deterministic approach.

Table 1 Sample mean and standard deviation of the $V_k^*(0)$.

	$k = 30$	$k = 60$	$k = 90$	$k = 120$
Mean	18.4715	18.4688	18.4667	18.4693
Std. Dev.	0.2375	0.1685	0.1403	0.1192

From Theorem 4 we know that, given $\varepsilon > 0$, there exist some positive constants C_ε and D_ε such that

$$\mathbb{P}_\mu\{|V_k^*(0) - V^*(0)| > \varepsilon\} \leq C_\varepsilon \exp\{-D_\varepsilon k\} \quad \text{for all } k \geq 1. \tag{17}$$

To check this inequality, we intend to approximate $\mathbb{P}_\mu\{|V_k^*(0) - V^*(0)| > \varepsilon\}$, but note that $V^*(0)$ is in fact unknown. However, we can replace it with the mean \bar{v} of the 10000 observations of $V_{120}^*(0)$ (which is presumably the most precise estimation of v^* we have at hand). For the particular case of the simulations we have made, we will let $\bar{v} = 18.4693$. Moreover, the distribution of $V_k^*(0)$ is not known, although we

have 10000 realizations of this random variable. Hence, we will compute for how many of the 10000 samples we have $|V_k^*(0) - \tilde{v}| > \epsilon$ and then we will divide it by 10000, so as to obtain the estimation $p_{k,\epsilon} \sim \mathbb{P}_\mu \{|V_k^*(0) - V^*(0)| > \epsilon\}$.

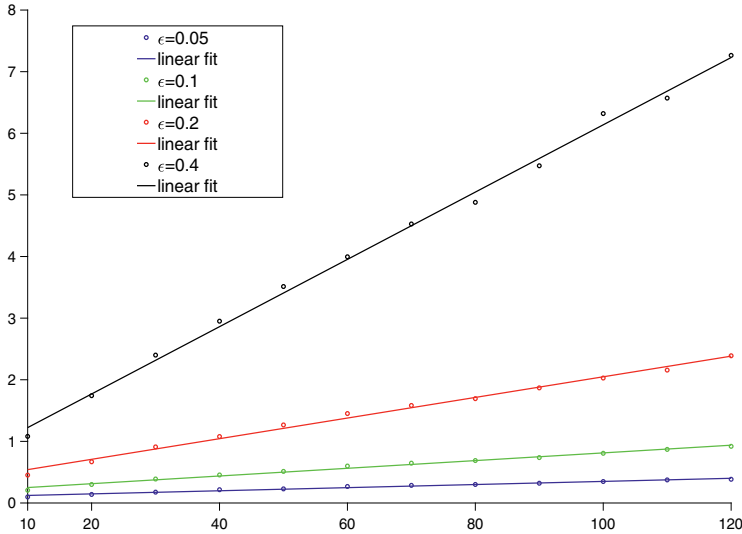


Fig. 3 Linear regression $-\log p_{k,\epsilon} \sim a_\epsilon + b_\epsilon k$.

For a choice of four values for the parameter ϵ , namely, $\epsilon = 0.05, 0.1, 0.2, 0.4$ we perform a linear regression of the form

$$-\log p_{k,\epsilon} \sim a_\epsilon + b_\epsilon k \quad \text{for } k = 10, 20, \dots, 120$$

(here, note that ϵ remains fixed while k varies).

Table 2 Linear fit of $-\log p_{k,\epsilon}$.

	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.4$
R -squared	0.9806	0.9910	0.9935	0.9968
ordinate a_ϵ	0.0971	0.1882	0.3740	0.6775
slope b_ϵ	0.0025	0.0062	0.0167	0.0546

Table 2 shows the coefficient of determination and the coefficients of the regression analysis for the particular sample we have taken, while in Figure 3 we show the corresponding linear fit. The linear fit is indeed very good and it yields the approximations for the estimation error shown in Table 3.

Table 3 Approximation of the estimation error $\mathbb{P}_\mu\{|V_k^*(0) - V^*(0)| > \varepsilon\}$.

$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.4$
$0.9074 \cdot e^{-0.0025k}$	$0.8285 \cdot e^{-0.0062k}$	$0.6880 \cdot e^{-0.0167k}$	$0.5079 \cdot e^{-0.0546k}$

This shows, empirically, not only that the bound (17) is satisfied, but that it is a tight bound. It is worth stressing that this is a non-asymptotic bound, which is accurate even for small values of the sample size k . The multiplicative constants \mathbf{C}_ε are reasonably low and they show the expected behavior (they decrease as ε grows), while the constant \mathbf{D}_ε in the exponent grows with ε , which was also to be expected.

Acknowledgements Research supported by grant MTM2016-75497-P from the Spanish Ministerio de Economía, Industria y Competitividad.

References

1. Anselmi, J., Dufour, F., Prieto-Rumeau, T.: Computable approximations for continuous-time Markov decision processes on Borel spaces based on empirical measures. *J. Math. Anal. Appl.* **443**, 1323–1361 (2016)
2. Anselmi, J., Dufour, F., Prieto-Rumeau, T.: Computable approximations for average Markov decision processes in continuous time. *J. Appl. Probab.* **55**, 571–592 (2018)
3. Bertsekas, D.P., Tsitsiklis, J.N.: *Neuro-Dynamic Programming*. Athena Scientific, Belmont (1996)
4. Boissard, E.: Simple bounds for convergence of empirical and occupation measures in 1-Wasserstein distance. *Electron. J. Probab.* **16**, 2296–2333 (2011)
5. Chang, H.S., Fu, M.C., Hu, J., Marcus, S.I.: *Simulation-based algorithms for Markov decision processes*. Communications and Control Engineering Series. Springer, London (2007)
6. Dufour, F., Prieto-Rumeau, T.: Approximation of Markov decision processes with general state space. *J. Math. Anal. Appl.* **388**, 1254–1267 (2012)
7. Dufour, F., Prieto-Rumeau, T.: Finite linear programming approximations of constrained discounted Markov decision processes. *SIAM J. Control Optim.* **51**, 1298–1324 (2013)
8. Dufour, F., Prieto-Rumeau, T.: Stochastic approximations of constrained discounted Markov decision processes. *J. Math. Anal. Appl.* **413**, 856–879 (2014)
9. Dufour, F., Prieto-Rumeau, T.: Approximation of average cost Markov decision processes using empirical distributions and concentration inequalities. *Stochastics* **87**, 273–307 (2015)
10. Guo, X., Zhang, W.: Convergence of controlled models and finite-state approximation for discounted continuous-time Markov decision processes with constraints. *European J. Oper. Res.* **238**, 486–496 (2014)
11. Hernández-Lerma, O., Lasserre, J.B.: *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, New York (1996)
12. Jacod, J.: *Calcul Stochastique et Problèmes de Martingales*. Lecture Notes in Mathematics **714**. Springer, Berlin (1979)
13. Kechris, A.S.: *Classical Descriptive Set Theory*. Springer, New York (1995)
14. Piunovskiy, A., Zhang, Y.: Discounted continuous-time Markov decision processes with unbounded rates and randomized history-dependent policies: the dynamic programming approach. *4OR* **12**, 49–75 (2014)
15. Powell, W.B.: *Approximate Dynamic Programming*. Wiley Interscience, Hoboken (2007)

16. Prieto-Rumeau, T., Hernández-Lerma, O.: Discounted continuous-time controlled Markov chains: convergence of control models. *J. Appl. Probab.* **49**, 1072–1090 (2012)
17. Prieto-Rumeau, T., Lorenzo, J.M.: Approximating ergodic average reward continuous-time controlled Markov chains. *IEEE Trans. Automat. Control* **55**, 201–207 (2010)
18. Saldi, N., Linder, T., Yüksel, S.: Asymptotic optimality and rates of convergence of quantized stationary policies in stochastic control. *IEEE Trans. Automat. Control* **60**, 553–558 (2015)
19. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, Massachusetts (1998)
20. Van Roy, B.: Neuro-dynamic programming: overview and recent trends. In: *Handbook of Markov decision processes*. *Internat. Ser. Oper. Res. Management Sci.* **40**, pp. 431–459. Kluwer Acad. Publ., Boston (2002)



Some Linear-Quadratic Stochastic Differential Games Driven by State Dependent Gauss-Volterra Processes^{*}

Tyrone E. Duncan and Bozenna Pasik-Duncan

Abstract In this paper a two player zero sum stochastic differential game in a finite dimensional space having a linear stochastic equation with a state dependent Gauss-Volterra noise is formulated and solved with a quadratic payoff for the two players and a finite time horizon. The control strategies are continuous linear state feedbacks. A Nash equilibrium is verified for the game and the two optimal strategies are obtained using a direct method that does not require solving nonlinear partial differential equations or forward-backward stochastic differential equations. The Gauss-Volterra processes are singular integrals of a standard Brownian motion and include various types of fractional Brownian motions as well as some other Gaussian processes.

Keywords. Stochastic differential games, fractional Brownian motions, Gauss-Volterra processes

Mathematics Subject Classification. Primary 93E20, 91A15; Secondary 91A35.

1 Introduction

A major difficulty for a solution of a stochastic differential game is determining explicit optimal strategies for the players. This difficulty is significantly increased if the noise process and thereby the state process are not Markov processes or martingales. The method used in this paper to determine the optimal strategies that are continuous feedback operators is direct so it does not require solving partial differential equations or special types of forward-backward stochastic differential equa-

Department of Mathematics, University of Kansas, Lawrence, KS 66045. e-mail: duncan@ku.edu and e-mail: bozenna@ku.edu

^{*} Research supported by NSF grant DMS 1411412 and AFOSR grant FA9550-12-1-0384.

tions. The noise processes are assumed to be scalar Gauss-Volterra processes that enter the stochastic equation as a product with the state so the noise is state dependent. These processes are represented as singular integrals of stochastic integrals with respect to a standard Brownian motion and they include fractional Brownian motions for the Hurst parameter $H \in (\frac{1}{2}, 1)$, Liouville fractional Brownian motions and multi-fractional Brownian motions.

A significant amount of literature exists for determining optimal strategies for stochastic differential games as well as general treatises e.g. [2], [18]. Isaacs [15] obtained a pair of nonlinear partial differential equations that become one equation if the game has a value e.g. [12] though the noise is required to be a Markov process to use these PDE methods. Another approach is to solve an appropriate forward-backward stochastic differential equation e.g. [13]. If the players' strategies are determined from the family of adapted processes of the state then the optimal strategies for these Gauss-Volterra processes will typically be functionals of the past of the state e.g. [6], [5]. Since these strategies are often not desirable practically, it is natural to restrict the strategies to be linear feedback functions of the system state to provide practicality. This latter choice of strategies is made for the present work. The authors are not aware of any other results for stochastic differential games driven by general Gauss-Volterra processes. This work was motivated by the optimal control results in [9] and the stochastic differential game results in [5].

2 Stochastic Differential Game Formulation

Initially the stochastic equation for the two person differential game is given. The state of the game is described by the following stochastic equation.

$$dX(t) = A(t)X(t)dt + B(t)U(t)dt + C(t)V(t)dt + \sigma(t)X(t)db(t) \quad (1)$$

$$X(0) = x_0 \quad (2)$$

where $X(t) \in \mathbb{R}^n$, $A(t) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, $B(t) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^{k_1})$, $U(t) \in \mathbb{R}^{k_1}$, $C(t) \in \mathcal{L}(\mathbb{R}^{k_2}, \mathbb{R}^n)$, $V(t) \in \mathbb{R}^{k_2}$, $\sigma(t) \in \mathbb{R}$, A, B, C, σ are continuous functions, (k_1, k_2) are positive integers, $(b(t), t \geq 0)$, is a real-valued Gauss-Volterra stochastic process that is defined subsequently, $x_0 \in \mathbb{R}^n$.

The payoff functional, J_T , is defined as

$$J_T(U, V) = \mathbb{E} \left[\int_0^T (\langle Q(t)X(t), X(t) \rangle + \langle R(t)U(t), U(t) \rangle - \langle S(t)V(t), V(t) \rangle) dt + \langle GX(T), X(T) \rangle \right] \quad (3)$$

where Q, R, S are continuous, symmetric and positive linear transformations and G is positive and symmetric.

Player 1 with strategy U seeks to minimize J_T and player 2 with strategy V seeks to maximize J_T . The families of strategies, \mathcal{U} and \mathcal{V} , for the two players are defined

now.

(A1)

$\mathcal{U} = \{U : [0, T] \rightarrow \mathbb{R}^{k_1} \text{ where } U(t) = K_1(t)X(t) \text{ and } K_1 \text{ is continuous}\}$

$\mathcal{V} = \{V : [0, T] \rightarrow \mathbb{R}^{k_2} \text{ where } V(t) = K_2(t)X(t) \text{ and } K_2 \text{ is continuous}\}$

Thus the strategies are restricted to be continuous linear feedbacks. The Riccati equation used to obtain optimal strategies for the two players for this stochastic differential game is

$$\frac{dP(t)}{dt} = -A^T P - PA + PBR^{-1}B^T P - PCS^{-1}C^T P - Q + \alpha P \tag{4}$$

$$P(0) = G \tag{5}$$

where α is defined subsequently. It is assumed that the Riccati equation has a unique, positive solution. Note that this Riccati equation differs from the standard one for a linear stochastic equation with a Brownian motion and a quadratic payoff. A sufficient condition for the uniqueness and positivity of the solution of the Riccati equation is that $BR^{-1}B^T - CS^{-1}C^T > 0$.

A definition of a Gauss-Volterra process is given now with some examples to indicate its range of applicability. The scalar process $(b(t), t \geq 0)$ is a Gauss-Volterra process with zero mean, that can be described by its covariance function, R , as

$$R(t, s) = \mathbb{E}[b(t)b(s)] := \int_0^{\min(t,s)} K(t, r)K(s, r)dr \tag{6}$$

where the kernel $K : \mathbb{R}_+^2 \rightarrow \mathbb{R}$ satisfies the following four conditions

- (K1) $K(t, s) = 0$ for $s > t$, $K(0, 0) = 0$, and $K(t, \cdot) \in L^2(0, t)$ for each $t \in \mathbb{R}_+$.
- (K2) For each $T > 0$ there are positive constants C, β such that

$$\int_0^T (K(t, r) - K(s, r))^2 dr \leq C|t - s|^\beta, \quad t, s \in (0, T]. \tag{7}$$

- (K3) (i) $K = K(t, s)$ is differentiable in the first variable in $\{0 < s < t < \infty\}$, both K and $\frac{\partial}{\partial t} K$ are continuous and $K(s+, s) = 0$ for each $s \in [0, \infty)$
 - (ii) $|\frac{\partial K}{\partial t}(t, s)| \leq c_T(t - s)^{\alpha-1}(\frac{t}{s})^\alpha$
 - (iii) $\int_0^t (K(t, u))^2 du \leq c_T(t - s)^{1-2\alpha}$
 on the set $\{0 < s < t < T\}$, $T < \infty$, for some constants $c_T > 0$ and $\alpha \in (0, \frac{1}{2})$.
- (K4) Let $\alpha(t) := \frac{\partial}{\partial t} (\int_0^t (\mathcal{K}_t^* \sigma)^2(r) dr)$ and assume that $\alpha \in C(\mathbb{R}_+)$ where α is in the Riccati equation (4) and

$$(\mathcal{K}_T^* \varphi)(s) := K(s+, s)\varphi(s) + \int_s^T \varphi(r)K(dr, s) \tag{8}$$

and \mathcal{K}_T^* is injective.

If σ is a constant then α simplifies as follows

$$\alpha(t) = \sigma^2 \frac{\partial}{\partial t} R(t, t) \tag{9}$$

where R is the covariance function. The kernel K has causality and continuity properties from the above first two conditions. It is assumed that there is a real-valued standard Wiener process $(W(t), t \geq 0)$ such that

$$b(t) = \int_0^t K(t, r) dW(r), \quad t \in \mathbb{R}_+ \tag{10}$$

(conditions when a Volterra process admits such a representation (10) have been obtained, cf. [11], [14]). From (K2) it easily follows by the Kolmogorov sample path continuity test that $(b(t), t \geq 0)$ has a continuous modification, which is the version that is chosen for the subsequent discussion. It is assumed that for all $s \in [0, T)$, $T > 0$, $K(\cdot, s)$ has bounded variation on the interval (s, T) and

$$\int_0^T |K|((s, T], s)^2 ds < \infty \tag{11}$$

where $|K|$ denotes the variation of K . Three examples of Gauss-Volterra processes are

(i) A fractional Brownian motion (FBM) with the Hurst parameter $H \in (\frac{1}{2}, 1)$. In this case

$$\begin{aligned} K(t, s) &= C_H s^{1/2-H} \int_s^t (u-s)^{H-3/2} u^{H-1/2} du, & s < t, \\ &= 0 & t \leq s. \end{aligned} \tag{12}$$

The kernel satisfies conditions (K1)–(K3) with $\alpha = H - \frac{1}{2}$.

ii) The Liouville fractional Brownian motion (LFBM, cf. [3]) for $H \in (\frac{1}{2}, 1)$, in which case

$$K(t, s) = C_H (t-s)^{H-\frac{1}{2}} 1_{(0,t]}(s), \quad t > s, \quad t, s \in \mathbb{R}_+ \tag{13}$$

satisfies (K1)–(K3) with $\alpha = H - \frac{1}{2}$.

iii) The multifractional Brownian motion (MBM). A simplified version analogous to LFBM in the above Example (ii) is considered. The kernel $K : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is defined as

$$K(t, s) = (t-s)^{H(t)-\frac{1}{2}} 1_{(0,t]}(s), \quad t, s \in \mathbb{R}_+,$$

where $H : \mathbb{R}_+ \rightarrow [\frac{1}{2}, 1)$ is the “time-dependent Hurst parameter”.

The evolution operator to determine the solution for (1) can be factored as described now. Let $T_{K_1, K_2}(t, s), t \geq s$ be the evolution operator for the linear operator family $(A(t) + B(t)K_1(t) + C(t)K_2(t) - \frac{1}{2}\alpha(t)I, t \geq 0)$. There is an additional term for the solution of (1) that depends explicitly on the noise term $(b(t), t \geq 0)$ as follows

$$T_b(t) = \exp\left(\int_0^t \sigma(t)db(t)\right) \tag{14}$$

The (Wiener-type) stochastic integral in the exponential can be simply defined from Malliavin calculus.

The evolution operator for the solution of (1) is the product of the above two evolution operators, T_{K_1, K_2} and T_b . The product of the last term in T_{k_1, K_2} and the operator T_b is the solution of a geometric stochastic equation for $(b(t), t \geq 0)$ using the Ito formula in [1] in analogy to the well known solution for a geometric equation for a Brownian motion and its solution that provides a geometric Brownian motion.

3 Optimal Strategies

The main result in this paper is the following theorem that provides optimal feedback gain control strategies for the two players and shows that these strategies form a Nash equilibrium in the families of admissible strategies..

Theorem 1. *Let K1-K4 be satisfied. The stochastic differential game described by the equation (1) and the payoff functional (3) for the families of strategies (A1) has optimal feedback strategies, (K_1^*, K_2^*) , given by*

$$K_1^*(t) = -R^{-1}(t)B(t)P(t) \tag{15}$$

$$K_2^*(t) = S^{-1}(t)C(t)P(t) \tag{16}$$

for $t \in [0, T]$. The optimal payoff, $J_T(U^*, V^*)$, where $U^* = K_1^*X, V^* = K_2^*X$, is

$$J_T(U^*, V^*) = \langle P(0)x_0, x_0 \rangle \tag{17}$$

where P is the unique positive, symmetric solution of the Riccati equation (4). These two optimal strategies form a Nash equilibrium for this stochastic differential game.

Proof. The determination of the optimal strategies for the two players follows the methods in [5] for a Brownian motion noise using the Riccati equation (4). Let $Y(t) = \langle P(t)X(t), X(t) \rangle$ for $t \geq 0$. Apply the Ito formula [1] to $(Y(t), t \geq 0)$ to obtain the following equality.

$$\begin{aligned} Y(T) - Y(0) = & \int_0^T [2(\langle P(t)X(t), A(t)X(t) \rangle + \langle P(t)X(t), B(t)K_1(t)X(t) \\ & + C(t)K_2(t)X(t) \rangle)dt + \langle P(t)X(t), \sigma(t)X(t) \rangle db(t)) \\ & + \langle \frac{dP}{dt}X(t), X(t) \rangle dt - \alpha(t)\langle X(t), X(t) \rangle dt] \end{aligned} \tag{18}$$

Using the Riccati equation to replace $\frac{dP}{dt}$ and adding and subtracting $\langle RX, X \rangle - \langle SX, X \rangle$ in the integrand (18) it follows that the payoff has the following expression.

$$\begin{aligned}
J_T(K_1, K_2) = \mathbb{E} \int_0^T & |R^{\frac{1}{2}}(K_1X + R^{-1}BPX)|^2 - |S^{\frac{1}{2}}(K_2X - S^{-1}CPX)|^2 dt \\
& + 2\mathbb{E} \int_0^T \langle PX, X \rangle \sigma db + \langle P(0)x_0, x_0 \rangle
\end{aligned} \tag{19}$$

The stochastic integral term in (19) has expectation zero because the stochastic integral is a Skorokhod integral so the minimization and the maximization for the two players are clear and thus the optimal strategies are

$$K_1^*(t) = -R^{-1}(t)B(t)P(t) \tag{20}$$

$$K_2^*(t) = S^{-1}(t)C(t)P(t) \tag{21}$$

The optimal payoff follows from (19). Furthermore it follows from (19) that these feedback strategies form a Nash equilibrium and this completes the proof.

4 Concluding Remarks

The result on optimal strategies in this paper demonstrates that optimal strategies can be obtained for a wide variety of Gaussian processes. A natural generalization of this stochastic game problem is to consider some noise processes that are not Gaussian. Probably the simplest of these processes is the family of Rosenblatt processes e.g. [17]. Some initial work on linear-quadratic control with an infinite time horizon for these latter processes is given in [4]. Furthermore it seems that these results can be extended to mean field games following [10].

References

1. E. Alos, O. Mazet, and D. Nualart: Stochastic calculus with respect to Gaussian processes. *Ann. Probab.*, 29 (2001), 766–801.
2. T. Basar and P. Bernhard, *H[∞]-Optimal Control and Related Minimax Design Problems*, Birkhauser, Boston, 1995
3. Z. Brzeźniak, J. van Neerven and D. Salopek, Stochastic evolution equations driven by Liouville Fractional Brownian Motion, *Czechoslovak Math J.* 62, (2012) 1–27.
4. P. Čoupek, T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, An infinite time horizon linear-quadratic control problem with a Rosenblatt process, *Proc. 57th IEEE Conf. Decision and Control*, 4973–4977, Miami, Dec. 2018.
5. T. E. Duncan, Linear-quadratic stochastic differential games with general noise processes, *Models and Methods in Economics and Management Science: Essays in Honor of Charles S. Tapiero*, (eds. F. El Ouardighi and K. Kogan), Operations Research and Management Series, Springer Intern. Publishing, Switzerland, Vol. 198, 2014, 17–26.
6. T. E. Duncan and B. Pasik-Duncan, Linear-quadratic fractional Gaussian control, *SIAM J. Control Optim.*, 51 (2013), 4604–4619.
7. T. E. Duncan and B. Pasik-Duncan, Explicit strategies for some linear and nonlinear stochastic differential games, *J. Math. Engrg. Sci. Aerospace*, 7 (2016), 83–92.

8. T.E. Duncan, B. Maslowski and B. Pasik-Duncan: Ergodic control of linear stochastic equations in a Hilbert space with fractional Brownian motion. Stochastic analysis, 91-102, *Banach Center Publ.*, 105, Polish Acad. Sci. Inst. Math., Warsaw, 2015
9. T. E. Duncan, B. Maslowski and B. Pasik-Duncan, Linear stochastic differential equations driven by Gauss-Volterra processes and related linear-quadratic control problems, *Appl. Math. Optim.*, 2018, to appear.
10. T. E. Duncan and H. Tembine, Linear-quadratic mean-field-type games: A direct approach, *Games*, 9 (1), Feb. 2018.
11. M. Erraoui and E. H. Essaky: Canonical representation for Gaussian processes, *Séminaire de Probabilités*, XLII (2009), 365-381.
12. W. H. Fleming and D. Hernandez-Hernandez, On the value of stochastic differential games, *Commun. Stoch. Anal.* 5 (2011), 241-251.
13. S. Hamadene, J. P. Lepeltier, and S. Peng, BSDEs with continuous coefficients and stochastic differential games, in *Backward Stochastic Differential Equations* N. El Karoui et al., eds., Pitman Res. Notes Math, 1997, 115-128.
14. T. Hida, Canonical representations of Gaussian processes and their applications, *Memoirs of College of Science*, Kyoto University, XXXIII (1960), 109-155.
15. R. Isaacs, *Differential Games*, J. Wiley, New York 1965.
16. J. Nash, Non-cooperative games, *Ann. Math.* 54 (1951), 286-295.
17. M. Rosenblatt, Independence and dependence, *Proc. 4th Berkeley Symp. Math. Stat. and Probab.* Vol. II, Univ. Cal. Press 1961.
18. J. Yong, *Differential Games. A Concise Introduction*, World Scientific, 2015



Correlated Equilibria for Infinite Horizon Nonzero-Sum Stochastic Differential Games

Beatris A. Escobedo-Trujillo and Héctor Jasso-Fuentes

Abstract This chapter is about two-person nonzero-sum stochastic differential games with discounted and long-run average (a.k.a. ergodic) payoffs. Our aim is to give conditions for the existence of *feedback* correlated randomized equilibria for each aforementioned payoff that are natural generalizations of the well-known Nash equilibria. To do so, we rewrite our original problem in terms of an auxiliary zero-sum game, so that the way to find correlated equilibria is based on some properties of this later game. Key ingredients to achieve the desired results are the continuity properties of the payoffs.

1 Introduction

Nash equilibrium is a very useful concept in game theory, however it is well known that under standard conditions the existence of Nash equilibria in nonzero-sum games with uncountable state-action spaces is not necessarily guaranteed within the set of randomized strategies.

During the past decades, there has been works that have dealt to game models with particular features in order to ensure the existence of Nash equilibria; for instance, games with an additive structure (see, e.g. [8, 16, 17]). Other works have explored an alternative method, which consists of “relaxing” the idea of Nash equilibrium. The idea is to extend the set of strategies into a bigger one, giving rise to the concept of *correlated strategies* as well as to the concept of *correlated equilibria*.

Beatris A. Escobedo-Trujillo
Engineering Faculty. Universidad Veracruzana, Coatzacoalcos, Ver., México. Tel. (921) 218-77-83, 211-57-07. e-mail: bescobedo@uv.mx

Héctor Jasso-Fuentes
Departamento de Matemáticas, CINVESTAV-IPN. Apartado Postal 14-740, México. D.F., 07000, México. e-mail: hjasso@math.cinvestav.mx

ria (see e.g. [4, 5, 20, 21]). This approach is the one we have focused on in this manuscript, whose details will be explained in later sections.

Recall that the Nash equilibrium concept means that if one player tries to alter his strategy unilaterally, he cannot improve his performance by such a change. If players choose their strategies according to the Nash equilibrium concept, they are said to play non-cooperatively, i.e., each player is only interested in maximizing his own utility. The correlated equilibrium concept means that all players, before taking a decision over the strategies, receive a global (or joint) recommendation, that is drawn randomly according to a joint distribution μ , then no player has an incentive to divert from the recommendation, provided that all other players follow theirs.

The main distinguishing feature of the concept of correlated equilibrium, unlike the definition of Nash equilibrium is that those recommendations do not need to be independent; i.e., the joint distributions do not need to be a product of marginal distributions. It turns out that a correlated equilibrium μ is a Nash equilibrium if and only if μ is a product measure.

It is well recognized that correlated equilibria were introduced by Aumann in 1974 for nonzero-sum games in normal form, extending the Nash equilibrium concept, [4, 5]. There exists a vast number of manuscripts that are focused on correlated strategies concept providing conditions for the existence of correlated equilibria [4, 5, 10, 20, 21, 22, 23, 24], this is, in some part, because it is easier to prove the existence and characterize correlated equilibria compared with Nash equilibria.

The work is inspired by the paper [20] which deals with correlated relaxed equilibria in nonzero-sum linear differential games with finite-horizon payoffs. Our aim here is to prove the existence of *feedback* correlated equilibria for a more general dynamic when the payoffs are of the (infinite horizon) discounted and average type. A key point to obtain our desired equilibria is to guarantee the continuity to both payoff functions (discounted and average payoffs) within the set of correlated strategies.

Although we restrict ourselves to the case of two players, its relatively easy to extend our results to the more general context of N players.

The main novelty of the manuscript lies in the fact that we are working with infinite-horizon (discounted and ergodic) payoff criteria under a considerable general diffusion process. Furthermore, the set of correlated equilibria are shown to be *feedback*, meaning that they are dependent on the current state of the game. To the best of our knowledge, this treatment has not been already studied in the current literature.

This chapter is organized as follows: In section 2, we introduce both the game and the payoffs we are trying to optimize. We also define the Nash equilibrium concept. Section 3 is devoted to the introduction of correlated strategies. We extend the domain of our payoffs over these strategies and will prove the continuity of those criteria. By last, in section 4 we introduce the concept of correlated equilibria and prove the existence of them. To do so, we rewrite the original game as a zero-sum game and we explore some of its properties. Correlated equilibria will be obtained through the use of some min-max theorems as well as for the continuity of our payoffs.

Notation and terminology.

- For some $m, n \in \mathbb{N}$, let $\mathcal{O} \subset \mathbb{R}^m$ and $V \subset \mathbb{R}^n$ be given open and Borel sets, respectively. We define:
 - $\mathbb{W}^{l,p}(\mathcal{O})$ the Sobolev space consisting of all real-valued measurable functions h on \mathcal{O} such that $D^\lambda h$ exists for all $|\lambda| \leq l$ in the weak sense and it belongs to $\mathbb{L}^p(\mathcal{O})$, where

$$D^\lambda h := \frac{\partial^{|\lambda|} h}{\partial x_1^{\lambda_1}, \dots, \partial x_m^{\lambda_m}} \quad \text{with } \lambda = (\lambda_1, \dots, \lambda_m), \quad \text{and } |\lambda| := \sum_{i=1}^m \lambda_i.$$

- $\mathbb{C}^k(\mathcal{O})$ the space of all real-valued continuous functions on \mathcal{O} with continuous l -th partial derivative in $x_i \in \mathbb{R}$, for $i = 1, \dots, m, l = 0, 1, \dots, k$. In particular, when $k = 0$, $\mathbb{C}^0(\mathcal{O})$ stands for the space of real-valued continuous functions on \mathcal{O} .
- $\mathbb{C}^{k,\beta}(\mathcal{O})$ the subspace of $\mathbb{C}^k(\mathcal{O})$ consisting of all those functions h such that $D^\lambda h$ satisfies a Hölder condition with exponent $\beta \in (0, 1]$, for all $|\lambda| \leq k$.
- $\mathbb{C}_b(\mathcal{O} \times V)$ the space consisting of all continuous bounded functions on $\mathcal{O} \times V$.

- For vectors x and matrices A we use the usual Euclidean norms

$$|x|^2 := \sum_i x_i^2 \quad \text{and} \quad |A|^2 := Tr(AA') = \sum_{i,j} A_{ij}^2,$$

where A' and $Tr(\cdot)$ denote the transpose and the trace of a square matrix, respectively.

- For any two strategies, say π^1 and π^2 , the notation $\pi^1 \times \pi^2$ means the product measure associated to this pair.

2 The game model

Consider an m -dimensional diffusion process $x(\cdot)$ controlled by two players and evolving according to the stochastic differential equation

$$dx(t) = b(x(t), u_1(t), u_2(t))dt + \sigma(x(t))dW(t), \quad x(0) = x_0, \quad (1)$$

where $b : \mathbb{R}^m \times U_1 \times U_2 \rightarrow \mathbb{R}^m$, $\sigma : \mathbb{R}^m \rightarrow \mathbb{R}^{m \times d}$ are given functions, and $W(\cdot)$ is a d -dimensional standard Brownian motion. The sets $U_1 \subset \mathbb{R}^{m_1}$ and $U_2 \subset \mathbb{R}^{m_2}$ are Borel sets called the action set for player 1 and player 2, respectively. Moreover, for $k = 1, 2$, $u_k(\cdot)$ is a non-anticipative U_k -valued stochastic process representing the control actions of player k at each time $t \geq 0$.

For $(u_1, u_2) \in U_1 \times U_2$, and h in $\mathbb{W}^{2,p}(\mathbb{R}^m)$, let

$$L^{u_1, u_2} h(x) := \sum_{i=1}^m b_i(x, u_1, u_2) \frac{\partial h}{\partial x_i}(x) + \frac{1}{2} \sum_{i,j}^m a^{ij}(x) \frac{\partial^2 h}{\partial x_i \partial x_j}(x), \quad (2)$$

where b_i is the i -th component of b , and a^{ij} is the (i, j) -component of the matrix $a(\cdot) := \sigma(\cdot)\sigma'(\cdot)$.

Let us now proceed to define the sets of strategies allowed for each player.

Randomized Markov strategies. Let $\mathcal{B}(U_1)$ be the Borel σ -algebra of U_1 , and let $\mathcal{P}(U_1)$ be the space of probability measures on U_1 . In the same way, we define $\mathcal{B}(U_2)$ and $\mathcal{P}(U_2)$ associated to player 2.

Definition 1. A *randomized Markov strategy* for player k ($k = 1, 2$) is defined as a family $\pi^k := \{\pi_t^k : t \geq 0\}$ of stochastic kernels on $\mathcal{B}(U_k) \times \mathbb{R}^m$; that is:

- (a) for each $t \geq 0$ and $x \in \mathbb{R}^m$, $\pi_t^k(\cdot|x)$ is in $\mathcal{P}(U_k)$, satisfying $\pi_t^k(U_k|x) = 1$;
- (b) for each $D \in \mathcal{B}(U_k)$ and $t \geq 0$, $\pi_t^k(D|\cdot)$ is a Borel function on \mathbb{R}^m ; and
- (c) for each $B \in \mathcal{B}(U_k)$ and $x \in \mathbb{R}^m$, the function $t \mapsto \pi_t^k(B|x)$ is a Borel measurable function.

Definition 2. A randomized strategy $\pi^k = \{\pi_t^k : t \geq 0\}$ ($k = 1, 2$) is said to be *stationary* if there is a stochastic kernel π^k on $\mathcal{B}(U_k) \times \mathbb{R}^m$ such that $\pi_t^k(\cdot|x) = \pi^k(\cdot|x)$ for all $x \in \mathbb{R}^m, t \geq 0$.

The set of randomized stationary strategies for player k is denoted by $\Pi_k, k = 1, 2$.

Next we define the payoff functions that each player wants to “optimize.”

Payoff rates. For each player $k = 1, 2$, let $r_k : \mathbb{R}^m \times U_1 \times U_2 \rightarrow \mathbb{R}$ be a measurable function, which we will call the payoff rate of player k ; in this sense, at each $t \geq 0$, $r_k(x(t), u_1, u_2)$ is the payoff of player k at time t , when the actions $u_1 \in U_1$ and $u_2 \in U_2$ are decided by players 1 and 2, respectively.

Throughout this manuscript we will use the notation $\pi^1 \times \pi^2$, representing the product measure of π^1 and π^2 .

Let the function ψ be either b or $r_k, k = 1, 2$. When players use a stationary randomized strategy $(\pi^1 \times \pi^2) \in \Pi_1 \times \Pi_2$, we write:

$$\psi(x, \pi^1 \times \pi^2) := \int_{U_1} \int_{U_2} \psi(x, u_1, u_2) \pi^1(du_1|x) \pi^2(du_2|x), \quad x \in \mathbb{R}^m.$$

With the above notation, the infinitesimal generator (2) is written as

$$L^{\pi^1 \times \pi^2} h(x) := \sum_{i=1}^m b_i(x, \pi^1 \times \pi^2) \frac{\partial h}{\partial x_i}(x) + \frac{1}{2} \sum_{i,j}^m a^{ij}(x) \frac{\partial^2 h}{\partial x_i \partial x_j}(x), \quad x \in \mathbb{R}^m.$$

Assume for the moment the existence of the probability measure $\mathbb{P}_x^{\pi^1 \times \pi^2}$ for each $x \in \mathbb{R}^m$ and $\pi^1 \times \pi^2 \in \Pi_1 \times \Pi_2$, associated to the process $x(\cdot)$. We will also denote by $\mathbb{E}_x^{\pi^1 \times \pi^2}(\cdot)$ its respective expectation. Next define the payoff criteria each player would be interested to optimize.

Definition 3 (Discounted payoff criterion). Let $\alpha > 0$, and consider the payoff rates r_1 and r_2 . For each player $k = 1, 2$, the expected α -discounted payoff for player k when players use the strategy $(\pi^1 \times \pi^2) \in \Pi_1 \times \Pi_2$ given the initial state $x \in \mathbb{R}^m$, is

$$V_k(x, \pi^1 \times \pi^2) := \mathbb{E}_x^{\pi^1 \times \pi^2} \left[\int_0^\infty e^{-\alpha t} r_k(x(t), \pi^1 \times \pi^2) dt \right]. \quad (3)$$

Definition 4 (Average payoff criterion). For each player $k = 1, 2$, the expected average payoff for player k when players use the strategy $(\pi^1 \times \pi^2) \in \Pi_1 \times \Pi_2$ given the initial state $x \in \mathbb{R}^m$, is

$$J_k(x, \pi^1 \times \pi^2) := \liminf_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_x^{\pi^1 \times \pi^2} \left[\int_0^T r_k(x(t), \pi^1 \times \pi^2) dt \right]. \quad (4)$$

In a noncooperative N -person nonzero-sum stochastic differential game, each player tries to maximize (or minimize) his/her individual performance criterion (in particular, criteria of type (3) and (4)). A Nash equilibrium, in this case, is a strategy such that once it is chosen by the players, no player will profit unilaterally by simply changing his/her own strategy. More specifically, in the maximization context, we have the next definition:

Definition 5 (Nash equilibrium). Let $F_k, k = 1, 2$, be either the discounted payoff in (3) or the average payoff (4). A randomized pair of strategies $(\pi_*^1 \times \pi_*^2) \in \Pi_1 \times \Pi_2$ is a Nash equilibrium if and only if

$$\begin{aligned} F_1(x, \pi_*^1 \times \pi_*^2) &\geq F_1(x, \pi^1 \times \pi_*^2), \quad \forall \pi^1 \in \Pi_1, \\ F_2(x, \pi_*^1 \times \pi_*^2) &\geq F_2(x, \pi_*^1 \times \pi^2), \quad \forall \pi^2 \in \Pi_2. \end{aligned}$$

It is also well recognized that the existence of Nash equilibria in nonzero-sum games with uncountable state-action spaces is not necessarily guaranteed in the set of randomized policies $\Pi_1 \times \Pi_2$ under standard conditions (for example, the conditions established in this chapter); however, such existence is achieved under special cases. For instance, we can assume the drift b in the dynamic (1) and the payoff rates r_k both satisfy an additive structure property (see, for instance, [8, 16, 17]). One alternative to avoid restrictive assumptions to the model, is to extend the concept of Nash equilibrium into a bigger set of $\Pi_1 \times \Pi_2$, which entails, in particular, to the concept of *correlated strategies* and its corresponding *correlated equilibrium*.

In the next section we will provide, among other things, conditions so that the dynamic (1) can attain a unique solution in some sense and that the payoffs in (3) and (4) are finite valued on the set of correlated strategies.

3 Correlated strategies.

In this work we extend the set of strategies available to the players. These strategies allow players to correlate their decisions during a pre-play communication process (see Section 4).

Definition 6 (Correlated strategy). A correlated (stationary) randomized strategy μ is a stochastic kernel on $\mathcal{B}(U_1 \times U_2) \times \mathbb{R}^m$ such that:

- (a) for each $x \in \mathbb{R}^m$, $\mu(\cdot|x)$ is a joined probability measure on $U_1 \times U_2$ and such that $\mu(U_1 \times U_2|x) = 1$.
- (b) For each $D \in \mathcal{B}(U_1 \times U_2)$, $\mu(D|\cdot)$ is Borel measurable on \mathbb{R}^m .

We will denote by Γ the set of all correlated randomized strategies. On the other hand, the marginal distributions of μ are defined as:

$$\mu_1(B_1|x) := \mu(B_1 \times U_2|x) \quad \text{and} \quad \mu_2(B_2|x) := \mu(U_1 \times B_2|x),$$

for each Borel set $B_k \in \mathcal{B}(U_k)$, $k = 1, 2$, and $x \in \mathbb{R}^m$.

Let ψ be either b , r_1 or r_2 . When players use a correlated strategy $\mu \in \Gamma$, we write

$$\psi(x, \mu) := \int_{U_1 \times U_2} \psi(x, u_1, u_2) \mu(d(u_1, u_2)|x),$$

Furthermore, the generator (2) turns out to be

$$L^\mu h(x) := \sum_{i=1}^m b_i(x, \mu) \frac{\partial h}{\partial x_i}(x) + \frac{1}{2} \sum_{i,j} a^{ij}(x) \frac{\partial^2 h}{\partial x_i \partial x_j}(x), \quad x \in \mathbb{R}^m. \quad (5)$$

We denote by Γ_k the set of k -marginal measures associated to Γ , for $k = 1, 2$.

Remark 1. Throughout this work we will assume that the players choose only randomized *stationary* strategies. The reason is that, even when it is possible to work in a more general class of strategies (for instance that of the so-named non-anticipative policies), our present hypotheses (stated later on) ensure the existence of optimal policies in the class of stationary strategies for all players. Further, it is worth to mention that recurrence and ergodicity properties of the state system (1) can be easily verified through the use of stationary strategies, but for general non-anticipative strategies, the corresponding state system might be time-inhomogeneous; a fact that can be hard to handle.

Assumption 1 Recall the elements of the dynamic (1). We assume:

- (a) The action sets U_1 and U_2 are compact.
- (b) The function $b : \mathbb{R}^m \times U_1 \times U_2 \rightarrow \mathbb{R}^m$ satisfies the following conditions:
 - (i) it is continuous on $\mathbb{R}^m \times U_1 \times U_2$.

(ii) it satisfies a Lipschitz condition uniformly in $(u_1, u_2) \in U_1 \times U_2$; that is, there exists a positive constant K_1 such that, for all $x, y \in \mathbb{R}^m$,

$$\sup_{(u_1, u_2) \in U_1 \times U_2} |b(x, u_1, u_2) - b(y, u_1, u_2)| \leq K_1 |x - y|.$$

(c) There exists a positive constant K_2 such that for all $x, y \in \mathbb{R}^m$,

$$|\sigma(x) - \sigma(y)| \leq K_2 |x - y|.$$

(d) (Uniform ellipticity). The matrix $a(x) = \sigma(x)\sigma'(x)$ satisfies that, for some constant $c_0 > 0$

$$xa(y)x' \geq c_0 |x|^2 \text{ for all } x, y \in \mathbb{R}^m.$$

Remark 2. (a) Assumption 1 ensures that there exists an almost surely unique strong solution of (1), for each strategy $\mu \in \Gamma$, which is a Markov–Feller process and whose infinitesimal generator coincides with L^μ in (5). (For more details, see the arguments of [2, Theorem 2.2.7]).

(b) The aforementioned existence and uniqueness remain valid for special types of joint kernels of either form $\mu = \pi_1 \times \pi_2$, or $\mu = \pi_1 \times \mu_2$ or $\mu = \mu_1 \times \pi_2$, for every $\pi_1 \in \Pi_1$, $\pi_2 \in \Pi_2$, $\mu_1 \in \Gamma_1$, $\mu_2 \in \Gamma_2$. This implies that the dynamic (1) is well defined even when a usual pair of strategies $(\pi_1 \times \pi_2) \in \Pi_1 \times \Pi_2$ as that introduced in Definition 2 is applied.

The following assumption is a Lyapunov–like condition that guaranties, in particular, that the discounted payoff criterion (3) is finite, among other facts such as the positive recurrence property of the diffusion (1) and the existence of an invariant measure, each of them for a suitable type of controls (or strategies) $u_1(\cdot)$ and $u_2(\cdot)$.

Assumption 2 There exists a function $w \in \mathbb{C}^2(\mathbb{R}^m)$, with $w \geq 1$, and constants $d \geq c > 0$ such that

(i) $\lim_{|x| \rightarrow \infty} w(x) = +\infty$, and

(ii) $L^{u_1, u_2} w(x) \leq -cw(x) + d$ for each $(u_1, u_2) \in U_1 \times U_2$ and $x \in \mathbb{R}^m$.

Remark 3. An easy application of Ito’s formula to $e^{ct}w(x(t))$ along with Assumption 2(ii), give us that

$$\sup_{\mu \in \Gamma} \mathbb{E}_x^\mu (w(x(t))) \leq e^{-ct}w(x) + \frac{d}{c}(1 - e^{-ct}).$$

Definition 7. Let $w \geq 1$ be the function in Assumption 2 and $\mathcal{O} \subset \mathbb{R}^m$ be an open set. We define the Banach space $\mathbb{B}_w(\mathcal{O})$ consisting of real–valued measurable functions h on \mathcal{O} with finite w –norm defined as follows:

$$\|h\|_w := \sup_{x \in \mathcal{O}} \frac{|h(x)|}{w(x)}.$$

We also include another set of hypotheses related to the payoffs r_k that uses the above definition.

Assumption 3 (a) *The function $r_k(x, u_1, u_2)$ is continuous on $\mathbb{R}^m \times U_1 \times U_2$ and locally Lipschitz in x uniformly with respect to $(u_1, u_2) \in U_1 \times U_2$; that is, for each $R > 0$, there exists a constant $K(R) > 0$ such that*

$$\sup_{(u_1, u_2) \in U_1 \times U_2} |r_k(x, u_1, u_2) - r_k(y, u_1, u_2)| \leq K(R)|x - y| \text{ for all } |x|, |y| \leq R.$$

(b) *$r_k(\cdot, u_1, u_2)$ is in $\mathbb{B}_w(\mathbb{R}^m)$ uniformly in (u_1, u_2) ; that is, there exists $M > 0$ such that for all $x \in \mathbb{R}^m$*

$$\sup_{(u_1, u_2) \in U_1 \times U_2} |r_k(x, u_1, u_2)| \leq Mw(x).$$

We will extend the discounted payoff criteria (3) and (4) on the set Γ .

Extended discounted criterion: Given r_k as in Assumption 3, $k = 1, 2$, and for any initial state $x \in \mathbb{R}^m$, the extended α -discounted payoff for player k when the strategy $\mu \in \Gamma$ is applied is defined as

$$V_k(x, \mu) := \mathbb{E}_x^\mu \left[\int_0^\infty e^{-\alpha t} r_k(x(t), \mu) dt \right]. \tag{6}$$

The following proposition is a direct consequence of Assumptions 2, 3(b), and Remark 3, so we shall omit the proof; similar arguments can be founded in [9, Proposition 9.1].

Proposition 1. *Under the Assumptions 1, 2, 3, the payoff (6) belongs to the space $\mathbb{B}_w(\mathbb{R}^m)$ for each correlated strategy μ ; in fact, for each x in \mathbb{R}^m we have*

$$|V_k(x, \mu)| \leq M(\alpha)w(x) \tag{7}$$

with $M(\alpha) := M \frac{(\alpha+d)}{\alpha c}$. Here, c and d are the constants in Assumption 2(b), and M is the constant in Assumption 3(b).

Extended average criterion: Let r_k be as in Assumption 3, $k = 1, 2$, and $x \in \mathbb{R}^m$. We define the extended average payoff for player k when the strategy $\mu \in \Gamma$ is used and initial state x , as follows

$$J_k(x, \mu) := \liminf_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_x^\mu \left[\int_0^T r_k(x(t), \mu) dt \right]. \tag{8}$$

Note that the above limit always exists. Actually, we will impose an ergodicity condition so that this limit becomes a constant in some sense.

From the arguments in [2, 12], for each $\mu \in \Gamma$, the Markov process $x(\cdot)$ is positive recurrent and admits a unique invariant probability measure η_μ , for which

$$\eta_\mu(w) := \int_{\mathbb{R}^m} w(x) \eta_\mu(dx) < \infty, \tag{9}$$

where w is the function defined in Assumption 2. The next assumption corresponds to the well-known asymptotic behaviour of $x(t)$ when t goes to the infinite. Sufficient conditions for this assumption can be seen in Theorem 2.7 in [14].

Assumption 4 *For every $\mu \in \Gamma$, the process $x(\cdot)$ is uniformly w -exponentially ergodic; that is, there exist positive constants k_1 and k_2 such that*

$$\sup_{\mu \in \Gamma} \left| \mathbb{E}_x^\mu [v(x(t))] - \eta_\mu(v) \right| \leq k_1 \|v\|_w e^{-k_2 t} w(x), \tag{10}$$

for all $x \in \mathbb{R}^m$, $t \geq 0$, and $v \in \mathbb{B}_w(\mathbb{R}^m)$.

In (10), the notation $\eta_\mu(v)$ has the same meaning as (9) with v instead of w .

Remark 4. Under Assumptions 1, 2, 3, and 4, the extended average payoff criterion (8) satisfies the following: For each $k = 1, 2$:

- (a) $J_k(x, \mu) = \int_{\mathbb{R}^m} r_k(y, \mu) \eta_\mu(dy)$, for all $x \in \mathbb{R}^m$, $\mu \in \Gamma$; actually the limit in (8) does exist in a strong sense (i.e., $\liminf = \limsup$) and does not depend on the initial condition x .
- (b) $\sup_{\mu \in \Gamma} |J_k(x, \mu)| \leq M \cdot d/c$, for all $x \in \mathbb{R}^m$, with M and d, c the constants appearing in our previous Assumptions 3 and 2, respectively.

For a proof of these two assertions, we can quote Section 3 in [9] or Section 2 in [14].

3.1 Continuity properties

In this part we will ensure that the functions $\mu \mapsto V_k(x, \mu)$ and $\mu \mapsto J_k(x, \mu)$ are continuous. To this end, we endow the set Γ with the topology of joint strategies (see e.g. [3, Lemma 3.4] or [6]).

Definition 8 (Topology of joint strategies). We say that a sequence $\{\mu_n : n = 1, 2, \dots\} \subset \Gamma$ converges to $\mu \in \Gamma$ (and we will denote such convergence as $\mu_n \xrightarrow{W} \mu$) if and only if for all $h \in \mathbb{C}_b(\mathbb{R}^m \times U_1 \times U_2)$ and $g \in \mathbb{L}^1(\mathbb{R}^m)$

$$\int_{\mathbb{R}^m} g(x) \int_{U_1 \times U_2} h(x, u_1, u_2) \mu_n(d(u_1, u_2)|x) dx \xrightarrow{n \rightarrow \infty} \int_{\mathbb{R}^m} g(x) \int_{U_1 \times U_2} h(x, u_1, u_2) \mu(d(u_1, u_2)|x) dx.$$

Remark 5. The space Γ is a convex compact metric space endowed with the previous topology; see [25, Theorem IV.3.11] or [6, Section 3]. Furthermore, as was mentioned in [16, Remark 2.11(b)], the set $\Pi_1 \times \Pi_2$ is compact too. The convexity of this last product set easily follows from the convexity of Π_k , $k = 1, 2$.

The following proposition gives a characterization for the α -discounted reward (6). For a proof we quote [15, Proposition 3.1.5].

Proposition 2. *Assume that the Assumptions 1, 2, 3 hold true. Then, for every $\mu \in \Gamma$, the associated total expected α -discounted function $V_k(\cdot, \mu)$ ($k = 1, 2$) is in $\mathbb{W}^{2,p}(\mathbb{R}^m) \cap \mathbb{B}_w(\mathbb{R}^m)$ and it satisfies the equation*

$$\alpha V_k(x, \mu) = r_k(x, \mu) + L^\mu V_k(x, \mu). \quad (11)$$

Conversely, if some function $\varphi_k \in \mathbb{W}^{2,p}(\mathbb{R}^n) \cap \mathbb{B}_w(\mathbb{R}^m)$ verifies (11), then

$$\varphi_k(x) = V_k(x, \mu) \quad \text{for all } x \in \mathbb{R}^m.$$

Moreover, if the equality in (11) is replaced by “ \leq ” or “ \geq ”, then (11) holds with the respective inequality.

The following result addresses a continuity property of the total expected α -discounted payoffs.

Proposition 3 (Continuity of V_k). *For $k = 1, 2$, the mapping $\mu \mapsto V_k(x, \mu)$ is continuous on Γ , for each $x \in \mathbb{R}^m$.*

Proof. Let $\{\mu_n\} \in \Gamma$ such that $\mu_n \xrightarrow{W} \mu$. Observe that Proposition 2 ensures that, for each $n \geq 1$, $V_k(x, \mu_n)$ satisfies the equation

$$\alpha V_k(x, \mu_n) = r_k(x, \mu_n) + L^{\mu_n} V_k(x, \mu_n) \quad x \in \mathbb{R}^m. \quad (12)$$

This last equation in terms of the operator $\mathcal{L}_\alpha^{\mu_n}$ given in (30) becomes

$$0 = \mathcal{L}_\alpha^{\mu_n} V_k(x, \mu_n) \quad x \in \mathbb{R}^m. \quad (13)$$

Next we will check that the hypotheses (a)-(e) of Theorem 2 provided in the appendix of this chapter are satisfied.

- (a) This hypothesis trivially follows from (13) (or by (12)).
- (b) To prove this hypotheses, let $R > 0$, and take the ball $B_R := \{x \in \mathbb{R}^m \mid |x| < R\}$. By [13, Theorem 9.11], there exists a constant C_0 independent of R such that, for a fixed $p > m$ (m being the dimension of (1)), we have

$$\begin{aligned} \|V_k(\cdot, \mu_n)\|_{\mathbb{W}^{2,p}(B_R)} &\leq C_0 (\|V_k(\cdot, \mu_n)\|_{\mathbb{L}^p(B_{2R})} + \|r_k(\cdot, \mu_n)\|_{\mathbb{L}^p(B_{2R})}) \\ &\leq C_0 (M(\alpha)\|w\|_{\mathbb{L}^p(B_{2R})} + M\|w\|_{\mathbb{L}^p(B_{2R})}) \\ &\leq C_0 (M(\alpha) + M) |\bar{B}_{2R}|^{1/p} \max_{x \in \bar{B}_{2R}} w(x) < \infty, \end{aligned}$$

where $|\bar{B}_{2R}|$ represents the volume of the closed ball with radius $2R$, and M and $M(\alpha)$ are the constants in Assumption 3(b) and in (7), respectively.

- (c)-(e) The parts (c) and (d) of Theorem 2 trivially hold by taking $\xi_n \equiv 0$ and $\alpha_n \equiv \alpha$, whereas that part (e) is part of our hypotheses.

Then, for $k = 1, 2$, we get the existence of a function $h_\mu^k \in \mathbb{W}^{2,p}(B_R)$ together with a subsequence $\{n_j\}$ such that $V_k(\cdot, \mu_{n_j}) \rightarrow h_\mu^k(\cdot)$ uniformly in B_R and pointwise on \mathbb{R}^m as $j \rightarrow \infty$ and $\mu_{n_j} \xrightarrow{W} \mu$. Furthermore, h_μ^k satisfies

$$\alpha h_\mu^k(x) = r_k(x, \mu) + L^\mu h_\mu^k(x), \quad x \in B_R.$$

Since the radius $R > 0$ was arbitrary, we can extend our analysis to all of $x \in \mathbb{R}^m$. Thus, Proposition 2 asserts that $h_\mu^k(x)$ actually coincides with $V_k(x, \mu)$. This proves the continuity of V_k . \square

We are going to focus on the continuity of the extended average payoff (8). To begin with, we shall use a characterization of this criterion, whose proof is identical to that in [14, Lemma 4.1] (see also [9, Proposition 5.1]).

Proposition 4 (Poisson equation). *For each $k = 1, 2$, and each fixed strategy $\mu \in \Gamma$, we denote by $g^k(\mu) := \int_{\mathbb{R}^m} r_k(y, \mu) \eta_\mu(dy)$. Then, under the Assumptions 1, 2, 3, and 4, there exists a function $\varphi_\mu^k \in \mathbb{W}^{2,p}(\mathbb{R}^n) \cap \mathbb{B}_w(\mathbb{R}^m)$, such that the pair $(g^k(\mu), \varphi_\mu^k)$ satisfies the so-named Poisson equation*

$$g^k(\mu) = r_k(x, \mu) + L^\mu \varphi_\mu^k(x), \quad k = 1, 2, \quad x \in \mathbb{R}^m, \quad (14)$$

as long with the transversality condition

$$\int_{\mathbb{R}^m} \varphi_\mu^k(x) \eta_\mu(dx) = 0. \quad (15)$$

Moreover, $g^k(\mu)$ is equal to the extended average payoff $J_k(x, \mu)$, for all $x \in \mathbb{R}^m$.

Now let us show the continuity of $g^k(\mu)$:

Proposition 5 (Continuity of g^k). *For $k = 1, 2$, the mapping $\mu \mapsto g^k(\mu)$ is continuous on Γ .*

Proof. The proof is similar to that given in Proposition 3. Indeed, take again a ball B_R for some $R > 0$ and use $\mu_n \in \Gamma$ such that $\mu_n \xrightarrow{W} \mu$. By Proposition 4, for each n the pair $(g^k(\mu_n), \varphi_{\mu_n}^k)$ satisfies the equation (14) with $\varphi_{\mu_n}^k \in \mathbb{W}^{2,p}(\mathbb{R}^n) \cap \mathbb{B}_w(\mathbb{R}^m)$. This equation in terms of operator $\mathcal{L}_\alpha^{\mu_n}$ in (30) becomes

$$g^k(\mu_n) = \mathcal{L}_0^{\mu_n} \varphi_{\mu_n}^k(x). \quad (16)$$

We will check that hypotheses (a)-(e) of Theorem 2 are satisfied. For this end, note by Assumption 3(b) and Proposition 4, that the functions $r_k(\cdot, \mu_n)$ and $\varphi_{\mu_n}^k$ are both in $\mathbb{B}_w(\mathbb{R}^m)$. Thus, using again the result in [13, Theorem 9.11], we can ensure the existence of some \bar{C}_0 (independent of R) such that

$$\begin{aligned} \|\varphi_{\mu_n}^k\|_{\mathbb{W}^{2,p}(B_R)} &\leq \bar{C}_0(\|\varphi_{\mu_n}^k\|_{\mathbb{L}^p(B_{2R})} + \|r_k(\cdot, \mu_n)\|_{\mathbb{L}^p(B_{2R})}) \\ &\leq \bar{C}_0(M_1\|w\|_{\mathbb{L}^p(B_{2R})} + M\|w\|_{\mathbb{L}^p(B_{2R})}) \\ &\leq \bar{C}_0(M_1 + M)|\bar{B}_{2R}|^{1/p} \max_{x \in \bar{B}_{2R}} w(x) < \infty, \end{aligned} \quad (17)$$

where $|\bar{B}_{2R}|$ is defined as in the proof of Proposition 3 and M_1 is some given constant. The hypotheses (a) and (b) follows from (16) and (17), respectively. As for

part (c), we take $\xi_n = g^k(\mu_n)$ and noting that $|g^k(\mu_n)| \leq Md/c$ (see Remark 4(b)), we get the existence of a constant g^k such that $g^k(\mu_n) \rightarrow g^k$ (under a suitable subsequence), hence part (c) of Theorem 2 trivially holds. Also, part (d) is satisfied by taking $\alpha_n \equiv 0$. Part (e) is part of our hypotheses. In this way, Theorem 2 ensures the existence of a function $\varphi_\mu^k \in \mathbb{W}^{2,p}(B_R)$ together with a subsequence $\{n_j\}$ such that $\varphi_{\mu_{n_j}}^k(\cdot) \rightarrow \varphi_\mu^k(\cdot)$ uniformly in B_R and pointwise on \mathbb{R}^m as $j \rightarrow \infty$ and $\mu_{n_j} \xrightarrow{W} \mu$. Moreover, φ_μ^k satisfies

$$g^k = r_k(x, \mu) + L^\mu \varphi_\mu^k(x) = 0, \quad x \in B_R. \tag{18}$$

Since the radius $R > 0$ was arbitrary, we can extend our analysis to all of $x \in \mathbb{R}^m$.

Finally, let $\bar{\varphi}_\mu^k(\cdot)$ be the bias function of μ , see [9, Definition 5.1]. By [9, Proposition 5.1], the pair $(g^k(\mu), \bar{\varphi}_\mu^k(\cdot))$ is the unique solution of the Poisson equation (14), i.e., $\varphi_\mu^k(\cdot) = \bar{\varphi}_\mu^k(\cdot) + c$ for some constant $c \in \mathbb{R}$, and $g^k = g^k(\mu)$. This implies that

$$g^k(\mu) = r_k(\cdot, \mu) + \mathcal{L}^\mu \bar{\varphi}_\mu^k(x) = 0 \quad x \in \mathbb{R}^m.$$

Furthermore, [9, Proposition 5.1] also ensures that the bias $\bar{\varphi}_\mu^k(\cdot)$ satisfies the transversality condition (15). Hence, a simple use of Proposition 4 provides us the continuity of the mapping $\mu \mapsto g^k(\mu)$ on Γ . □

4 Correlated equilibria

As mentioned in [20], a correlated strategy limits the freedom of the players in selecting their strategies, because a process of pre-play communication is needed to carry out a correlated strategy. However, any player is free to choose any strategy, regardless of the results of the communication process.

Suppose that a correlated strategy $\mu \in \Gamma$ is fixed by the players during a pre-play communication process. Then players make their final decisions independently of each other. As a consequence, we obtain the following cases.

1. Both players accept $\mu \in \Gamma$, then the system (1) evolves by applying the control strategy μ .
2. Both players do not accept $\mu \in \Gamma$, then the system (1) evolves according to some $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2 \subset \Gamma$.
3. Player 1 does not accept $\mu \in \Gamma$ and decides to use a stationary randomized strategy $\pi^1 \in \Pi^1$ instead, while player 2 approves the use of μ . Then, π^1 and μ are taken into account and the system (1) evolves with the control strategy $(\pi^1, \mu_2) \in \Pi_1 \times \Gamma_2 \subset \Gamma$, with μ_2 as the marginal distribution of μ on U_2 .
4. Player 2 does not accept $\mu \in \Gamma$ and decides to use a stationary randomized strategy $\pi^2 \in \Pi^2$ instead, while player 1 approves the use of μ . Then, π^2 and μ are taken into account and the system (1) evolves according to the pair $(\mu_1, \pi^2) \in \Gamma_1 \times \Pi_2$, with μ_1 as the marginal distribution of μ on U_1 .

The next definition extends the concept of a Nash equilibrium for the larger set of join strategies Γ .

Definition 9 (Correlated equilibria). Let F_k be either payoffs V_k or J_k , defined in (6) and (8), respectively. A correlated randomized strategy $\mu \in \Gamma$ is a correlated equilibrium for the extended payoff F_k if and only if

$$F_1(x, \mu) \geq F_1(x, \pi^1 \times \mu_2) \quad \forall \pi^1 \in \Pi_1,$$

$$F_2(x, \mu) \geq F_2(x, \mu_1 \times \pi^2) \quad \forall \pi^2 \in \Pi_2.$$

Existence of correlated equilibria always exists under our present hypotheses as it is established next:

Theorem 1. (a) Under Assumptions 1, 2, and 3, there exists a correlated equilibrium associated to the payoff V_k in (6).
 (b) If Assumption 4 is also considered, then the existence of a correlated equilibrium for the payoff J_k in (8) is also achieved.

To prove this theorem, we are going to describe some auxiliary results.

The auxiliary zero-sum game: Consider the set

$$\Theta = \{(\pi^1 \times \pi^2, \lambda_1, \lambda_2) : \lambda_1, \lambda_2 > 0, \lambda_1 + \lambda_2 = 1, \pi^1 \times \pi^2 \in \Pi_1 \times \Pi_2\}.$$

We assume that we have two virtual players, say players A and B , so that the set of correlated randomized strategies Γ is the set of strategies for player A , whereas that Θ is the set of strategies for player B . The common payoff for both players is given by

$$G_F(x, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2) :=$$

$$\lambda_1 \left[\int_{U_1 \times U_2} F_1(x, u_1, u_2) \mu(d(u_1, u_2)|x) - \int_{U_1 \times U_2} F_1(x, u_1, u_2) \pi^1(du_1|x) \mu_2(du_2|x) \right]$$

$$+ \lambda_2 \left[\int_{U_1 \times U_2} F_2(x, u_1, u_2) \mu(d(u_1, u_2)|x) - \int_{U_1 \times U_2} F_2(x, u_1, u_2) \mu_1(du_1|x) \pi^2(du_2|x) \right],$$

(19)

where F_k denotes either V_k or J_k , for $k = 1, 2$ and the subscript F of G simply refers the dependence of G with the F_k 's.

In virtue of the notation in (6) or (8), the payoff given in (19) can be rewritten as

$$G_F(x, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2) = \lambda_1 [F_1(x, \mu) - F_1(x, \pi^1 \times \mu_2)]$$

$$+ \lambda_2 [F_2(x, \mu) - F_2(x, \mu_1 \times \pi^2)].$$

(20)

Value of the game: In zero-sum games, the functions

$$U(x) := \inf_{(\pi^1 \times \pi^2, \lambda_1, \lambda_2) \in \Theta} \sup_{\mu \in \Gamma} G_F(x, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2) \quad \text{and}$$

$$L(x) := \sup_{\mu \in \Gamma} \inf_{(\pi^1 \times \pi^2, \lambda_1, \lambda_2) \in \Theta} G_F(x, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2),$$

play an important role. The function L is called the game's *lower value*, and U is the game's *upper value*. Clearly, we have $L \leq U$. If the upper and lower values coincide, then the game is said to have a *value*, and the *value of the game*, denoted as V , is the common value of L and U , i.e.,

$$V := L = U.$$

Definition 10. Let X be a nonempty Hausdorff space and let $g : X \mapsto \mathbb{R}$ be a real-valued function. We say that g is affine-like function if and only if, for every $x_1, x_2 \in X$ and $\beta \in [0, 1]$, there exists $x_\beta \in X$ such that $g(x_\beta) = \beta g(x_1) + (1 - \beta)g(x_2)$.

The following proposition shows some properties of the payoff functions G_V and G_J .

Proposition 6. (a) *Suppose that Assumptions 1, 2 and 3 hold true. Then, the mapping $\mu \mapsto G_V(\cdot, \mu, \cdot, \cdot, \cdot)$ is continuous and affine-like on Γ . Furthermore, the mapping $(\pi^1 \times \pi^2, \lambda_1, \lambda_2) \mapsto G_V(\cdot, \cdot, \pi^1 \times \pi^2, \lambda_1, \lambda_2)$ is affine-like on Θ .*

(b) *If in addition Assumption 4 is satisfied, then the same assertion in (a) is true for the payoff G_J .*

Proof. (a) First, let us prove the continuity: Consider the sequence $\{\mu_n\} \subset \Gamma$ such that $\mu_n \xrightarrow{W} \mu$. Observe that

$$\begin{aligned} 0 &\leq |G_V(x, \mu_n, \pi^1 \times \pi^2, \lambda_1, \lambda_2) - G_V(x, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2)| \\ &\leq \lambda_1 |V_1(x, \mu_n) - V_1(x, \mu)| + \lambda_2 |V_2(x, \mu_n) - V_2(x, \mu)| + \lambda_1 |V_1(x, \pi^1 \times \mu_{2n}) \\ &\quad - V_1(x, \pi^1 \times \mu_2)| + \lambda_2 |V_2(x, \mu_{1n} \times \pi^2) - V_2(x, \mu_1 \times \pi^2)|. \end{aligned} \tag{21}$$

Then, in virtue of Proposition 3, the terms in the right-hand side of (21) converge to zero as $\mu_n \xrightarrow{W} \mu$. So, G_V is continuous in Γ .

On the other hand, it is well-known that the discount payoff V_k can be seen as a linear mapping between r_k and the so-named occupation measure $v[x; \mu]$; i.e., $V_k(x, \mu) = \int r_k d\nu[x; \mu]$, for every $x \in \mathbb{R}^m$ and $\mu \in \Gamma$, where $v[x; \mu]$ is defined as

$$\int r_k v[x; \mu] = \alpha \mathbb{E}_x^\mu \left[\int_0^\infty e^{-\alpha t} \int_{U_1 \times U_2} r_k(x(t), u_1, u_2) \mu(d(u_2, u_2)|x(t)) dt \right]. \tag{22}$$

The details of this last fact can be extracted from page 1191 in [11] or from page 102 in [7]. Then by rewritting the payoff function V_k in the way of (22), it can be proved (see, for instance [11], page 1195) that, for any two strategies $\mu, \bar{\mu} \in \Gamma$ and $\beta \in [0, 1]$, there exists another $\mu^\beta \in \Gamma$ so that

$$v[x; \mu^\beta] = \beta v[x; \mu] + (1 - \beta)v[x; \bar{\mu}]. \tag{23}$$

This last property together with (22) yield that $V_k(x, \mu^\beta) = \beta V_k(x, \mu) + (1 - \beta)V_k(x, \bar{\mu})$ and the choice of μ^β is independent of $k = 1, 2$. With the previous ingredients, let us use the strategy $\mu^\beta \in \Gamma$ obtained by the affine-like property for both criteria V_1 and V_2 , for some arbitrary choose of two strategies $\mu, \bar{\mu} \in \Gamma$ and $\beta \in [0, 1]$. Then, the following is satisfied

$$\begin{aligned}
G_V(x, \mu^\beta, \pi^1 \times \pi^2, \lambda_1, \lambda_2) &= \lambda_1 [V_1(x, \mu^\beta) - V_1(x, \pi^1 \times \mu_2^\beta)] + \\
&\quad \lambda_2 [V_2(x, \mu^\beta) - V_2(x, \mu_1^\beta \times \pi^2)] \\
&= \lambda_1 [\{\beta V_1(x, \mu) + (1 - \beta)V_1(x, \bar{\mu})\} - V_1(x, \pi^1 \times \mu_2^\beta)] + \\
&\quad + \lambda_2 [\{\beta V_2(x, \mu) + (1 - \beta)V_2(x, \bar{\mu})\} - V_2(x, \mu_1^\beta \times \pi^2)].
\end{aligned} \tag{24}$$

In addition, by following the same arguments of page 1195 in [11], it can be also verified that

$$\begin{aligned}
V_1(x, \pi^1 \times \mu_2^\beta) &= \beta V_1(x, \pi^1 \times \mu_2) + (1 - \beta)V_1(x, \pi^1 \times \bar{\mu}_2) \quad \text{and} \\
V_2(x, \mu_1^\beta \times \pi^2) &= \beta V_2(x, \mu_1 \times \pi^2) + (1 - \beta)V_2(x, \bar{\mu}_1 \times \pi^2).
\end{aligned} \tag{25}$$

Combining (25) with (24) we deduce

$$\begin{aligned}
G_V(x, \mu^\beta, \pi^1 \times \pi^2, \lambda_1, \lambda_2) &= \beta G_V(x, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2) \\
&\quad + (1 - \beta)G_V(x, \bar{\mu}, \pi^1 \times \pi^2, \lambda_1, \lambda_2),
\end{aligned}$$

for all $(\pi^1 \times \pi^2, \lambda_1, \lambda_2) \in \Theta$. This proves the affine-like property of G_V on Γ .

On the other hand, for any $\beta \in [0, 1]$ and any $(\pi^1 \times \pi^2, \lambda_1, \lambda_2)$, $(\bar{\pi}^1 \times \bar{\pi}^2, \bar{\lambda}_1, \bar{\lambda}_2) \in \Theta$ consider the following strategies $\pi_\beta^k \in \Pi_k$ and constants $\lambda_k^\beta \in \mathbb{R}$ ($k = 1, 2$):

$$\pi_\beta^k := \frac{\beta \lambda_k \pi^k + (1 - \beta) \bar{\lambda}_k \bar{\pi}_k}{\beta \lambda_k + (1 - \beta) \bar{\lambda}_k}, \quad \text{and} \quad \lambda_k^\beta := \beta \lambda_k + (1 - \beta) \bar{\lambda}_k.$$

Plugging these elements into G_V , it is easy to check that for each $\mu \in \Gamma$ and $x \in \mathbb{R}^m$,

$$\begin{aligned}
G_V(x, \mu, \pi_\beta^1 \times \pi_\beta^2, \lambda_1^\beta, \lambda_2^\beta) &= \beta G_V(x, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2) + \\
&\quad (1 - \beta)G_V(x, \mu, \bar{\pi}^1 \times \bar{\pi}^2, \bar{\lambda}_1, \bar{\lambda}_2).
\end{aligned}$$

This proves that G_V is affine-like on Θ .

(b) As for the continuity of G_J on Γ , the proof is the same as in part (a), the only difference lies in replacing V_k by J_k and just use Proposition 5 in lieu of Proposition 3. Furthermore, there are works asserting that the average payoff J_k ($k = 1, 2$) can be rewritten in terms of an occupation measure $\rho[\mu]$; i.e., for all $x \in \mathbb{R}^m$ and each $\mu \in \Gamma$, $J_k(x, \mu) = \int r_k d\rho[\mu]$, where $\rho[\mu](dy, du) := \eta_\mu(dy)\mu(du|y)$, with η_μ being the invariant measure defined in (9) (for further details see for instance, [2], page 87 or [7], page 91). Using the arguments as in page 92 of [7], we can obtain exactly the same property as (23) for ρ rather than v . Then, it is straightforward that the mapping $\mu \mapsto J_k(\cdot, \mu)$ is affine-like on Γ . To prove the affine-like property of G_J , we proceed in the same way as (24). We can use also the same procedures of page 92 of [7] to get a similar relation of (25) associated to J_k . These previous properties would prove that G_J is affine-like on Γ after doing basic estimates. The proof that G_J is affine-like on Θ is similar to the one presented for G_V so we shall omit it. \square

Proposition 7. *The upper value U is nonnegative.*

Proof. Clearly we know that

$$\begin{aligned} \sup_{\mu \in \Gamma} G_F(\cdot, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2) &\geq G_F(\cdot, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2) \\ &\forall \mu \in \Gamma, (\pi^1 \times \pi^2, \lambda_1, \lambda_2) \in \Theta. \end{aligned}$$

Then taking in particular $\hat{\mu} := \pi^1 \times \pi^2$, we obtain that $G_F(\cdot, \hat{\mu}, \pi^1 \times \pi^2, \lambda_1, \lambda_2) = 0$. Therefore,

$$\begin{aligned} \sup_{\mu \in \Gamma} G_F(\cdot, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2) &\geq G_F(\cdot, \hat{\mu}, \pi^1 \times \pi^2, \lambda_1, \lambda_2) = 0, \\ &\forall (\pi^1 \times \pi^2, \lambda_1, \lambda_2) \in \Theta. \end{aligned}$$

This implies that

$$U(\cdot) := \inf_{(\pi^1 \times \pi^2, \lambda_1, \lambda_2) \in \Theta} \sup_{\mu \in \Gamma} G_F(\cdot, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2) \geq 0. \quad (26)$$

□

Proof of Theorem 1. (a) First note that Proposition 6 gives the hypotheses to get the Isaac's condition (see, for instance pages 108-109 in [20])

$$\begin{aligned} &\inf_{(\pi^1 \times \pi^2, \lambda_1, \lambda_2) \in \Theta} \sup_{\mu \in \Gamma} G_V(x, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2) = \\ &= \sup_{\mu \in \Gamma} \inf_{(\pi^1 \times \pi^2, \lambda_1, \lambda_2) \in \Theta} G_V(x, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2), \quad x \in \mathbb{R}^m. \end{aligned} \quad (27)$$

Relations (27) and (26) gives us that

$$\sup_{\mu \in \Gamma} \inf_{(\pi^1 \times \pi^2, \lambda_1, \lambda_2) \in \Theta} G_V(x, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2) \geq 0 \quad \forall x \in \mathbb{R}^m.$$

As $\mu \mapsto G_V(\cdot, \mu, \cdot, \cdot, \cdot)$ is continuous, then it easy to verify that

$$\mu \mapsto \inf_{(\pi^1 \times \pi^2, \lambda_1, \lambda_2) \in \Theta} G_V(\cdot, \mu, \pi^1 \times \pi^2, \lambda_1, \lambda_2)$$

is upper semi-continuous. This last property together with the compactness of Γ imply the existence of $\mu^* \in \Gamma$ (that depends only of $x \in \mathbb{R}^m$) such that

$$G_V(x, \mu^*, \pi^1 \times \pi^2, \lambda_1, \lambda_2) \geq 0, \quad \forall (\pi^1 \times \pi^2, \lambda_1, \lambda_2) \in \Theta, x \in \mathbb{R}^m. \quad (28)$$

In virtue of (28), if we let $\lambda_2 \rightarrow 0$ in (19) or (20) (yielding that $\lambda_1 \rightarrow 1$), we get

$$V_1(x, \mu^*) - V_1(x, \pi^1 \times \mu_2^*) \geq 0, \quad \text{for all } \pi^1 \in \Pi^1.$$

Similarly, by letting $\lambda_1 \rightarrow 0$ (yielding that $\lambda_2 \rightarrow 1$), we can also deduce

$$V_2(x, \mu^*) - V_2(x, \mu_1^* \times \pi^2) \geq 0 \text{ for all } \pi^2 \in \Pi^2.$$

Thus, from the Definition 9, $\mu^* \in \Gamma$ becomes a correlated equilibrium.

The proof of part (b) is identical than (a), the only difference lies in the fact that we need Assumption 4 as an extra hypothesis to guarantee the continuity for G_J . \square

Appendix

The main objective of this appendix is to prove that the convergence $\mu_n \xrightarrow{W} \mu$, $\alpha_n \rightarrow \alpha$, and $h_n \rightarrow h$ (this later convergence in a suitable sense), yield that, for each $k = 1, 2$,

$$\lim_{n \rightarrow \infty} \{r_k(\cdot, \mu_n) + L^{\mu_n} h_n - \alpha_n h_n\} = r_k(\cdot, \mu) + L^\mu h - \alpha h. \quad (29)$$

Let \mathcal{O} be an open, bounded and connected subset of \mathbb{R}^m . We denote the closure of this set by $\bar{\mathcal{O}}$.

For every $x \in \mathbb{R}^m$, $\mu \in \Gamma$, $\alpha > 0$, h in $\mathbb{W}^{2,p}(\mathcal{O})$, we define

$$\begin{aligned} \hat{\Psi}(x, \mu, \alpha; h) &:= r_k(x, \mu) + \sum_{i=1}^n b_i(x, \mu) \frac{\partial h}{\partial x_i}(x) - \alpha h(x), \\ \mathcal{L}_\alpha^\mu h(x) &:= \hat{\Psi}(x, \mu, \alpha; h) + \frac{1}{2} \sum_{i,j=1}^m a^{ij}(x) \frac{\partial^2 h}{\partial x_i \partial x_j}(x), \end{aligned} \quad (30)$$

where b_i is the i -th component of the function b defined in (1) and a as in Assumption 1(d).

The following theorem establishes the limit result referred in (29).

Theorem 2. *Let \mathcal{O} be a bounded \mathcal{C}^2 domain. Suppose that there exist sequences $\{h_n\} \in \mathbb{W}^{2,p}(\mathcal{O})$, $\{\xi_n\} \in \mathbb{L}^p(\mathcal{O})$, with $p > m$ (m is the dimension of (1)), $\{\mu_n\} \in \Gamma$, and $\{\alpha_n\} \geq 0$, satisfying the following:*

- (a) $\mathcal{L}_{\alpha_n}^{\mu_n} h_n = \xi_n$ in \mathcal{O} for $n = 1, 2, \dots$
- (b) There exists a constant \tilde{M}_1 such that $\|h_n\|_{\mathbb{W}^{2,p}(\mathcal{O})} \leq \tilde{M}_1$ for $n = 1, 2, \dots$
- (c) ξ_n converges in $\mathbb{L}^p(\mathcal{O})$ to some function ξ .
- (d) α_n converges to some constant $\alpha \geq 0$.
- (e) $\mu_n \xrightarrow{W} \mu \in \Gamma$.

Then, there exist a function $h \in \mathbb{W}^{2,p}(\mathcal{O})$ and a subsequence $\{n_r\} \subset \{1, 2, \dots\}$ such that $h_{n_r} \rightarrow h$ in the norm of $\mathbb{C}^{1,\eta}(\bar{\mathcal{O}})$ for $\eta < 1 - \frac{m}{p}$ as $r \rightarrow \infty$. Moreover,

$$\mathcal{L}_\alpha^\mu h = \xi \text{ in } \mathcal{O}. \quad (31)$$

Proof. We first show that there exist a function h in $\mathbb{W}^{2,p}(\mathcal{O})$ and a subsequence $\{n_r\} \subset \{1, 2, \dots\}$ such that, as $r \rightarrow \infty$, $h_{n_r} \rightarrow h$ weakly in $\mathbb{W}^{2,p}(\mathcal{O})$ and strongly in

$\mathbb{C}^{1,\eta}(\bar{\mathcal{O}})$. Namely, since $\mathbb{W}^{2,p}(\mathcal{O})$ is reflexive (see [1, Theorem 3.5]), then, using Theorem 1.17 in the same reference [1], the ball

$$H := \left\{ h \in \mathbb{W}^{2,p}(\mathcal{O}) : \|h\|_{\mathbb{W}^{2,p}(\mathcal{O})} \leq \bar{M} \right\} \tag{32}$$

is weakly sequentially compact. On the other hand, since $p > m$, by [1, Theorem 6.2, Part III], the imbedding $\mathbb{W}^{2,p}(\mathcal{O}) \hookrightarrow \mathbb{C}^{1,\eta}(\bar{\mathcal{O}})$, for $0 \leq \eta < 1 - \frac{m}{p}$ is compact; hence, it is also continuous, and thus the set H in (32) is relatively compact in $\mathbb{C}^{1,\eta}(\bar{\mathcal{O}})$. This fact ensures the existence of a function $h \in \mathbb{W}^{2,p}(\mathcal{O})$ and a subsequence $\{h_{n_r}\} \equiv \{h_n\} \subset H$ such that

$$h_{n_r} \rightarrow h \text{ weakly in } \mathbb{W}^{2,p}(\mathcal{O}) \text{ and strongly in } \mathbb{C}^{1,\eta}(\bar{\mathcal{O}}). \tag{33}$$

The second step is to show that, as $n \rightarrow \infty$,

$$\int_{\mathcal{O}} g(x) \hat{\Psi}(x, \mu_n, \alpha_n; h_n) dx \rightarrow \int_{\mathcal{O}} g(x) \hat{\Psi}(x, \mu, \alpha; h) dx \quad \text{for all } g \in \mathbb{L}^1(\mathcal{O}). \tag{34}$$

To this end, given $x \in \mathcal{O}$, $k = 1, 2$, functions $h \in \mathbb{W}^{2,p}(\mathcal{O})$ and $h_n \in H$, $\mu, \mu_n \in \Gamma$, and constants $\alpha_n, \alpha \geq 0$, the following holds for all $g \in \mathbb{L}^1(\mathcal{O})$.

$$\begin{aligned} & \left| \int_{\mathcal{O}} g(x) \hat{\Psi}(x, \mu_n, \alpha_n; h_n) dx - \int_{\mathcal{O}} g(x) \hat{\Psi}(x, \mu, \alpha; h) dx \right| \\ & \leq \left| \int_{\mathcal{O}} g(x) [r_k(x, \mu_n) - r_k(x, \mu)] dx \right| \\ & + \sum_{i=1}^m \left| \int_{\mathcal{O}} g(x) \left[b_i(x, \mu_n) \frac{\partial h_n}{\partial x_i}(x) - b_i(x, \mu) \frac{\partial h}{\partial x_i}(x) \right] dx \right| \\ & + \left| \int_{\mathcal{O}} g(x) [\alpha_n h_n(x) - \alpha h(x)] dx \right| \\ & \leq \left| \int_{\mathcal{O}} g(x) r_k(x, \mu_n) dx - \int_{\mathcal{O}} g(x) r_k(x, \mu) dx \right| \\ & + \sum_{i=1}^m \left| \int_{\mathcal{O}} g(x) \frac{\partial h_n}{\partial x_i}(x) [b_i(x, \mu_n) - b_i(x, \mu)] dx \right| \\ & + \sum_{i=1}^m \left| \int_{\mathcal{O}} g(x) b_i(x, \mu_n) \left[\frac{\partial h_n}{\partial x_i}(x) - \frac{\partial h}{\partial x_i}(x) \right] dx \right| + |\alpha_n - \alpha| \left| \int_{\mathcal{O}} g(x) h_n(x) dx \right| \\ & + \alpha \left| \int_{\mathcal{O}} g(x) [h_n(x) - h(x)] dx \right|. \end{aligned}$$

Since the embedding $\mathbb{W}^{2,p}(\mathcal{O}) \hookrightarrow \mathbb{C}^{1,\eta}(\bar{\mathcal{O}})$ is continuous, hypothesis (b) together with the definition of the norm $\|\cdot\|_{\mathbb{C}^{1,\eta}(\bar{\mathcal{O}})}$, imply that there is a constant $\bar{M} > 0$ such that

$$\max \left\{ |h_n|, \max_{1 \leq i \leq m} \left| \frac{\partial h_n}{\partial x_i} \right| \right\} \leq \|h_n\|_{\mathbb{C}^{1,\eta}(\bar{\mathcal{O}})} \leq \bar{M} \|h_n\|_{\mathbb{W}^{2,p}(\mathcal{O})} \leq \bar{M} \bar{M}_1.$$

On the other hand, it is easy to verify that Assumptions 1 and 3, yield that $|b(\cdot, \mu)| + |r_k(\cdot, \mu)| \leq K(\bar{\mathcal{O}})$. Hence,

$$\begin{aligned} & \left| \int_{\mathcal{O}} g(x) \hat{\Psi}(x, \mu_n, \alpha_n; h_n) dx - \int_{\mathcal{O}} g(x) \hat{\Psi}(x, \mu, \alpha; h) dx \right| \leq \\ & \left| \int_{\mathcal{O}} g(x) r_k(x, \mu_n) dx - \int_{\mathcal{O}} g(x) r_k(x, \mu) dx \right| \\ & + \bar{M} \bar{M}_1 m \max_{1 \leq i \leq m} \left| \int_{\mathcal{O}} g(x) [b_i(x, \mu_n) - b_i(x, \mu)] dx \right| \\ & + \|g\|_{\mathbb{L}^1(\mathcal{O})} \|h_n - h\|_{\mathbb{C}^{1,\eta}(\bar{\mathcal{O}})} (mK(\bar{\mathcal{O}}) + \alpha) + |\alpha_n - \alpha| \bar{M} \bar{M}_1 \|g\|_{\mathbb{L}^1(\mathcal{O})}. \end{aligned} \tag{35}$$

Observe that $r_k(\cdot, \mu)$ $k = 1, 2$, and $b_i(\cdot, \mu)$ $i = 1, \dots, m$ are bounded on $\bar{\mathcal{O}}$. Then, hypotheses (d) to (e), together with (33), lead to the right hand side of (35) goes to zero as $n \rightarrow \infty$, thus proving (34).

The existence of the constant $K(\bar{\mathcal{O}})$ used for the analysis in (35) can be also used to get also that $|\sigma(x)| \leq K(\bar{\mathcal{O}})$, then we can affirm that for each g in $\mathbb{L}^{\frac{p}{p-1}}(\mathcal{O})$,

$$\begin{aligned} & \frac{1}{2} \left| \int_{\mathcal{O}} g(x) \left[\sum_{i,j=1}^m a^{ij}(x) \frac{\partial^2 h_n}{\partial x_i \partial x_j}(x) - \sum_{i,j=1}^m a^{ij}(x) \frac{\partial^2 h}{\partial x_i \partial x_j}(x) \right] dx \right| \\ & \leq \frac{m^2}{2} [K(\bar{\mathcal{O}})]^2 \sum_{i,j=1}^m \left| \int_{\mathcal{O}} g(x) \left[\frac{\partial^2 h_n}{\partial x_i \partial x_j}(x) - \frac{\partial^2 h}{\partial x_i \partial x_j}(x) \right] dx \right|. \end{aligned} \tag{36}$$

Thus the weak convergence of $\{h_n\}$ to h in $\mathbb{W}^{2,p}(\mathcal{O})$ yields that the right-hand side of (36) converges to zero as $n \rightarrow \infty$. Notice also that the convergence of (34) is also valid for all $g \in \mathbb{L}^{\frac{p}{p-1}}(\mathcal{O})$. The reason is because $\mathbb{L}^{\frac{p}{p-1}}(\mathcal{O}) \subset \mathbb{L}^1(\mathcal{O})$ (recall the Lebesgue measure on \mathcal{O} is bounded). This last fact together with (36) and hypothesis (c), yield that for every g in $\mathbb{L}^{\frac{p}{p-1}}(\mathcal{O})$,

$$\int_{\mathcal{O}} g(x) [\mathcal{L}_\alpha^\mu h(x) - \xi(x)] dx = \lim_{n \rightarrow \infty} \int_{\mathcal{O}} g(x) [\mathcal{L}_{\alpha_n}^{\mu_n}(x) - \xi_n(x)] dx = 0.$$

The above limit, along with Theorem 2.10 in [18], implies (31), i.e.

$$\mathcal{L}_\alpha^\mu h = \xi \text{ in } \mathcal{O}.$$

This completes the proof. □

References

1. Adams, R.A.: *Sobolev Spaces*. Academic Press. New York, (1975).
2. Arapostathis, A.; Ghosh, M.K.; Borkar, V.S.: *Ergodic control of diffusion processes*. Encyclopedia of Mathematics and its Applications, 143. Cambridge University Press, (2012).
3. Arapostathis, A.; Borkar, V. S. : Uniform recurrence properties of controlled diffusions and applications to optimal control. *SIAM J. Control Optim.* 48, 4181-4223 (2010).

4. Aumann, R. J.: *Subjectivity and correlation in randomized games*. *Econometrica* 30, 445-462, (1974).
5. Aumann, R. J.: *Correlated equilibrium as an expression of Bayesian rationality*. *Econometrica* 55, 118, (1987).
6. Borkar, V.S.: *A topology for Markov controls*. *Appl. Math. Optim.* 20, 55-62, (1989).
7. Borkar, V. S.; Ghosh, M. K.: *Controlled diffusion processes with constraints*. *J. Math. Anal. Appl.*, 152, 88-108. (1990)
8. Borkar, V. S.; Ghosh, M. K. (1992): *Stochastic differential games: occupation measure based approach*. *J. Optim. Theory Appl.*, 73, 359-385. Correction: 88, 251-252, (1996).
9. Escobedo-Trujillo, B.A.; López-Barrientos, J.D.; Hernández-Lerma, O.: *Bias and overtaking equilibria for zero-sum stochastic differential games*. *J. Optim. Theory Appl.* 153, 662-687, (2012).
10. Forges, F.: *Five legitimate definitions of correlated equilibrium in games with incomplete information*. *Theory and Decision* 35, 277-310, (1993).
11. Ghosh, M.K.; Arapostathis, A.; Marcus, S.I.: *Optimal control of switching diffusions with applications to flexible manufacturing systems*. *SIAM J. Control Optim.* 30, 1-23, (1992).
12. Ghosh, M.K.; Arapostathis, A.; Marcus, S.I.: *Ergodic control of switching diffusions*. *SIAM J. Control Optim.* 35, 1962-1988, (1997).
13. Gilbarg, D.; Trudinger, N.S.: *Elliptic partial differential equations of second order*. Reprinted version. Heidelberg, Springer, (1998).
14. Jasso-Fuentes, H.; Hernández-Lerma, O.: *Characterizations of overtaking optimality for controlled diffusion processes*. *Appl. Math. Optim.* 57, 349-369, (2008).
15. Jasso-Fuentes, H.; Yin, G.G.: *Advanced criteria for controlled Markov-modulated diffusions in an infinite horizon: overtaking, bias and Blackwell optimality*. Science Press, Beijing China, (2013).
16. Jasso-Fuentes, H.; López-Barrientos D.; Escobedo-Trujillo B.A.: *Infinite horizon nonzero-sum stochastic differential games with additive structure*. *IMA J. Math. Control Inform.* 34, 283-309, (2017).
17. Küenle, H. U.: *Equilibrium strategies in stochastic games with additive cost and transition structure*. *Int. Game Theory Rev.*, 1, 131-147, (1999).
18. Lieb, E.H.; Loss, M.: *Analysis*. Second Edition. AMS. Providence, Rhode Island, (2001)..
19. Neyman, A.: *Correlated Equilibrium and potential games*. *International Journal of Game Theory*. 26, 223-227, (1997).
20. Nowak, A. S.: *Correlated relaxed equilibria in nonzero-sum linear differential games*. *J. Math. Anal. Appl.* 163, 104-112, (1992).
21. Nowak, A. S.: *Correlated equilibria in nonzero-sum differential games*. *J. Math. Anal. Appl.* 174, 539-549, (1993) (1994).
22. Solan, E.: *Characterization of correlated equilibria in stochastic games*. *Int. J. Game Theory* 30, 259-277, (2001).
23. Soltz, G.; Lugosi G.: *Learning correlated equilibria in games with compact sets of strategies*. *Games and Economic Behavior* 59, 187-208, (2007).
24. Stein, N.D.; Parrilo P. A.; Ozdaglar A.: *Correlated equilibria in continuous games: Characterization and computation*. *Games and Economic Behavior* 71, 436-455, (2011).
25. Warga, J.: *Optimal control of differential and functional equations*. Academic Press, New York, (1972).



Lattice Dynamical Systems in the Biological Sciences

Xiaoying Han and Peter E. Kloeden

Abstract This chapter focuses on dynamical behavior of lattice models arising in the biological sciences, in particular, attractors for such systems. Three types of lattice dynamical systems are investigated; they are lattice reaction-diffusion systems, Hopfield neural lattice systems, and neural field lattice systems. For each system the existence of a global, nonautonomous, or random attractor is shown. The upper semi continuity of attractors for the Hopfield neural lattice model and the upper semi continuity of numerical attractors are also discussed.

1 Introduction

A lattice dynamical system corresponding to a reaction-diffusion equation on the one-dimensional domain \mathbb{R}

$$\frac{\partial u(x,t)}{\partial t} = v \frac{\partial^2 u(x,t)}{\partial x^2} - \lambda u + f(u(x,t)) + g(x), \quad \text{with } \lambda, v > 0,$$

is obtained by using a finite difference quotient to discretize the Laplacian operator. More precisely, applying a spatial scaling $u_i(t) = u(i\Delta x, t)$ and setting the scaled step size Δx to 1 leads to the infinite dimensional system of ordinary differential equations called a *lattice dynamical system* (LDS),

$$\frac{du_i}{dt} = v(u_{i-1} - 2u_i + u_{i+1}) - \lambda u_i + f(u_i) + g_i, \quad i \in \mathbb{Z}, \quad (1)$$

Xiaoying Han

Department of Mathematics and Statistics, 221 Parker Hall, Auburn University, Auburn, AL 36849, USA, e-mail: xzh0003@auburn.edu

Peter E. Kloeden

Department of Mathematics and Statistics, 221 Parker Hall, Auburn University, Auburn, AL 36849, USA, e-mail: kloeden@math.uni-frankfurt.de

where u_i and g_i correspond to $u(i, t)$ and $g(i)$, respectively, for each $i \in \mathbb{Z}$.

Not all LDS originate from discretizing an underlying partial differential equation (PDE) as above. In fact, they exist naturally in applications where the spatial structure has a discrete character, such as cellular neural networks [22, 23, 24], chemical reaction theory [28, 48, 55], and living cell systems [10, 11, 50, 51, 61, 62]. Moreover, LDS can be interpreted as an infinite-dimensional ordinary differential equation, functional differential equation, or an evolution equation on sequence spaces and hence allow clear insight into the dynamics of the models even when their corresponding PDEs are analytically intractable. Such advantages make studies on LDS attractive and important.

Sequence spaces

One typical technique to study LDS is to first reformulate them as ordinary differential equations on an appropriate sequence space. The most widely used sequence space is the Hilbert space of real-valued square summable bi-infinite sequences ℓ^2 with norm and inner product

$$\|\mathbf{u}\| := \left(\sum_{i \in \mathbb{Z}} u_i^2 \right)^{1/2}, \quad (\mathbf{u}, \mathbf{v}) := \sum_{i \in \mathbb{Z}} u_i v_i \quad \text{for } \mathbf{u} = (u_i)_{i \in \mathbb{Z}}, \mathbf{v} = (v_i)_{i \in \mathbb{Z}} \in \ell^2.$$

Similarly, ℓ^∞ is the Banach space of real-valued bounded bi-infinite sequences with norm $\|\mathbf{u}\|_\infty := \sup_{i \in \mathbb{Z}} |u_i|$.

Since $u_i \rightarrow 0$ as $i \rightarrow \pm\infty$ for $\mathbf{u} = (u_i)_{i \in \mathbb{Z}} \in \ell^2$, the Hilbert space ℓ^2 does not include traveling wave solutions or solutions with just bounded components. This excludes a large number of applications with non-vanishing values at distant. Weighted sequence spaces were introduced to handle such dynamical behaviour. For greater applicability these will be defined for weighted space of bi-infinite real-valued sequences with vectorial indices $\mathbf{i} = (i_1, \dots, i_d) \in \mathbb{Z}^d$.

In particular, given a bounded positive sequence of weights $(\rho_{\mathbf{i}})_{\mathbf{i} \in \mathbb{Z}^d}$, define the linear space

$$\ell_\rho^p := \left\{ \mathbf{u} = (u_{\mathbf{i}})_{\mathbf{i} \in \mathbb{Z}^d} : \sum_{\mathbf{i} \in \mathbb{Z}^d} \rho_{\mathbf{i}} u_{\mathbf{i}}^p < \infty, \quad u_{\mathbf{i}} \in \mathbb{R} \right\}$$

with the norm

$$\|\mathbf{u}\|_{\rho, p} := \left(\sum_{\mathbf{i} \in \mathbb{Z}^d} \rho_{\mathbf{i}} u_{\mathbf{i}}^p \right)^{1/p} \quad \text{for } \mathbf{u} = (u_{\mathbf{i}})_{\mathbf{i} \in \mathbb{Z}^d} \in \ell_\rho^p.$$

Then ℓ_ρ^p is a separable Banach space [41]. In particular, when $p = 2$, ℓ_ρ^2 is a separable Hilbert space with the inner product and norm

$$\langle \mathbf{u}, \mathbf{v} \rangle := \sum_{\mathbf{i} \in \mathbb{Z}^d} \rho_{\mathbf{i}} u_{\mathbf{i}} v_{\mathbf{i}}, \quad \|\mathbf{u}\|_\rho := \sqrt{\sum_{\mathbf{i} \in \mathbb{Z}^d} \rho_{\mathbf{i}} u_{\mathbf{i}}^2} \quad \text{for } \mathbf{u} = (u_{\mathbf{i}})_{\mathbf{i} \in \mathbb{Z}^d} \in \ell_\rho^2, \mathbf{v} = (v_{\mathbf{i}})_{\mathbf{i} \in \mathbb{Z}^d} \in \ell_\rho^2.$$

The weights ρ_i are often assumed to satisfy

Assumption 1 $\rho_i > 0$ for all $i \in \mathbb{Z}^d$ and $\rho_\Sigma := \sum_{i \in \mathbb{Z}^d} \rho_i < \infty$.

It is straightforward to show that ℓ^2_ρ contains bi-infinite sequences with just bounded components and that $\ell^2 \subseteq \ell^\infty \subseteq \ell^2_\rho$.

A brief literature review

Extensive studies have been done regarding various aspects of solutions of LDS, that can be mainly classified into three categories, traveling wave solutions (see e.g., [3, 7, 11, 12, 19, 20, 30, 32, 45, 56, 57, 59, 73]), chaotic properties of solutions (see e.g., [21, 27, 63]), and long term behavior of solutions (see, e.g., [1, 9, 49, 65, 70, 69, 71, 72]). More recently, nonautonomous and stochastic LDS have been studied by Abdallah [2], Bates *et al.* [8], Caraballo & Lu [14], Caraballo *et al.*, [15, 16], Fan & Wang [29], Han [33] and references therein, Huang [46], Wang [66], Zhao & Zhou [68], amongst others.

Most studies on LDS in the literature consider the linear diffusion operator $u_{i+1} - 2u_i + u_{i-1}$ as in (1). In the biological context such an operator assumes the simplest tri-diagonal interconnection structure that allows only uniform linear diffusion among cells within the nearest neighborhood, i.e., each cell interacts only with the cells which are adjacent to it in a uniform manner. These assumptions exclude numerous applications with different interconnection structures at different components (see, e.g., [18, 22, 24, 23, 60]), among which neural networks are a very important example.

The main focus of this chapter is the existence of attractors for LDS arising from biological sciences. Each of the systems considered in this chapter has a different character from those considered in the literature. It is well-known that the existence of attractors usually relies on the existence of closed absorbing sets and asymptotic compactness of the underlying dynamical system. For the reader’s convenience below we recall the definition of attractors and state the well-known results on existence of attractors, pullback attractors, and random attractors, along with the definition of asymptotic compactness. The reader is referred to [5, 13, 53] for basic concepts and theory of nonautonomous and random dynamical systems.

In what follows, let \mathfrak{X} be a complete metric space and let dist denote the Hausdorff semi-distance of \mathfrak{X} given by $\text{dist}(A, B) = \sup_{a \in A} \inf_{b \in B} |a - b|_{\mathfrak{X}}$ for $A, B \subset \mathfrak{X}$.

Definition 1. A nonempty subset \mathcal{A} is called a global attractor for a semigroup (dynamical system) of continuous operator $\{\mathcal{S}(t)\}_{t \geq 0}$ on \mathfrak{X} if

- (i) \mathcal{A} is compact;
- (ii) \mathcal{A} is invariant under \mathcal{S} , i.e., $\mathcal{S}(t)\mathcal{A} = \mathcal{A}$ for each $t \geq 0$;
- (iii) \mathcal{A} attracts every bounded sets of \mathfrak{X} , i.e., $\lim_{t \rightarrow \infty} \text{dist}(\mathcal{S}(t)X, \mathcal{A}) = 0$ for any bounded set $X \in \mathfrak{X}$.

Definition 2. A family of sets $\mathcal{A} = \{A(t)\}_{t \in \mathbb{R}}$ is called a pullback attractor for a continuous two-parameter semigroup (nonautonomous dynamical system) $\{\varphi(t, t_0)\}_{t \geq t_0}$ on \mathfrak{X} if

- (i) $A(t)$ is compact for all $t \in \mathbb{R}$;
- (ii) \mathcal{A} is invariant under φ , i.e., $\varphi(t, t_0)A(t_0) = A(t)$ for all $t \geq t_0$;
- (iii) \mathcal{A} pullback attracts all families of bounded subsets of \mathfrak{X} , i.e.,

$$\lim_{t_0 \rightarrow -\infty} \text{dist}(\varphi(t, t_0)X(t_0), A(t)) = 0 \quad \text{for any fixed } t \in \mathbb{R}.$$

Definition 3. A random set $\omega \mapsto \mathcal{A}(\omega)$ is called a global random attractor for a continuous random dynamical system $\{\psi(t, \omega)\}_{t \geq 0, \omega \in \Omega}$ on \mathfrak{X} if

- (i) $\mathcal{A}(\omega)$ is a compact set of \mathfrak{X} for a.e. $\omega \in \Omega$;
- (ii) \mathcal{A} is invariant under ψ , i.e., $\psi(t, \omega)\mathcal{A}(\omega) = \mathcal{A}(\theta_t \omega)$ for all $t \geq 0$ and a.e. $\omega \in \Omega$;
- (iii) \mathcal{A} pullback attracts all families of tempered random sets of \mathfrak{X} , i.e.,

$$\lim_{t \rightarrow \infty} \text{dist}(\psi(t, \theta_{-t} \omega)X(\theta_{-t} \omega), \mathcal{A}(\omega)) = 0 \quad \text{for any } X \in \mathcal{D}(\mathfrak{X}) \text{ and a.e. } \omega \in \Omega,$$

where $\mathcal{D}(\mathfrak{X})$ denotes the set of all tempered random sets of \mathfrak{X} .

Proposition 1. Let $\{\mathcal{S}(t)\}_{t \geq 0}$ be a semigroup (dynamical system) of continuous operator on \mathfrak{X} . If $\{\mathcal{S}(t)\}_{t \geq 0}$ has a bounded absorbing set Λ and is asymptotically compact, i.e., $\{\mathcal{S}(t_n)x_n\}$ is precompact in \mathfrak{X} for every bounded sequence $\{x_n\}$ in \mathfrak{X} and $t_n \rightarrow \infty$, then $\{\mathcal{S}(t)\}_{t \geq 0}$ has a global attractor in \mathfrak{X} .

Proposition 2. Let $\{\varphi(t, t_0)\}_{t \geq t_0}$ be a continuous two-parameter semigroup (nonautonomous dynamical system) on \mathfrak{X} . If $\{\varphi(t, t_0)\}_{t \geq t_0}$ has a family of compact pullback absorbing sets $\Lambda = \{\Lambda(t)\}_{t \in \mathbb{R}}$ and is asymptotically compact, i.e., for all $\tau \in \mathbb{R}$ the sequence $\{\varphi(t_n, \tau - t_n)x_n\}$ has a convergent subsequence in \mathfrak{X} for every $t_n \rightarrow \infty$ and $x_n \in \Lambda(\tau - t_n)$, then $\{\varphi(t, t_0)\}_{t \geq t_0}$ has a pullback attractor.

Proposition 3. Let $\{\psi(t, \omega)\}_{t \geq 0, \omega \in \Omega}$ be a continuous random dynamical system with state space \mathfrak{X} . If $\{\psi(t, \omega)\}_{t \geq 0, \omega \in \Omega}$ has a tempered random closed absorbing set $\Lambda(\omega)$ and is asymptotically compact, i.e., for a.e. $\omega \in \Omega$ each sequence $x_n \in \psi(t_n, \theta_{-t_n} \omega)\Lambda(\theta_{-t_n} \omega)$ with $t_n \rightarrow \infty$ has a convergence subsequence in \mathfrak{X} , then $\{\psi(t, \omega)\}_{t \geq 0, \omega \in \Omega}$ has a unique global random attractor.

Outline of this chapter

We are interested in the dynamical behaviour of lattice models that arise in the biological sciences, in particular, the existence of attractors in such system. The chapter only considers models that we have investigated and will include nonautonomous, random and set-valued attractors as well as the usual autonomous ones. The chapter is organized as follows. Section 2 introduces the original work on reaction-diffusion

LDS and discusses numerical approximation of attractors for reaction-diffusion LDS. In Section 3 a recently studied reaction-diffusion LDS with delayed recovery is introduced. In Section 4 and 5 we study Hopfield type neural lattice models and neural field lattice models arising from discretization of neural field PDE models, respectively. In the end some closing remarks are given in Section 6.

2 Reaction-diffusion lattice models

Reaction-diffusion lattice models are obtained from discretizing a reaction-diffusion type PDE on a one-dimensional domain; they are LDS with the leading operator $A : \ell^2 \rightarrow \ell^2$ defined by

$$(A\mathbf{u})_i = u_{i-1} - 2u_i + u_{i+1}, \quad i \in \mathbb{Z}. \quad (2)$$

Define the operators $B, B^* : \ell^2 \rightarrow \ell^2$ by

$$(B\mathbf{u})_i = u_{i+1} - u_i, \quad (B^*\mathbf{u})_i = u_{i-1} - u_i, \quad i \in \mathbb{Z}.$$

Then it is straightforward to check that $-A = BB^* = B^*B$ and that $(B^*\mathbf{u}, \mathbf{v}) = (\mathbf{u}, B\mathbf{v})$ for any $\mathbf{u}, \mathbf{v} \in \ell^2$, and hence $(A\mathbf{u}, \mathbf{u}) = -\|B\mathbf{u}\|^2 \leq 0$ for any $\mathbf{u} \in \ell^2$. Note that in ℓ^2 this means A is negative definite since $\|B\mathbf{u}\| = 0$ implies that all components u_i are identical and hence \mathbf{u} is zero in ℓ^2 .

2.1 Classical techniques for reaction-diffusion LDS

The original paper on the LDS (1) with the operator A by Bates, Lu & Wang [9] has had a seminal influence on the investigation of attractors in LDS. There it was assumed that $\mathbf{g} = (g_i)_{i \in \mathbb{Z}} \in \ell^2$ and f is a smooth nonlinear function satisfying

Assumption 2 $sf(s) \geq 0$ for all $s \in \mathbb{R}$.

Note that Assumption 2 implies $f(0) = 0$ since f is smooth. Then for any $\mathbf{u} \in \ell^2$, $F(\mathbf{u}) := (f(u_i))_{i \in \mathbb{Z}} \in \ell^2$ and hence the LDS (1) can be written as an ODE on ℓ^2 :

$$\frac{d\mathbf{u}(t)}{dt} = \nu A\mathbf{u} - \lambda\mathbf{u} + F(\mathbf{u}) + \mathbf{g}. \quad (3)$$

Assumption 2 also implies that F is locally Lipschitz from ℓ^2 to ℓ^2 . In addition it was shown that given any initial condition $\mathbf{u}(0) = \mathbf{u}_o = (u_{o,i})_{i \in \mathbb{Z}}$ and $T > 0$, a solution of (3) is always bounded on $t \in [0, T]$. Standard existence and uniqueness theorems for ODEs on Banach spaces (see e.g., Deimling [26]) ensure the global existence and uniqueness of a solution $\mathbf{u}(t; \mathbf{u}_o) \in \mathcal{C}([0, \infty), \ell^2)$ for the equation (3). Moreover, the solution $\mathbf{u}(t; \mathbf{u}_o)$ defines a semigroup (i.e., autonomous semi-

dynamical system) $\{\mathcal{S}(t)\}_{t \geq 0}$ that maps ℓ^2 to ℓ^2 by

$$\mathcal{S}(t)\mathbf{u}_o = \mathbf{u}(t; \mathbf{u}_o), \quad t \geq 0.$$

Existence of an absorbing set

It is easy to show that the semi-group $\{\mathcal{S}(t)\}_{t \geq 0}$ on ℓ^2 has a positive invariant absorbing set. In fact, taking the inner product of (3) with $\mathbf{u} \in \ell^2$ gives

$$\frac{d}{dt} \|\mathbf{u}\|^2 + 2\nu \|B\mathbf{u}\|_2^2 + 2\lambda \|\mathbf{u}\|_2^2 = 2(F(\mathbf{u}), \mathbf{u}) + 2(\mathbf{g}, \mathbf{u}) \leq -\lambda \|\mathbf{u}\|^2 + \frac{1}{\lambda} \|\mathbf{g}\|^2,$$

and hence

$$\frac{d}{dt} \|\mathbf{u}\|^2 \leq -\lambda \|\mathbf{u}\|^2 + \frac{1}{\lambda} \|\mathbf{g}\|^2.$$

The Gronwall inequality then gives

$$\|\mathbf{u}(t)\|^2 \leq \|\mathbf{u}_o\|^2 e^{-\lambda t} + \frac{1}{\lambda} \|\mathbf{g}\|^2$$

Hence the closed and bounded subset of ℓ^2

$$\Lambda := \left\{ \mathbf{u} \in \ell^2 : \|\mathbf{u}(t)\|^2 \leq 1 + \frac{1}{\lambda} \|\mathbf{g}\|^2 \right\}$$

is a positively invariant absorbing set for the semi-group $\{\mathcal{S}(t)\}_{t \geq 0}$ on ℓ^2 .

Asymptotic compactness

A significant contribution of the paper [9] was to show that the semi-group $\{\mathcal{S}(t)\}_{t \geq 0}$ generated by the LDS (3) is asymptotically compact, from which it follows that $\{\mathcal{S}(t)\}_{t \geq 0}$ has a global attractor \mathcal{A} in ℓ^2 . Their method of proof has since been adapted and used repeatedly in a large number of other papers including all of those discussed in this chapter. The key step of the proof is to derive an *asymptotic tail* estimate for the solution $\mathbf{u}(t; \mathbf{u}_o)$ of the LDS in Λ .

Lemma 1. *For every $\varepsilon > 0$ there exist $T(\varepsilon) > 0$ and $I(\varepsilon) \in \mathbb{N}$ such that*

$$\sum_{|i| > I(\varepsilon)} |\mathbf{u}(t; \mathbf{u}_o)_i|^2 \leq \varepsilon^2$$

for all $\mathbf{u}_o \in \Lambda$ and $t \geq T(\varepsilon)$.

The proof utilizes a smooth cut-off function $\xi_m : \mathbb{Z}_+ \rightarrow [0, 1]$ with

$$\xi_m(|i|) := \xi\left(\frac{|i|}{m}\right) \text{ for } m \in \mathbb{Z}_+, i \in \mathbb{Z} \text{ where } \xi(s) \begin{cases} = 0, & 0 \leq s \leq 1 \\ \in [0, 1], & 1 \leq s \leq 2 \\ = 1, & s \geq 2 \end{cases}. \quad (4)$$

For a large fixed m (eventually determined in the proof) multiplying equation (1) by $v_i(t) = \xi_m(|i|)u_i(t)$ and summing over $i \in \mathbb{Z}$ gives

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \sum_{i \in \mathbb{Z}} \xi_m(|i|) |u_i(t)|^2 + v(Bu, Bv) + \lambda \sum_{i \in \mathbb{Z}} \xi_m(|i|) |u_i(t)|^2 \\ = \sum_{i \in \mathbb{Z}} \xi_m(|i|) u_i(t) f(u_i(t)) u_i(t) + \sum_{i \in \mathbb{Z}} \xi_m(|i|) g_i. \end{aligned}$$

After some skillful estimates this leads to

$$\frac{d}{dt} \sum_{i \in \mathbb{Z}} \xi_m(|i|) |u_i(t)|^2 + \lambda \sum_{i \in \mathbb{Z}} \xi_m(|i|) |u_i(t)|^2 \leq \frac{C}{m} + \frac{1}{\lambda} \sum_{|i| \geq m} g_i^2 \leq \frac{1}{2} \varepsilon$$

for $m \geq I(\varepsilon)$ since $\mathbf{g} = (g_i)_{i \in \mathbb{Z}} \in \ell^2$. Finally, by the Gronwall inequality,

$$\sum_{|i| \geq 2m} |u_i(t)|^2 \leq \sum_{i \in \mathbb{Z}} \xi_m(|i|) |u_i(t)|^2 \leq \varepsilon$$

for $t \geq T(\varepsilon)$ (to handle the initial condition) and $m \geq I(\varepsilon)$.

To obtain asymptotic compactness a sequence $\mathbf{u}(t_n; \mathbf{u}_{o,n})$ with $\mathbf{u}_{o,n} \in \Lambda$ and $t_n \rightarrow \infty$ is considered. Since Λ is closed and bounded convex subset of the Hilbert space ℓ^2 it is weakly compact. This gives a weakly convergent subsequence with a limit in Λ . The asymptotic tail estimate is then used to separate a finite number of terms from the small tail to show that the weak limit is in fact a strong limit. The existence of a global attractor then follows from Proposition 1. The reader is urged to study [9] carefully, where everything is clearly explained.

Finite dimensional approximations

Bates, Lu & Wang [9] also show that the $2N + 1$ -dimensional approximations of the lattice system (1) also have global attractors \mathcal{A}_N which converge upper semi continuously to the attractor \mathcal{A} in the Hausdorff semi-distance on ℓ^2 , i.e.,

$$\lim_{N \rightarrow \infty} \text{dist}_{\ell^2}(\mathcal{A}_N, \mathcal{A}) = 0. \quad (5)$$

There they consider vectors $\mathbf{x} = (x_{-N}, \dots, x_0, \dots, x_N)$ in \mathbb{R}^{2N+1} which can be extended naturally to elements of ℓ^2 with components set to zero for indices $|i| > N$. The proof of upper semi continuous convergence uses similar ideas to those for the tail estimates above.

Note that do not simply truncate the LDS (1) but assume that it has periodic boundary conditions, i.e., with $x_N(t) = x_{-N-1}(t)$ and $x_{-N}(t) = x_{N+1}(t)$. Specifically, they consider the finite-dimensional system of ODEs

$$\begin{cases} \frac{dx_{-N}}{dt} = v(x_N - 2x_{-N} + x_{-N+1}) - \lambda x_{-N} + f(x_{-N}) + g_{-N}, \\ \vdots \\ \frac{dx_i}{dt} = v(x_{i-1} - 2x_i + x_{i+1}) - \lambda x_i + f_i(x_i) + g_i, \quad i = -N+1, \dots, N-1, \\ \vdots \\ \frac{dx_N}{dt} = v(x_{N-1} - 2x_N + x_{-N}) - \lambda x_N + f(x_N) + g_N. \end{cases} \quad (6)$$

2.2 Numerical approximation of lattice attractors

A general theorem of Kloeden & Lorenz [52] (see also [36]) can be applied to conclude that a one-step numerical scheme with constant time stepsize h applied to the ODE system (6) has an attractor $\mathcal{A}_N^{(h)}$, which converges upper semi continuously to \mathcal{A}_N for each N , i.e.,

$$\lim_{h \rightarrow 0+} \text{dist}_{\mathbb{R}^{2N+1}} \left(\mathcal{A}_N^{(h)}, \mathcal{A}_N \right) = 0. \quad (7)$$

Thus, combining the convergences (5) and (7), we see that $\mathcal{A}_N^{(h)}$ can be used as an approximation for the lattice attractor \mathcal{A} for the LDS (1) when h is small enough and N is large enough.

Han, Kloeden & Sonner [39] also consider the numerical approximation of the attractor of the LDS (1). They focus on the implicit Euler scheme (IES) and first apply it to ODE (3) in the space ℓ^2 , where it has the form

$$\mathbf{u}_{n+1}^{(h)} = \mathbf{u}_n^{(h)} + hA\mathbf{u}_{n+1}^{(h)} - h\lambda\mathbf{u}_{n+1}^{(h)} + hF\left(\mathbf{u}_{n+1}^{(h)}\right) + h\mathbf{g}, \quad n \in \mathbb{N}, \quad \mathbf{u}_o \in \ell^2. \quad (8)$$

They show that the IES (8) is uniquely solvable for small enough stepsize, hence generates a discrete-time semi-dynamical system. Moreover, this numerical system has an absorbing sets and an attractor $\mathcal{A}_\infty^{(h)}$, which converges upper semi continuously to \mathcal{A} as $h \rightarrow 0+$.

The attractor $\mathcal{A}_\infty^{(h)}$ is useful for theoretical purposes, but actual computations must be done in finite dimensions. In [39] the implicit Euler scheme is also applied to the finite dimensional system of ODEs (6) and shown to have an attractor $\mathcal{A}_N^{(h)}$. By a compactness argument similar to the asymptotic tail estimates, it is shown that the attractors $\mathcal{A}_N^{(h)}$ converge upper semi continuously to $\mathcal{A}_\infty^{(h)}$ for a fixed stepsize h , i.e.,

$$\lim_{N \rightarrow \infty} \text{dist}_{\ell^2} \left(\mathcal{A}_N^{(h)}, \mathcal{A} \right) = 0.$$

Notice that the finite dimensional IES have common range of step-sizes and common absorbing set when extended to ℓ^2 , which is independent of the dimension $2N + 1$.

There are thus two paths for computing an approximation of the LDS attractor, essentially first approximating space then time or approximating time then space. This is illustrated in Figure 1.

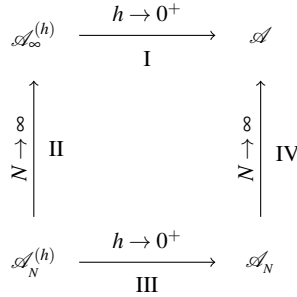


Fig. 1 Convergence paths for the approximated numerical attractor to the analytical attractor.

3 A lattice reaction-diffusion model with delayed recovery

Motivated by the appearance of switching effects and recovery delays in systems of excitable cells [47, 64] and time-dependent structure of reactions, Han & Kloeden [34] studied a non-autonomous LDS with a reaction term which is switched off when a certain threshold is exceeded and restored after a suitable recovery time:

$$\dot{u}_i = \nu(A\mathbf{u})_i + f_i(t, u_i) \cdot \mathfrak{h}(\Theta_i - \max_{-\tau \leq s \leq 0} u_i(t+s)), \quad i \in \mathbb{Z}, \quad t > t_0 \quad (9)$$

$$u_i(s) = \phi_i(s - t_0), \quad \forall s \in [t_0 - \tau, t_0], \quad i \in \mathbb{Z}, \quad t_0 \in \mathbb{R}.$$

Here $\nu > 0$ is reciprocal of the inter-cellular resistance [50], $(A\mathbf{u})_i$ is the discretized Laplacian operator as in (2), and \mathfrak{h} is the heaviside function defined by

$$\mathfrak{h}(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}. \quad (10)$$

For each $i \in \mathbb{Z}$: $u_i \in \mathbb{R}$ represents the membrane potential of the cell at the i -th active site; $\Theta_i \in \mathbb{R}$ is the threshold triggering the switch-off at the i -th site; $u_i(t + \cdot) \in$

$\mathcal{C}([-\tau, 0], \mathbb{R})$ is the segment of u_i on time interval $[t - \tau, t]$ where τ is a positive constant representing the time-delay.

The LDS (23) describes a reaction-diffusion system with or without delay depending on the maximum value achieved at each location during the past τ period of time. More specifically, each u_i evolves with respect to a reaction-diffusion equation with a delay in the reaction term as long as u_i stays below the threshold Θ_i starting from t_0 . Once the value of u_i reaches or exceeds Θ_i at some time, the reaction will be switched off, and stay off for at least τ period of time. The delay in the reaction term can be recovered later, when u_i evolves back to smaller values than Θ_i and remains smaller than Θ_i for more than τ period of time.

The novelty and difficulty of system (9) lies in that the switching off and on of the reaction term leads to a relaxation effect, essentially multiplication by a Heaviside function, and thus to the formulation of the system as a set-valued system, i.e., inclusion differential equation. In view of the recovery time prescribed between switching off and back on again introduces a delay term, which appears only upper semi continuously.

Existence of solutions

Consider the function spaces

$$E_1 = \mathcal{C}([-\tau, 0], \mathbb{R}), \quad E_{\mathcal{C}} = \mathcal{C}([-\tau, 0], \ell^2), \quad E_{\infty} = \mathcal{C}([-\tau, 0], \ell^{\infty}),$$

with norms

$$\|\cdot\|_{E_1} = \max_{s \in [-\tau, 0]} |\cdot(s)|, \quad \|\cdot\|_{E_{\mathcal{C}}} = \max_{s \in [-\tau, 0]} \|\cdot(s)\|, \quad \|\cdot\|_{E_{\infty}} = \max_{s \in [-\tau, 0]} \|\cdot(s)\|_{\infty}.$$

In addition, denote by $E_{\mathcal{B}} = \mathcal{C}_{\mathcal{B}}(\mathbb{R}, \ell^2)$ the space of all continuous bounded functions from \mathbb{R} into ℓ^2 with norm $\|\cdot\|_{E_{\mathcal{B}}} = \sum_{t \in \mathbb{R}} \|\cdot(t)\|$.

Denote by $\mathbf{u}_t = (u_{it})_{i \in \mathbb{Z}}$ where $u_{it}(s) = u_i(t + s)$, $s \in [-\tau, 0]$ and define the set-valued mapping $\mathfrak{F} : \mathbb{R} \times E_{\mathcal{C}} \rightarrow \mathcal{P}(\ell^2)$ by $\mathfrak{F}(t, \mathbf{u}_t) := (F_i(t, u_{it}))_{i \in \mathbb{Z}}$, where

$$F_i(t, u_{it}) = \begin{cases} f_i(t, u_i), & \max_{s \in [-\tau, 0]} u_{it}(s) < \Theta_i \\ f_i(t, u_i) \cdot [0, 1], & \max_{s \in [-\tau, 0]} u_{it}(s) = \Theta_i \\ 0, & \max_{s \in [-\tau, 0]} u_{it}(s) > \Theta_i \end{cases}$$

Writing the initial condition as $\boldsymbol{\phi}(\cdot) := (\phi_i(\cdot))_{i \in \mathbb{Z}}$, then the lattice equation (9) can be written as the delay differential inclusion equation

$$\frac{d\mathbf{u}}{dt} \in \mathbf{v}A\mathbf{u} + \mathfrak{F}(t, \mathbf{u}_t), \quad \mathbf{u}(s) = \boldsymbol{\phi}(s - t_0), \quad \forall s \in [t_0 - \tau, t_0]. \tag{11}$$

Recall that the bounded linear operator $A : \ell^2 \rightarrow \ell^2$ defined by (2) generates a uniformly continuous semigroup $\{\mathcal{S}(t)\}_{t \geq 0}$.

Definition 4. A continuous function $\mathbf{u} : [t_0, T] \rightarrow \ell^2$ is called a (strong) solution of problem (11) if there exists a selection function

$$f(t) \in \mathfrak{M}_{\mathfrak{F}} := \{f \in \mathcal{L}^1(t_0, T; \ell^2) : f(t) \in \mathfrak{F}(t, \mathbf{u}_t), \text{ for a.e. } t \in [t_0, T]\}$$

such that \mathbf{u} is a (strong) solution to the corresponding auxiliary problem, i.e., if \mathbf{u} is absolutely continuous on any compact subinterval of $[t_0, T]$ with $\mathbf{u}(t_0) = \mathbf{u}_o$ and $\mathbf{u}'(t) = \mathbf{v}\mathbf{A}\mathbf{u}(t) + f(t)$ in ℓ^2 for Lebesgue-a.e. $t \in (t_0, T)$.

To ensure equation (11) has a (strong) solution, the following assumptions on functions f_i are imposed in [34]:

Assumption 3 $f_i : \mathbb{R} \rightarrow \mathbb{R}$ is continuous in t for each $i \in \mathbb{Z}$;

Assumption 4 there exists a constant $a \geq 0$, and $\mathbf{b}(t) = (b_i(t))_{i \in \mathbb{Z}} \in E_{\mathcal{B}}$ such that

$$f_i^2(t, x) \leq ax^2 + b_i^2(t), \quad \forall s \in \mathbb{R};$$

Assumption 5 for any $x \in \mathbb{R}$, there exist $\delta > 0$ and $\mathbf{L}(t) = (L_i(t))_{i \in \mathbb{Z}} \in E_{\mathcal{B}}$ such that

$$|f_i(t, y) - f_i(t, x)| \leq |L_i(t)||y - x|, \quad i \in \mathbb{Z}$$

for all $|y - x| < \delta$.

It is straightforward to show that the map $\mathfrak{F} : \mathbb{R} \times E_{\mathcal{E}} \rightarrow \mathcal{P}(\ell^2)$ is bounded, closed and convex, but more difficult to show that \mathfrak{F} is upper semi-continuous. To that end, the integer set \mathbb{Z} is divided into three time-dependent sub-indices,

$$\begin{aligned} J_0^h &= \left\{ i \in \mathbb{Z} : \max_{s \in [-\tau, 0]} h_i(s) > \Theta_i \right\}, \\ J_1^h &= \left\{ i \in \mathbb{Z} : \max_{s \in [-\tau, 0]} h_i(s) < \Theta_i \right\}, \\ J_{\Theta}^h &= \left\{ i \in \mathbb{Z} : \max_{s \in [-\tau, 0]} h_i(s) = \Theta_i \right\}, \end{aligned}$$

on each of which it was shown that $\text{dist}_{\ell^2}(\mathfrak{F}(t, h_1), \mathfrak{F}(t, h_2))$ is small provided $\|h_1 - h_2\|_{E_{\mathcal{E}}}$ is small. The next theorem then follows from known results on differential inclusion equations in Banach spaces.

Theorem 1. Assume that Assumptions 3 – 5 hold. Then for any $t_0 \in \mathbb{R}$, problem (11) has at least one solution, $\mathbf{u}(\cdot) = \mathbf{u}(\cdot; t_0, \phi)$. Moreover, the solutions define a two-parameter set-valued semi-group or nonautonomous set-valued dynamical system ϕ on $E_{\mathcal{E}}$ by

$$\phi(t, t_0)\phi = \{\mathbf{u}_t(\cdot; t_0, \phi) \in E_{\mathcal{E}} : \mathbf{u}(\cdot) \text{ is a solution to (11) with } \phi \in E_{\mathcal{E}}\}.$$

Existence of a global attractor

To obtain the existence of a global attractor the following dissipative conditions are needed:

Assumption 6 *there exists $\alpha > 0$ with $\alpha^2 > a > 0$ and $\beta(t) := (\beta_i(t))_{i \in \mathbb{Z}} \in H_{\mathcal{B}}$ such that*

$$f_i(t, x)x \leq -\alpha x^2 + \beta_i^2(t), \quad i \in \mathbb{Z}.$$

Under Assumptions 3 – 6 it can be shown that every solution $\mathbf{u}(\cdot)$ to (11) satisfies

$$\|\mathbf{u}_t\|_{E_{\mathcal{C}}}^2 \leq \|\phi\|_{E_{\mathcal{C}}}^2 e^{\lambda(\tau+t_0-t)} + \frac{2\alpha\|\beta\|^2 + \|b\|^2}{\lambda} (1 - e^{\lambda(t_0-t)}).$$

Hence the closed and bounded set

$$\Lambda := \left\{ u \in E_{\mathcal{C}} : \|u\|_{E_{\mathcal{C}}}^2 \leq 1 + \frac{2\alpha\|\beta\|^2 + \|b\|^2}{\lambda} \right\}$$

is a positive invariant absorbing set for φ .

Asymptotic tail estimations can also be constructed by using the cut-off function ξ defined in (4), but estimations have to be done separately on the three time-dependent sub-indices J_0^h, J_1^h and J_{Θ}^h respectively. Then a modification of asymptotic compactness arguments in [9] shows that φ is asymptotic compact. It then follows from Proposition 2 that the nonautonomous dynamical system $\{\varphi(t, t_0)\}_{t \geq t_0}$ has a pullback attractor in $E_{\mathcal{C}}$.

4 Hopfield neural lattice models

In 1984 John Hopfield [44] introduced a system of n ordinary differential equations (ODEs) to model the interaction between a network of n neurons. It has since found many diverse applications and, in particular, is one of the most popular mathematical models for investigating neural networks in artificial intelligence. It is now referred to as the *Hopfield neural network* and given by

$$\mu_i \frac{du_i(t)}{dt} = -\frac{u_i(t)}{\gamma_i} + \sum_{j=1}^n \lambda_{i,j} f_j(u_j(t)) + g_i, \quad i = 1, \dots, n, \quad (12)$$

where u_i represents the voltage on the input of the i th neuron at time t ; $\mu_i > 0$ and $\gamma_i > 0$ represents the neuron amplifier input capacitance and resistance of the i th neuron, respectively; and g_i is the constant external forcing on the i th neuron.

Here n is the total number of neurons coupled by an $n \times n$ matrix $(\lambda_{i,j})_{1 \leq i, j \leq n}$, where $\lambda_{i,j}$ represents the connection strength between the i th and the j th the neuron. More precisely, for each pair of $i, j = 1, \dots, n$, $\lambda_{i,j}$ is the synapse efficacy between neurons i and j , and thus $\lambda_{i,j} > 0$ ($\lambda_{i,j} < 0$, resp.) means the output of neuron j ex-

cites (inhibits, resp.) neuron i . The term $\lambda_{i,j}f_j(u_j(t))$ represents the electric current input to neuron i due to the present potential of neuron j , in which the function f_j is neuron activation functions and assumed to be a sigmoid type function.

We are interested in studying dynamics of the above Hopfield neural network model when its size becomes increasingly large, i.e., $n \rightarrow \infty$. To this end, we extend the n dimensional ODE system (12) to an infinite dimensional lattice system, that models the dynamics of an infinite number of neurons indexed by $i \in \mathbb{Z}$, in which each neuron is still connected with other neurons within its finite n neighborhood. More precisely, the i th neuron is connected to the $(i - n)$ th, \dots , $(i + n)$ th neurons through the strength matrix $(\lambda_{i,j})_{i-n \leq j \leq i+n}$ and the activation functions f_j for $j = i - n, \dots, i + n$. System (12) then becomes the following LDS, namely the Hopfield neural lattice model:

$$\mu_i \frac{du_i(t)}{dt} = -\frac{u_i(t)}{\gamma_i} + \sum_{j=i-n}^{i+n} \lambda_{i,j}f_j(u_j(t)) + g_i, \quad i \in \mathbb{Z}. \tag{13}$$

The model parameters are assumed to satisfy:

Assumption 7 *the efficacy between each pair of neurons is finite, i.e., there exists $M_\lambda > 0$ such that*

$$\sup_{i,j \in \mathbb{Z}} |\lambda_{i,j}| \leq M_\lambda;$$

Assumption 8 *the neuron amplifier input capacitance and resistance are uniformly bounded, i.e., there exist positive constants m_μ, M_ν, m_γ , and M_γ , such that*

$$m_\mu \leq \mu_i \leq M_\mu, \quad m_\gamma \leq \gamma_i \leq M_\gamma, \quad \forall i \in \mathbb{Z}.$$

For existence of a global solution and global attractor, the forcing term and the neuron activation function are assumed to satisfy:

Assumption 9 *the neuron activation function satisfies $f_i \in \mathcal{C}^1(\mathbb{R}, \mathbb{R})$ and $f_i(0) = 0$ for all $i \in \mathbb{Z}$. Moreover, there exists a continuous non-decreasing function $L(r) \in \mathcal{C}(\mathbb{R}_+, \mathbb{R}_+)$ such that*

$$\sup_{i \in \mathbb{Z}} \max_{s \in [-r,r]} |f'_i(s)| \leq L(r) \quad \forall r \in \mathbb{R}_+;$$

Assumption 10 *for each $i \in \mathbb{Z}$ there exist $\alpha > 0$ and $\boldsymbol{\beta} = (\beta_i)_{i \in \mathbb{Z}} \in \ell^2$ such that*

$$sf_i(s) \leq -\alpha s^2 + \beta_i^2, \quad \forall s \in \mathbb{R}.$$

Assumption 11 *the aggregated forcing on the whole network is finite in the sense that $(g_i)_{i \in \mathbb{Z}} \in \ell^2$.*

Existence of solutions

Note that Assumption 9 essentially requires each neuron activation function f_j to be locally Lipschitz continuous uniformly in $i \in \mathbb{Z}$. In fact, given any $\mathbf{u} = (u_i)_{i \in \mathbb{Z}} \in \ell^2$, for each $i \in \mathbb{Z}$ there exists $\zeta_i \in \mathbb{R}$ with $|\zeta_i| \leq |u_i|$ such that

$$|f_i(u_i)| = |f'_i(\zeta_i)u_i| \leq L(|u_i|)|u_i| \leq L(\|\mathbf{u}\|)|u_i|, \quad \forall i \in \mathbb{Z}. \quad (14)$$

Given any $\mathbf{u} = (u_i)_{i \in \mathbb{Z}}$, define the operators $\Gamma \mathbf{u} = ((\Gamma \mathbf{u})_i)_{i \in \mathbb{Z}}$, $F \mathbf{u} = ((F \mathbf{u})_i)_{i \in \mathbb{Z}}$ and function \mathbf{g} by

$$(\Gamma \mathbf{u})_i = \frac{u_i}{\gamma_i \mu_i}; \quad (F \mathbf{u})_i = \sum_{j=i+n}^{i+n} \frac{\lambda_{i,j}}{\mu_i} f_j(u_j); \quad \mathbf{g} = \left(\frac{g_i}{\mu_i} \right)_{i \in \mathbb{Z}}. \quad (15)$$

By Assumption 8 and 11, $\mathbf{g} \in \ell^2$ and $\Gamma : \ell^2 \rightarrow \ell^2$. Moreover, by the inequality (14),

$$\begin{aligned} \|F \mathbf{u}\|^2 &= \sum_{i \in \mathbb{Z}} \left(\sum_{j=i+n}^{i+n} \frac{\lambda_{i,j}}{\mu_i} f_j(u_j) \right)^2 \leq \frac{M_\lambda^2}{m_\mu^2} \sum_{i \in \mathbb{Z}} \left(\sum_{j=i+n}^{i+n} f_j(u_j) \right)^2 \\ &\leq \frac{M_\lambda^2}{m_\mu^2} \sum_{i \in \mathbb{Z}} [(2n+1)L(\|\mathbf{u}\|)|u_i|]^2 = (2n+1)^2 \frac{M_\lambda^2}{m_\mu^2} L^2(\|\mathbf{u}\|) \cdot \|\mathbf{u}\|^2, \end{aligned}$$

which implies that F maps ℓ^2 to ℓ^2 . Therefore the LDS (13) can be written as an ODE on ℓ^2 :

$$\frac{d\mathbf{u}}{dt} = -\Gamma \mathbf{u} + F \mathbf{u} + \mathbf{g} := P(\mathbf{u}) \quad (16)$$

It follows from Assumptions 8 and 9 that $P(\mathbf{u})$ is locally Lipschitz and thus given any initial condition $\mathbf{u}_o = (u_{o,i})_{i \in \mathbb{Z}}$ the equation (16) has a unique local solution $\mathbf{u}(t; \mathbf{u}_o) \in \mathcal{C}([0, T], \ell^2)$. The dissipativity condition 10 ensures that the local solution $\mathbf{u}(t; \mathbf{u}_o)$ is always bounded on $t \in [0, T]$ for every finite T . Hence the solution $\mathbf{u}(t; \mathbf{u}_o)$ of equation (16) exists globally in time, i.e., $\mathbf{u}(t; \mathbf{u}_o) \in \mathcal{C}([0, \infty), \ell^2)$. Moreover, the solution depends continuously on the initial data. Therefore the solution $\mathbf{u}(t; \mathbf{u}_o)$ defines a continuous dynamical system $\{\mathcal{S}(t)\}_{t \geq 0}$ that maps ℓ^2 to ℓ^2 by

$$\mathcal{S}(t)\mathbf{u}_o = \mathbf{u}(t; \mathbf{u}_o), \quad t \geq 0, \quad \mathbf{u}_o \in \ell^2.$$

Existence of attractor

To construct an absorbing set for $\{\mathcal{S}(t)\}_{t \geq 0}$, multiply the (13) by $u_i(t)$ and sum over all $i \in \mathbb{Z}$ to obtain

$$\frac{1}{2} \frac{d\|\mathbf{u}(t)\|^2}{dt} = - \sum_{i \in \mathbb{Z}} \frac{u_i^2}{\mu_i \gamma_i} + \sum_{i \in \mathbb{Z}} \frac{u_i}{\mu_i} \sum_{j=i-n}^{i+n} \lambda_{i,j} f_j(u_j) + \sum_{i \in \mathbb{Z}} \frac{u_i}{\mu_i} g_i.$$

The main technical difficulty is to estimate the term $\sum_{i \in \mathbb{Z}} \frac{u_i}{\mu_i} \sum_{j=i-n}^{i+n} \lambda_{i,j} f_j(u_j)$, because each $\lambda_{i,j}$ can be either positive or negative and thus the dissipativity of f_j may either stay dissipative or give rise to a growth. Thus the exciting neuron pairs ($\lambda_{i,j} > 0$) and inhibiting neuron pairs ($\lambda_{i,j} < 0$) have to be analyzed separately. Another difficulty comes from the mismatched product term $u_i f_j(u_j)$, to which the dissipative assumption 10 is not directly applicable. In fact,

$$\begin{aligned} u_i f_j(u_j) &= (u_i - u_j) f_j(u_j) + u_j f_j(u_j) \leq L \left(u_j^2 + \frac{Lu_i^2}{2\alpha} + \frac{\alpha u_j^2}{2L} \right) - \alpha u_j^2 + \beta_j^2 \\ &\leq -\frac{\alpha}{2} u_j^2 + Lu_j^2 + \frac{L^2}{2\alpha} u_i^2 + \beta_j^2, \quad \forall i, j \in \mathbb{Z}. \end{aligned} \tag{17}$$

To avoid further complications assume that $L(r) \equiv L$, then the following Lemma holds [43].

Lemma 2. *Assume that assumptions 7–11 hold. Then the continuous dynamical system $\{\mathcal{S}(t)\}_{t \geq 0}$ generated by solutions to the system (16) has a positive invariant bounded absorbing set Λ provided*

$$\inf_{i \in \mathbb{Z}} \min_{\substack{|j-i| \leq n \\ \lambda_{i,j} > 0}} \lambda_{i,j} > 0 \quad \text{and} \quad \frac{1}{M_\mu M_\gamma} + (2n + 1) \left[\frac{\alpha m_\lambda}{M_\mu} - \frac{LM_\lambda}{m_\mu} \left(\frac{9}{2} + \frac{L}{\alpha} \right) \right] > 0.$$

The asymptotic compactness of Λ under $\{\mathcal{S}(t)\}_{t \geq 0}$ is also established following a tail estimate by using a continuous, increasing and sub-additive cut-off function $\xi_m : \mathbb{Z}_+ \rightarrow [0, 1]$ satisfying (4). The existence of a global attractor then follows from Proposition 1.

Upper semi continuous convergence of attractors

The upper semi continuity of global attractors is of crucial importance when numerical simulations of an LDS are sought. More precisely, a numerical simulation of an (infinite dimensional) LDS requires a simplified finite dimensional approximation of the original infinite system, and the convergence of the global attractor ensures that numerical solutions based on the simplified finite dimensional system mimic the solutions for the infinite lattice system in the limit.

To investigate the upper semi continuity of the global attractor for syste (13), consider the following $(2N + 1)$ -dimensional system of ODEs obtained from directly truncating the lattice system (13):

$$\mu_i \frac{d}{dt} u_i(t) = -\frac{u_i(t)}{\gamma_i} + \sum_{j=i-n}^{i+n} \lambda_{i,j} f_j(u_j(t)) + g_i, \quad i = -N, \dots, 0, \dots, N. \tag{18}$$

Since every neuron is interacting with $2n$ neurons in the neighborhood ordered by their indices, we assume that $N \geq n$ to capture enough dynamics of the network.

In order for the finite-dimensional ODE system (18) to be well-posed, boundary conditions, i.e., on terms out of the range $i = -N, \dots, N$, need to be imposed. Here we assume a Dirichlet type boundary conditions:

Assumption 12 $u_i(t) \equiv 0$ for $i = -N - n, \dots, -N - 1$ and $i = N + 1, \dots, N + n$.

For any $N \in \mathbb{N}$, denote by $\mathbf{u}(t) = (u_{-N}, \dots, u_0, \dots, u_N) \in \mathbb{R}^{2N+1}$, and

$$\Gamma^N := \begin{pmatrix} -\frac{1}{\mu_{-N}\gamma_{-N}} & 0 & \dots & 0 & 0 \\ 0 & -\frac{1}{\mu_{-N+1}\gamma_{-N+1}} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \dots & 0 & -\frac{1}{\mu_N\gamma_N} \end{pmatrix} \in \mathbb{R}^{(2N+1) \times (2N+1)}.$$

In addition, let

$$\mathbf{g}^N = \left(\frac{g_{-N}}{\mu_{-N}}, \dots, \frac{g_N}{\mu_N} \right) \in \mathbb{R}^{2N+1}.$$

Then under Assumption 12 and 9 the $(2N + 1)$ -dimensional ODE system (18) becomes

$$\frac{d\mathbf{u}(t)}{dt} = \Gamma^N \mathbf{u} + F^N(\mathbf{u}) + \mathbf{g}^N, \tag{19}$$

where $F^N(\mathbf{u}) = (F_i^N(\mathbf{u}))_{i=-N, \dots, N}$ is defined by

$$F_i^N = \begin{cases} \frac{1}{\mu_i} \sum_{j=-N}^{i+n} \lambda_{i,j} f_j(\mathbf{u}_j), & i = -N, \dots, -N + n - 1 \\ \frac{1}{\mu_i} \sum_{j=i-n}^{i+n} \lambda_{i,j} f_j(\mathbf{u}_j), & i = -N + n, \dots, N - n \\ \frac{1}{\mu_i} \sum_{j=i-n}^N \lambda_{i,j} f_j(\mathbf{u}_j), & i = N - n + 1, \dots, N \end{cases}.$$

The above ODE system is well-posed and thus given any initial condition $\mathbf{u}(0) = \mathbf{u}_o$ with $\mathbf{u}_o = (u_{o,-N}, \dots, u_{o,0}, \dots, u_{o,N}) \in \mathbb{R}^{2N+1}$, the equation (19) has a unique solution $\mathbf{u}(t; \mathbf{u}_o) \in \mathcal{C}([0, \infty), \mathbb{R}^{2N+1}) \cap \mathcal{C}^1((0, \infty), \mathbb{R}^{2N+1})$. Moreover, the solution defines a dynamical system $\{\mathcal{S}_N(t)\}_{t \geq 0}$ that maps \mathbb{R}^{2N+1} to \mathbb{R}^{2N+1} by $\mathcal{S}_N(t)\mathbf{u}_o = \mathbf{u}(t; \mathbf{u}_o)$.

To construct an absorbing set, denoting by $|\cdot|$ the Euclidean norm of \mathbb{R}^{2N+1} and taking the inner product of (19) with \mathbf{u} in \mathbb{R}^{2N+1} to get

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} |u(t)|^2 = & - \sum_{i=-N}^N \frac{1}{\mu_i \gamma_i} u_i^2(t) + \sum_{i=-N}^N \frac{u_i}{\mu_i \gamma_i} g_i + \underbrace{\sum_{i=-N}^{-N+n-1} \frac{u_i}{\mu_i} \sum_{j=-N}^{i+n} \lambda_{i,j} f_j(u_j)}_{(i)} \\ & + \underbrace{\sum_{i=-N+n}^{N-n} \frac{u_i}{\mu_i} \sum_{j=i-n}^{i+n} \lambda_{i,j} f_j(u_j)}_{(ii)} + \underbrace{\sum_{i=N-n+1}^N \frac{u_i}{\mu_i} \sum_{j=i-n}^N \lambda_{i,j} f_j(u_j)}_{(iii)}, \end{aligned}$$

in which

$$- \sum_{i=-N}^N \frac{1}{\mu_i \gamma_i} u_i^2(t) + \sum_{i=-N}^N \frac{u_i}{\mu_i \gamma_i} g_i \leq - \frac{1}{2M_\mu M_\gamma} |u(t)|^2 + \frac{M_\mu M_\gamma}{2m_\mu^2} |g^N|^2.$$

Then using Assumption 9 with $L(r) \equiv L$, Assumption 10, the inequality (17), and computing terms with $\lambda_{i,j} > 0$ and $\lambda_{i,j} < 0$ separately we have

$$\begin{aligned} (i) & \leq - \frac{\alpha m_\lambda}{M_\mu} n |u(t)|^2 + \frac{3LM_\lambda}{2m_\mu} n |u(t)|^2 + \frac{LM_\lambda}{m_\mu} n \left(\frac{L}{\alpha} + 1 \right)^{-N+n-1} \sum_{i=-N}^{i+n-1} u_i^2 + \frac{nM_\lambda}{m_\mu} \|\beta\|^2 \\ (ii) & \leq \frac{3LM_\lambda}{2m_\mu} (2n+1) |u(t)|^2 + \frac{LM_\lambda}{m_\mu} (2n+1) \left(\frac{L}{\alpha} + 1 \right)^{\sum_{i=-N+n}^{N-n}} u_i^2 + \frac{(2n+1)M_\lambda}{m_\mu} \|\beta\|^2, \\ (iii) & \leq - \frac{\alpha m_\lambda}{M_\mu} n |u(t)|^2 + \frac{3LM_\lambda}{2m_\mu} n |u(t)|^2 + \frac{LM_\lambda}{m_\mu} n \left(\frac{L}{\alpha} + 1 \right)^{\sum_{i=N-n+1}^N} u_i^2 + \frac{nM_\lambda}{m_\mu} \|\beta\|^2. \end{aligned}$$

Summarizing the above results in

$$\frac{1}{2} \frac{d}{dt} |u(t)|^2 \leq -C |u(t)|^2 + \frac{M_\mu M_\gamma}{2m_\mu^2} \|g\|^2 + (4n+1) \frac{M_\lambda}{m_\mu} \|\beta\|^2,$$

where

$$C = \frac{1}{2M_\mu M_\gamma} + 2n \frac{\alpha m_\lambda}{M_\mu} - \frac{3LM_\lambda}{2m_\mu} (4n+1) - \frac{LM_\lambda}{m_\mu} (2n+1).$$

Therefore, provided $C > 0$, the dynamical system $\{\mathcal{S}_N(t)\}_{t \geq 0}$ has an absorbing set in \mathbb{R}^{2N+1}

$$\Lambda_N := \left\{ u \in \mathbb{R}^{2N+1} : |u| \leq \frac{M_\mu M_\gamma}{4Cm_\mu^2} \|g\|^2 + \frac{4n+1}{2C} \frac{M_\lambda}{m_\mu} \|\beta\|^2 + 1 \right\}.$$

Notice that Λ_N depends only on model parameters, but not N . It then follows directly from Proposition 1 that the dynamical system $\{\mathcal{S}_N(t)\}_{t \geq 0}$ has a global attractor $\mathcal{A}_N \subset \Lambda_N$. Moreover, using a contradiction argument, it can be shown that the global attractors \mathcal{A}_N , with natural embedding in ℓ^2 , converge to the global attractor \mathcal{A} of the LDS (13) upper semi continuously, as $N \rightarrow \infty$. More precisely,

$$\lim_{N \rightarrow \infty} \text{dist}_{\ell^2}(\mathcal{A}_N, \mathcal{A}) = 0,$$

where $\text{dist}_{\ell^2}(\mathcal{A}_N, \mathcal{A}) = \sup_{a \in \mathcal{A}_N} \text{dist}_{\ell^2}(a, \mathcal{A}) = \sup_{a \in \mathcal{A}_N} \inf_{b \in \mathcal{A}} \|a - b\|$.

A random Hopfield neural lattice model

To take into account random perturbations of the environment, we introduce a noise in the equations in (12) by replacing each constant input g_i by a random forcing $g_i(\theta_t, \omega)$ represented by a measure-preserving driving dynamical system $\{\theta_t\}_{t \in \mathbb{R}}$ acting on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For basic concepts of the driving dynamical system $\{\theta_t\}_{t \in \mathbb{R}}$ and random dynamical system the reader is referred to [5]. The LDS (13) then becomes

$$\mu_i \frac{du_i(t)}{dt} = -\frac{u_i(t)}{\gamma_i} + \sum_{j=i-n}^{i+n} \lambda_{i,j} f_j(u_j(t)) + g_i(\theta_t, \omega), \quad i \in \mathbb{Z}. \tag{20}$$

The assumptions 8 through 10 remain the same, but due to the randomness of g_i , Assumption 11 needs to be replaced by

Assumption 13 $(g_i(\omega))_{i \in \mathbb{Z}} \in \ell^2, \quad \forall \omega \in \Omega$.

The LDS (20) can then be written as a random ordinary differential equation (RODE) on ℓ^2 :

$$\frac{d\mathbf{u}}{dt} = -\Gamma \mathbf{u} + F \mathbf{u} + \mathbf{g}(\theta_t, \omega), \tag{21}$$

where $\Gamma \mathbf{u}$ and $F \mathbf{u}$ are as defined in (15) and $\mathbf{g}(\theta_t, \omega) := (g_i(\theta_t, \omega)/\gamma_i)_{i \in \mathbb{Z}}$.

Following similar computations for the ODE (16) and using existence theorem for RODEs [35], it can be shown that given initial condition $\mathbf{u}(t_0) = \mathbf{u}_o \in \ell^2$, a unique solution $\mathbf{u}(\cdot; t_0, \omega, \mathbf{u}_o) \in \mathcal{C}([t_0, \infty), \ell^2)$ for (21) exists globally in time for any $t_0 \in \mathbb{R}$ and $\omega \in \Omega$. Moreover the solution is continuous in $\mathbf{u}_o \in \ell^2$ and satisfies

$$\mathbf{u}(t + t_0; t_0, \omega, \mathbf{u}_o) = \mathbf{u}(t; 0, \theta_{t_0} \omega, \mathbf{u}_o), \quad \forall t \geq 0, \mathbf{u}_o \in \ell^2, \omega \in \Omega.$$

Therefore the solution of the RODE (21) defines a continuous random dynamical system $\{\psi(t, \omega)\}_{t \geq 0, \omega \in \Omega}$ by

$$\psi(t, \omega) \mathbf{u}_o = \mathbf{u}(t; 0, \omega, \mathbf{u}_o), \quad \forall t \geq 0, \mathbf{u}_o \in \ell^2, \omega \in \Omega.$$

The major difference in obtaining the existence of a random absorbing set for the random dynamical system $\{\psi(t, \omega)\}_{t \geq 0, \omega \in \Omega}$ defined by the solution of (21) and the existence of an absorbing set for the dynamical system $\{\mathcal{S}(t)\}_{t \geq 0}$ defined by the solution of (16) is due to the random forcing $\mathbf{g}(\theta_t, \omega)$. More precisely, all estimations for $\{\psi(t, \omega)\}_{t \geq 0, \omega \in \Omega}$ need to be done in the pullback sense, i.e., with the initial time $t_0 \rightarrow -\infty$ and current time t held fixed. The reader is referred to [35] and [13] and references therein for detailed explanations of pullback and forward attraction. In

[42] it was shown that

$$\begin{aligned} \|\psi(t, \theta_{-t}\omega, \mathbf{u}_o)\|^2 &\leq e^{-ct} \|\mathbf{u}_o\|^2 + 2(2n+1) \frac{M_\lambda}{m_\mu c} \|\boldsymbol{\beta}\|^2 \\ &\quad + \frac{M_\mu M_\gamma}{m_\mu^2} \int_{-t}^0 \sum_{i \in \mathbb{Z}} g_i^2(\theta_s \omega) e^{-c(t-s)} ds, \end{aligned}$$

where c is a positive constant depend on $M_\mu, m_\mu, M_\lambda, m_\lambda, M_\gamma, n$, and α in Assumption 10. Due to Assumption 13, $\int_{-t}^0 \sum_{i \in \mathbb{Z}} g_i^2(\theta_s \omega) e^{-c(t-s)} ds$ is a tempered random variable and thus the random dynamical system $\{\psi(t, \omega)\}_{t \geq 0, \omega \in \Omega}$ possesses a random tempered absorbing set

$$\Lambda(\omega) = \left\{ \mathbf{u} \in \ell^2 : \|\mathbf{u}\| \leq \left(2(2n+1) \frac{M_\lambda}{m_\mu c} \|\boldsymbol{\beta}\|^2 + \frac{M_\mu M_\gamma}{m_\mu^2} \mathcal{J}(\theta, \omega) \right)^{1/2} \right\},$$

where

$$\mathcal{J}(\theta, \omega) = \int_{-t}^0 \sum_{i \in \mathbb{Z}} g_i^2(\theta_s \omega) e^{-c(t-s)} ds.$$

The asymptotic compactness of $\Lambda(\omega)$ under the RDS $\{\psi(t, \omega)\}_{t \geq 0, \omega \in \Omega}$ can also be shown by using the cut-off function ξ_m and computations analog to those for the LDS (13), and the existence of a random attractor then follows directly from Proposition 3.

5 Neural field lattice models

Neural field models are often represented as evolution equations generated as continuum limits of computational models of neural field theory. They are tissue level models that describe the spatio-temporal evolution of coarse grained variables such as synaptic or firing rate activity in populations of neurons. See Coombes et al. [25] and the literature therein. A particularly influential model is that proposed by S. Amari in [4] (see also Chapter 3 of Coombes *et al.* [25] by Amari):

$$\partial_t u(t, x) = -u(t, x) + \int_{\Omega} K(x-y) \mathfrak{h}(u(t, y) - \Theta) dy, \quad x \in \Omega \subset \mathbb{R}, \quad (22)$$

where $\Theta > 0$ is a given threshold and $\mathfrak{h}(x) = 1$ for $x \geq 0$ and $\mathfrak{h}(x) = 0$ for $x < 0$ is the Heaviside function as defined in (10).

The continuum neural models may lose their validity in capturing detailed dynamics at discrete sites when the discrete structures of neural systems become dominant. Lattice models can then used to describe dynamics at each site of the neural field (see, e.g., [30, 37]. In particular, Han & Kloeden [37] introduced and investigated the following lattice version of the Amari model with time-dependent external

forcing:

$$\frac{d}{dt} u_i(t) = f_i(u_i(t)) + \sum_{j \in \mathbb{Z}^d} k_{i,j} h(u_j(t) - \Theta) + g_i(t), \quad i \in \mathbb{Z}^d. \quad (23)$$

The infinite dimensional matrix $(k_{i,j})_{i,j \in \mathbb{Z}^d}$ in (23) is the discrete counterpart of the kernel function K in (22) and the term $k_{i,j} h(u_j(t) - \Theta)$ describes the nonlocal interactions between the i th and j th neurons. More precisely, the membrane potential of the i th neuron is affected by those neurons with membrane potential above a certain threshold Θ . The matrix $(k_{i,j})_{i,j \in \mathbb{Z}^d}$ is assumed to satisfy

Assumption 14 $k_{i,j} \geq 0$, and $\sum_{j \in \mathbb{Z}^d} k_{i,j} \leq \kappa$ for all $i, j \in \mathbb{Z}^d$ for some $\kappa > 0$,

which essentially puts a constraint on the aggregate structure of the interactions among neurons.

The main difficulty to analyze the above LDS lies in the discontinuity introduced by the heaviside function h . One way to handle such discontinuity is to replace the Heaviside function by a set-valued mapping χ defined on \mathbb{R} :

$$\chi(s) = \begin{cases} \{0\}, & s < 0, \\ [0, 1], & s = 0, \\ \{1\}, & s > 0, \end{cases} \quad s \in \mathbb{R}. \quad (24)$$

Then the lattice system (23) can be reformulated as the lattice differential inclusion

$$\frac{du_i(t)}{dt} \in f_i(u_i(t)) + \sum_{j \in \mathbb{Z}^d} k_{i,j} \chi(u_j(t) - \Theta) + g_i(t). \quad (25)$$

Another way to handle the discontinuity is to approximate the Heaviside function by a simplifying sigmoidal function such as

$$\sigma_\varepsilon(s) = \frac{1}{1 + e^{-s/\varepsilon}}, \quad s \in \mathbb{R}, \quad 0 < \varepsilon < 1. \quad (26)$$

This avoids the need to introduce a differential inclusion as above.

5.1 Neural field lattice inclusion model

In this subsection we summarize the analysis to obtain an attractor for the lattice differential inclusion (25) given in [37]. To include a wider range of solutions, we study the system (25) in the weighted space of bi-infinite sequences ℓ_ρ^2 as defined in Section 1 with the weights ρ_i 's satisfying Assumption 1.

For each $i \in \mathbb{Z}^d$ the function f_i is assumed to satisfy

Assumption 15 $f_i : \mathbb{R} \rightarrow \mathbb{R}$ is continuously differentiable with weighted equi-locally bounded derivatives, i.e., there exists a non-decreasing function $L(\cdot) \in \mathcal{C}(\mathbb{R}^+, \mathbb{R}^+)$ such that

$$\max_{s \in [-r, r]} |f'_i(s)| \leq L(\rho_i r), \quad \forall r \in \mathbb{R}^+, i \in \mathbb{Z}^d;$$

Assumption 16 $f_i(0) = 0$ and there exist constants $\alpha > 0$ and $\beta := (\beta_i)_{i \in \mathbb{Z}^d} \in \ell^2_\rho$ such that

$$s f_i(s) \leq -\alpha |s|^2 + \beta_i^2, \quad \forall s \in \mathbb{R}, \quad \forall i \in \mathbb{Z}^d.$$

In addition, the time-dependent forcing term $g_i(t)$ is assumed to satisfy:

Assumption 17 $\mathbf{g}(\cdot) := (g_i(\cdot))_{i \in \mathbb{Z}^d} \in \mathcal{C}_{\mathcal{B}}(\mathbb{R}, \ell^2_\rho)$ and $\bar{g}(\cdot) \in \mathcal{L}^1_{loc}(\mathbb{R}) \cap \mathcal{L}^2(\mathbb{R}, \ell^2_\rho)$, where $\bar{g}(t) := \sup_{i \in \mathbb{Z}^d} |g_i(t)|$.

For $\mathbf{u} \in \ell^2_\rho$ define the reaction operator F by $F(\mathbf{u}) = (f_i(u_i))_{i \in \mathbb{Z}^d}$. Then under Assumption 15, F maps ℓ^2_ρ to ℓ^2_ρ and is locally Lipschitz with

$$\|F(\mathbf{u}) - F(\mathbf{v})\|_\rho \leq L(\sqrt{\rho_\Sigma}(\|\mathbf{u}\|_\rho + \|\mathbf{v}\|_\rho))\|\mathbf{u} - \mathbf{v}\|_\rho.$$

Define the interconnection term as the the set-valued operator $\mathfrak{H}(\mathbf{u}) := (\mathfrak{H}_i(\mathbf{u}))_{i \in \mathbb{Z}^d}$ for every $\mathbf{u} = (u_i)_{i \in \mathbb{Z}^d} \in \ell^2_\rho$ given componentwise by

$$\mathfrak{H}_i(\mathbf{u}) = \sum_{j \in \mathbb{Z}^d} k_{i,j} \chi(u_j - \Theta),$$

where χ is the set-valued mapping defined in (24). Then by Assumption 1 and 14 the operator \mathfrak{H} maps an element in ℓ^2_ρ into a set in ℓ^2_ρ .

The lattice differential inclusion (25) can be rewritten as a differential inclusion on ℓ^2_ρ as

$$\dot{\mathbf{u}}(t) \in \mathfrak{G}(\mathbf{u}(t), t) := F(\mathbf{u}(t)) + \mathfrak{H}(\mathbf{u}(t)) + \mathbf{g}(t). \tag{27}$$

Solutions of the above differential inclusion is defined componentwise as follows.

Definition 5. An absolutely continuous function $\mathbf{u}(t) = (u_i(t))_{i \in \mathbb{Z}^d} : [t_0, t_0 + T) \rightarrow \ell^2_\rho$ is called a solution to the differential inclusion (27) if

$$\dot{u}_i(t) \in f_i(u_i(t)) + \mathfrak{H}_i(\mathbf{u}(t)) + g_i(t), \quad \forall i \in \mathbb{Z}^d, \text{ a.e.}$$

Existence theorems in the literature for an inclusion like (27) require the set-valued mapping \mathfrak{H} to be upper semicontinuous in the space ℓ^2_ρ .

It is quite easy to show that the components \mathfrak{H}_i are upper semicontinuous on ℓ^2_ρ . However, the weighted norm of the space ℓ^2_ρ makes it difficult to extend this result to the full mapping. Instead Han & Kloeden [37] constructed a solution as the limit of solutions of approximating systems, summarized below.

Approximation of the set-valued operator

Let $\mathbb{Z}_N^d := \{i = (i_1, \dots, i_d) \in \mathbb{Z}^d : |i_1|, \dots, |i_d| \leq N\}$ and define the truncated set-valued operator

$$\mathfrak{H}_i^N(\mathbf{u}) = \sum_{j \in \mathbb{Z}_N^d} k_{i,j} \chi(u_j - \Theta), \quad \mathbf{u} = (u_i)_{i \in \mathbb{Z}^d} \in \ell_\rho^2.$$

Lemma 3. *For every $i \in \mathbb{Z}^d$, the set-valued mapping $\mathbf{u} \mapsto \mathfrak{H}_i^N(\mathbf{u})$ is upper semi continuous from ℓ_ρ^2 into the nonempty compact convex subsets of \mathbb{R}^1 , i.e.,*

$$\text{dist}_{\mathbb{R}^1}(\mathfrak{H}_i^N(\mathbf{u}^n), \mathfrak{H}_i^N(\hat{\mathbf{u}})) \rightarrow 0 \text{ as } \mathbf{u}^n \rightarrow \hat{\mathbf{u}} \text{ in } \ell_\rho^2.$$

The proof uses the inequality for nonempty compact subsets of \mathbb{R}^d

$$\text{dist}_{\mathbb{R}^d}(A_1 + B_1, A_2 + B_2) \leq \text{dist}_{\mathbb{R}^d}(A_1, A_2) + \text{dist}_{\mathbb{R}^d}(B_1, B_2)$$

and the fact that \mathfrak{H}_i^N is the finite sum of terms involving the upper semi continuous set-valued mapping $s \mapsto \chi(s - \Theta)$.

Lemma 4. *For each $i \in \mathbb{Z}^d$ and every $\varepsilon > 0$ there exists $\mathfrak{N}(\varepsilon)$ such that*

$$\text{dist}_{\mathbb{R}^1}(\mathfrak{H}_i(\mathbf{u}), \mathfrak{H}_i^N(\mathbf{u})) \leq \varepsilon \text{ for all } N \geq \mathfrak{N}(\varepsilon, i), \mathbf{u} \in \ell_\rho^2.$$

To prove this write $\mathfrak{H}_i(\mathbf{u}) := \mathfrak{H}_i^N(\mathbf{u}) + \mathfrak{E}_i^N(\mathbf{u})$, where $\mathfrak{E}_i^N(\mathbf{u}) := \sum_{j \in \mathbb{Z}^d \setminus \mathbb{Z}_N^d} k_{i,j} \chi(u_j - \Theta)$. Then for each $i \in \mathbb{Z}^d$ and all $\mathbf{u} \in \ell_\rho^2$

$$\begin{aligned} \text{dist}_{\mathbb{R}^1}(\mathfrak{H}_i(\mathbf{u}), \mathfrak{H}_i^N(\mathbf{u})) &= \text{dist}_{\mathbb{R}^1}(\mathfrak{H}_i^N(\mathbf{u}) + \mathfrak{E}_i^N(\mathbf{u}), \mathfrak{H}_i^N(\mathbf{u}) + \{0\}) \\ &\leq \text{dist}_{\mathbb{R}^1}(\mathfrak{E}_i^N(\mathbf{u}), \{0\}) \leq \sum_{j \in \mathbb{Z}^d \setminus \mathbb{Z}_N^d} k_{i,j} \leq \varepsilon \quad \forall N \geq \mathfrak{N}(\varepsilon), \end{aligned}$$

because $\|\chi(u_j - \Theta)\| \leq 1$ for all $\mathbf{u} \in \ell_\rho^2$.

Finite dimensional lattice inclusion

For each $i \in \mathbb{Z}_N^d$ and $\mathbf{u}^N(t) = (u_i^N(t))_{i \in \mathbb{Z}_N^d} \in \mathbb{R}^{(2N+1)^d}$ consider the finite dimensional lattice inclusion

$$\frac{d}{dt} u_i^N(t) \in \mathfrak{G}_i^N(\mathbf{u}^N(t), t) := f_i(u_i^N(t)) + \mathfrak{H}_i^N(\mathbf{u}^N(t)) + g_i(t).$$

The set-valued mapping $\mathfrak{G}_i^N(\mathbf{u}^N, t)$ is nonempty, compact, convex valued as well as upper semicontinuous in \mathbf{u}^N and measurable in t . Moreover, it satisfies a bounded growth condition. Hence by standard existence theorems for finite dimensional in-

clusions, e.g., see Aubin & Cellina [6], there exists a solution

$$\mathbf{u}^N(t; t_0, \mathbf{u}_o^N) = (\mathbf{u}_i^N(t; t_0, \mathbf{u}_o^N))_{i \in \mathbb{Z}_N^d}.$$

This implies that for each $i \in \mathbb{Z}_N^d$ there exists a selection $\sigma_i^N(t) \in \mathfrak{H}_i^N(\mathbf{u}^N(t; t_0, \mathbf{u}_o^N))$, a.e., that is such that

$$\frac{d}{dt} u_i^N(t; t_0, \mathbf{u}_o^N) = f_i(u_i^N(t; t_0, \mathbf{u}_o^N)) + \sigma_i^N(t) + g_i(t), \quad i \in \mathbb{Z}_N^d, \text{ a.e..}$$

Componentwise convergent subsequence

First, extend the solution $\mathbf{u}^N(t; t_0, \mathbf{u}_o^N) = (u_i^N(t; t_0, \mathbf{u}_o^N))_{i \in \mathbb{Z}_N^d}$ to $\mathbf{v}^N(t) = (v_i^N(t))_{i \in \mathbb{Z}^d}$ in ℓ_p^2 with zero elements and modify σ_i^N, g_i similarly. Then $\mathbf{v}^N(t)$ satisfies the infinite dimensional lattice ODE, a.e.,

$$\frac{d}{dt} v_i^N(t) = f_i(v_i^N(t)) + \tilde{\sigma}_i^N(t) + g_i^N(t), \quad i \in \mathbb{Z}^d.$$

It follows from the properties of the coefficients functions that

$$|v_i^N(t)| \leq \mu_{i,T}, \quad \left| \frac{d}{dt} v_i^N(t) \right| \leq \tilde{\mu}_{i,T} \quad \text{for all } t \in [t_0, t_0 + T], N \in \mathbb{N}, i \in \mathbb{Z}_N^d,$$

i.e., $\{v_i^N(\cdot)\}_{N \in \mathbb{N}}$ is uniformly bounded and equi-Lipschitz continuous on $[t_0, t_0 + T]$ for each $i \in \mathbb{Z}^d$. Hence the Ascoli-Arzelà Theorem for each $i \in \mathbb{Z}^d$, there is a $v_i^*(\cdot) \in \mathcal{C}([t_0, t_0 + T], \mathbb{R}_+)$ and a convergent subsequence $\{v_i^{N_n}(\cdot)\}_{n \in \mathbb{N}}$ and such that

$$v_i^{N_n}(\cdot) \rightarrow v_i^*(\cdot) \text{ in } \mathcal{C}([t_0, t_0 + T], \mathbb{R}_+) \text{ and } \frac{d}{dt} v_i^{N_n}(\cdot) \rightarrow \frac{d}{dt} v_i^*(\cdot) \text{ in } \mathcal{L}^1([t_0, t_0 + T], \mathbb{R}).$$

The limit function $v_i^*(\cdot)$ shares the equi-Lipschitz continuity of the subsequence $\{v_i^{N_n}(\cdot)\}_{n \in \mathbb{N}}$ and hence is absolutely continuous on $[t_0, t_0 + T]$.

The argument above can be strengthened to obtain a common diagonal subsequence that converges for all $i \in \mathbb{Z}^d$.

Convergent subsequence in ℓ_p^2

By the dissipativity Assumption 16 and more work it can be shown that for some $C_T > 0$

$$\|\mathbf{v}^N(t)\|_\rho \leq C_T, \quad \left\| \frac{d}{dt} \mathbf{v}^N(t) \right\|_\rho^2 \leq C_T \quad \forall t \in [t_0, t_0 + T], N \in \mathbb{N}.$$

Hence by Ascoli-Arzelà Theorem in $\mathcal{C}([t_0, t_0 + T], \ell_\rho^2)$ there exists a $\hat{\mathbf{v}}(\cdot) \in \mathcal{C}([t_0, t_0 + T], \ell_\rho^2)$ and a convergent subsequence $\{\mathbf{v}^{N_n}(\cdot)\}_{n \in \mathbb{N}}$ such that

$$\sup_{t \in [t_0, t_0 + T]} \|\mathbf{v}^{N_n}(t) - \hat{\mathbf{v}}(t)\|_\rho \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Equivalence of limit points

It can be assumed that the two convergent subsequences above for the component-wise limit are the same. Since $\mathbf{v}^{N_n}(t) \rightarrow \hat{\mathbf{v}}(t)$ in ℓ_ρ^2 , $\forall \varepsilon > 0 \exists N(\varepsilon)$ such that

$$\|\mathbf{v}^{N_n}(t) - \hat{\mathbf{v}}(t)\|_\rho^2 = \sum_{\mathbf{i} \in \mathbb{Z}^d} \rho_{\mathbf{i}} |v_{\mathbf{i}}^{N_n}(t) - \hat{v}_{\mathbf{i}}(t)|^2 < \varepsilon^2, \quad n \geq N(\varepsilon).$$

It then follows that

$$|v_{\mathbf{i}}^{N_n}(t) - \hat{v}_{\mathbf{i}}(t)| < \varepsilon / \sqrt{\rho_{\mathbf{i}}}, \quad n \geq N(\varepsilon), \mathbf{i} \in \mathbb{Z}^d.$$

Thus, for every fixed $\mathbf{i} \in \mathbb{Z}^d$,

$$|\hat{v}_{\mathbf{i}}(t) - v_{\mathbf{i}}^*(t)| \leq |v_{\mathbf{i}}^{N_n}(t) - v_{\mathbf{i}}^*(t)| + |v_{\mathbf{i}}^{N_n}(t) - \hat{v}_{\mathbf{i}}(t)| \leq \varepsilon / \sqrt{\rho_{\mathbf{i}}} + \varepsilon.$$

Thus $\hat{v}_{\mathbf{i}}(t) = v_{\mathbf{i}}^*(t)$ for every $\mathbf{i} \in \mathbb{Z}^d$ and $t \in [t_0, t_0 + T]$.

The limit as solution of the lattice inclusion

Rearranging the ODE for the convergent subsequence $\{\mathbf{v}^{N_n}(\cdot)\}_{n \in \mathbb{N}}$ gives

$$\sigma_{\mathbf{i}}^{N_n}(t) = \frac{d}{dt} v_{\mathbf{i}}^{N_n}(t) - f_{\mathbf{i}}(v_{\mathbf{i}}^{N_n}(t)) - g_{\mathbf{i}}(t), \quad \mathbf{i} \in \mathbb{Z}^d, \text{ a.e.},$$

so with the limits $v_{\mathbf{i}}^*(\cdot)$ constructed above define

$$\sigma_{\mathbf{i}}^*(t) := \frac{d}{dt} v_{\mathbf{i}}^*(t) - f_{\mathbf{i}}(v_{\mathbf{i}}^*(t)) - g_{\mathbf{i}}(t), \quad \mathbf{i} \in \mathbb{Z}^d, \text{ a.e.} \quad (28)$$

The terms on the right side of the above equation converge in $\mathcal{L}^1([t_0, t_0 + T], \mathbb{R})$ for each $\mathbf{i} \in \mathbb{Z}^d$. Hence $\sigma_{\mathbf{i}}^{N_n}(\cdot) \rightarrow \sigma_{\mathbf{i}}^*(\cdot)$ in $\mathcal{L}^1([t_0, t_0 + T], \mathbb{R})$ as $n \rightarrow \infty$ for each $\mathbf{i} \in \mathbb{Z}^d$.

It remains to show that $\sigma_{\mathbf{i}}^*(t) \in \mathfrak{H}_{\mathbf{i}}(\mathbf{v}^*(t))$ each $\mathbf{i} \in \mathbb{Z}^d$. In fact, for each $N \in \mathbb{N}$

$$\begin{aligned} \text{dist}_{\mathbb{R}}(\sigma_{\mathbf{i}}^*(t), \mathfrak{H}_{\mathbf{i}}(\mathbf{v}^*(t))) &\leq |\sigma_{\mathbf{i}}^*(t) - \sigma_{\mathbf{i}}^{N_n}(t)| + \text{dist}_{\mathbb{R}}(\sigma_{\mathbf{i}}^{N_n}(t), \mathfrak{H}_{\mathbf{i}}^N(\mathbf{v}^{N_n}(t))) \\ &\quad + \text{dist}_{\mathbb{R}}(\mathfrak{H}_{\mathbf{i}}^N(\mathbf{v}^{N_n}(t)), \mathfrak{H}_{\mathbf{i}}^N(\mathbf{v}^*(t))) + \text{dist}_{\mathbb{R}}(\mathfrak{H}_{\mathbf{i}}^N(\mathbf{v}^*(t)), \mathfrak{H}_{\mathbf{i}}(\mathbf{v}^*(t))). \end{aligned}$$

Estimating the integral of each term from t_0 to $t_0 + T$ then gives

$$\int_{t_0}^{t_0+T} \text{dist}_{\mathbb{R}}(\sigma_i^*(t), \mathfrak{H}_i(\mathbf{v}^*(t))) dt = 0,$$

from which it follows that $\sigma_i^*(t) \in \mathfrak{H}_i(\mathbf{v}^*(t))$ for $t \in [t_0, t_0 + T]$ a.e.. Finally, the equation (28) for σ_i^* can be rewritten as

$$\frac{d}{dt} v_i^*(t) = f_i(v_i^*(t)) + \sigma_i^*(t) + g_i(t), \quad i \in \mathbb{Z}^d, \text{ a.e..}$$

Since $\sigma_i^*(t) \in \mathfrak{H}_i(\mathbf{v}^*(t))$ this implies that $\mathbf{v}^*(t) = (v_i^*(t))_{i \in \mathbb{Z}^d}$ is a solution of the lattice differential inclusion (25). This completes the existence proof.

Remark 1. The proof is rather long and indirect. However, it generalizes without difficulty to include delay and random terms.

Nonautonomous set-valued dynamical system

The attainability set for the lattice inclusion

$$\varphi(t, t_0, \mathbf{u}_o) := \left\{ \mathbf{v} \in \ell_\rho^2 : \exists \text{ a solution } \mathbf{u}(\cdot; t_0, \mathbf{u}_o) \text{ with } \mathbf{u}(t_0; t_0, \mathbf{u}_o) = \mathbf{u}_o \text{ such that } \mathbf{v} = \mathbf{u}(t; t_0, \mathbf{u}_o) \right\}$$

generates a two-parameter set-valued semi-group or nonautonomous set-valued dynamical system. Under the above Assumptions $\varphi(t, t_0, \mathbf{u}_o)$ is a nonempty compact subset of ℓ_ρ^2 for any $t > t_0$ and $\mathbf{u}_o = (u_{i,o})_{i \in \mathbb{Z}^d} \in \ell_\rho$. Moreover, the set-valued mapping $\mathbf{u}_o \mapsto \varphi(t, t_0, \mathbf{u}_o)$ is upper semi continuous in \mathbf{u}_o in ℓ_ρ for any $t \geq t_0$.

It follows from the dissipativity Assumption 16 that the set-valued dynamical system $\varphi(t, t_0, \mathbf{u}_o)$ has a nonautonomous pullback absorbing family of closed and bounded component sets

$$\Lambda(t) := \left\{ \mathbf{u} \in \ell_\rho^2 : \|\mathbf{u}\|_\rho^2 \leq R(t) \right\}, \quad \forall t \in \mathbb{R},$$

where

$$R(t) := \frac{2}{\alpha} \left(\|\beta\|_\rho^2 + \frac{\kappa^2}{\alpha} \rho_\Sigma + \rho_\Sigma \int_{-\infty}^t \bar{g}^2(s) e^{-\alpha(t-s)} ds \right) + 1.$$

These sets $\Lambda(t)$, $t \in \mathbb{R}$, are positively invariant, i.e., $\mathcal{S}(t, t_0, \Lambda(t_0)) \subset \Lambda(t)$ for all $t \geq t_0$.

Recall [53] that a pullback attractor $\mathcal{A} = \{A(t)\}_{t \in \mathbb{R}}$ for \mathcal{S} consists of nonempty compact subsets $A(t)$ of ℓ_ρ^2 which are invariant, i.e., $\varphi(t, t_0, A(t_0)) = A(t)$ for all $t \geq t_0$, and pullback attract the absorbing family, i.e.,

$$\lim_{s \rightarrow \infty} \text{dist}_{\ell_\rho^2}(\varphi(t, t-s, \Lambda(t-s)), A(t)) = 0.$$

The asymptotic tails and asymptotic compactness argument of Bates, Lu & Wang [9] can be adapted to show that the set-valued dynamical system $\varphi(t, t_0, \mathbf{u}_o)$ is asymptotically upper semi compact. From this it follows that the set-valued dynamical system $\varphi(t, t_0, \mathbf{u}_o)$ systems generated by the neural lattice model (25) possesses a unique *pullback attractor* $\mathcal{A} = \{A(t)\}_{t \in \mathbb{R}}$ with components given by

$$A(t) = \bigcap_{s \geq 0} \overline{\bigcup_{t \geq t_0 + s} \varphi(t, t_0, \Lambda(t_0))}.$$

Forward omega limit sets

Pullback attractors involve information about the dynamics of the system in the *past*. They need not be asymptotically stable. Nonautonomous omega limit sets involve information about the dynamics in the *future*.

The lattice inclusion system $\varphi(t, t_0, \mathbf{u}_o)$ also has a positively invariant forward absorbing set

$$\Lambda_0 := \left\{ \mathbf{u} \in \ell_\rho^2 : \|\mathbf{u}\|_\rho^2 \leq R_0 \right\}$$

where

$$R_0 := \frac{2}{\alpha} \left(\|\beta\|_\rho^2 + \frac{\kappa^2}{\alpha} \rho_\Sigma + \hat{g} \right) + 1, \quad \hat{g} := \sup_{t \geq 0} e^{-\alpha t} \int_{t_0}^t \|\mathbf{g}(s)\|_\rho^2 e^{\alpha s} ds < \infty.$$

Similarly to the pullback case it can be shown that $\varphi(t, t_0, \mathbf{u}_o)$ is forward asymptotic compact in Λ_0 . Hence for each $t_0 \in \mathbb{R}$ the *nonautonomous omega limit set*

$$\omega_{t_0, \Lambda_0} = \left\{ \mathbf{u} \in \ell_\rho^2 : \exists t_n \rightarrow \infty, \mathbf{u}_n \in \varphi(t_n, t_0, \Lambda_0), \mathbf{u}_n \rightarrow \mathbf{u} \text{ as } n \rightarrow \infty \right\}$$

which is a nonempty and compact subset of Λ_0 . Moreover,

$$\text{dist}_{\ell_\rho^2} \left(\varphi(t_n, t_0, \Lambda_0), \omega_{t_0, \Lambda_0} \right) \rightarrow 0 \text{ as } t \rightarrow \infty.$$

5.2 Neural field lattice model with sigmoidal function

In this subsection we approximate the Heaviside function H in (23) by the sigmoidal function σ_ϵ defined as in (26). Note that this sigmoidal function is globally Lipschitz with the Lipschitz constant $L_\sigma = \frac{1}{\epsilon}$. Replacing H by σ_ϵ and assuming constant external forcing g_i in (23) results in

$$\frac{du_i}{dt} = f_i(u_i) + \sum_{j \in \mathbb{Z}^d} k_{i,j} \sigma_\epsilon(u_j(t) - \Theta) + g_i, \quad i \in \mathbb{Z}^d. \tag{29}$$

The lattice differential equation (29) and lattice differential inclusion (23) can be connected through an inflated lattice differential inclusion

$$\frac{du_i(t)}{dt} \in f_i(u_i(t)) + \mathfrak{H}_i^\varepsilon(\mathbf{u}(t)) + g_i, \quad i \in \mathbb{Z}^d, \quad (30)$$

where

$$\mathfrak{H}_i^\varepsilon(\mathbf{u}) := \sum_{j \in \mathbb{Z}^d} k_{i,j} \chi_\varepsilon(u_j - \Theta) \text{ with } \chi_\varepsilon(s) = \begin{cases} [0, \varepsilon], & s < -b(\varepsilon), \\ [0, 1], & -b(\varepsilon) \leq s \leq b(\varepsilon), \\ [1 - \varepsilon, 1], & s > b(\varepsilon), \end{cases} \quad s \in \mathbb{R}$$

where $b(\varepsilon) > 0$ solves the algebraic equation

$$\sigma_\varepsilon(b(\varepsilon)) = 1 - \varepsilon, \quad \text{i.e.,} \quad \frac{1}{1 + e^{-b(\varepsilon)/\varepsilon}} = 1 - \varepsilon.$$

Under Assumption 15, 16 and

Assumption 18 $\mathbf{g} := (g_i)_{i \in \mathbb{Z}^d} \in \ell_p^2$,

Han, Kloeden & Wang [40] showed that the Heaviside system (23), the inflated system (30) and the sigmoidal system (29), have global attractors \mathfrak{A} , \mathfrak{A}^ε and \mathcal{A}^ε , respectively with

$$\mathcal{A} = \mathcal{A}^0 \subset \mathcal{A}^\varepsilon, \quad \mathfrak{A}^\varepsilon \subset \mathcal{A}^\varepsilon, \quad \forall \varepsilon \in [0, 1]$$

Moreover,

$$\text{dist}_{\ell_p^2}(\mathcal{A}^\varepsilon, \mathcal{A}) \rightarrow 0, \quad \text{and} \quad \text{dist}_{\ell_p^2}(\mathfrak{A}^\varepsilon, \mathcal{A}) \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0.$$

With the consideration that delays are often included in neural field models to account for the transmission time of signals between neurons (see, e.g., [58]), Wang, Kloeden & Yang [67] considered the autonomous neural field lattice system with the sigmoidal function and delays

$$\frac{d}{dt} u_i(t) = f_i(u_i(t)) + \sum_{j \in \mathbb{Z}^d} k_{i,j} \sigma_\varepsilon(u_j(t - \tau_{i,j}) - \Theta) + g_i, \quad i \in \mathbb{Z}^d. \quad (31)$$

The delays $\tau_{i,j} > 0$ are assumed to be uniformly bounded, i.e., satisfy

Assumption 19 *there exists a constant $\tau \in (0, \infty)$ that $0 \leq \tau_{i,j} \leq \tau$ for all $i \in \mathbb{Z}^d$.*

Denote by E_ρ the Banach space $\mathcal{C}([-\tau, 0], \ell_\rho^2)$ of all continuous functions from $[-\tau, 0]$ to ℓ_ρ^2 with the norm $\|\cdot\|_{E_\rho} = \max_{s \in [-\tau, 0]} \|\cdot(s)\|_\rho$. For any $\mathbf{u}(t) = (u_i(t))_{i \in \mathbb{Z}^d} \in \ell_\rho^2$, \mathbf{u}_t represents the segment in E_ρ defined by $\mathbf{u}_t(s) = \mathbf{u}(t + s)$ for $s \in [-\tau, 0]$.

Similar to the previous subsection, let $F(\mathbf{u}) := (f_i(u_i))_{i \in \mathbb{Z}^d}$ and $\mathbf{g} = (g_i)_{i \in \mathbb{Z}^d}$. In addition, for any $\boldsymbol{\eta} = (\eta_i)_{i \in \mathbb{Z}^d} \in E_\rho$ define the interaction operator K_τ by $K_\tau(\boldsymbol{\eta}) =$

$(K_{\tau,i}(\boldsymbol{\eta}))_{i \in \mathbb{Z}^d}$ by

$$K_{\tau,i}(\boldsymbol{\eta}) = \sum_{j \in \mathbb{Z}^d} k_{i,j} \sigma_E(\eta_j(-\tau_{i,j}) - \theta), \quad \forall i \in \mathbb{Z}^d.$$

Then the operator K_τ maps E_ρ to ℓ_ρ^2 . As a result, the lattice delay differential equation (31) can be written as a functional differential equation on ℓ_ρ^2 :

$$\frac{d}{dt} \mathbf{u}(t) = G_\tau(t, \mathbf{u}_t) := F(\mathbf{u}) + K_\tau(\mathbf{u}_t) + \mathbf{g}. \tag{32}$$

To ensure that the operator $K_\tau : E_\rho \rightarrow \ell_\rho^2$ is Lipschitz continuous, assume that

Assumption 20 *there exists a constant $\tilde{\kappa} > 0$ such $\sum_{j \in \mathbb{Z}^d} \frac{k_{i,j}^2}{\rho_j} \leq \tilde{\kappa}$ for each $i \in \mathbb{Z}^d$.*

Let Assumption 1, 15, 16, 18 and 20 hold. Then given any initial data $u_i(s) = \phi_i(s)$ for $s \in [-\tau, 0]$ with $\boldsymbol{\phi}(\cdot) = (\phi_i(\cdot))_{i \in \mathbb{Z}^d} \in E_\rho$ the existence and uniqueness of solutions to the delay differential equation (32) follows directly from a result of Caraballo *et al.* [17]. Moreover, it defines continuous semigroup $\{\mathcal{S}(t)\}_{t \geq 0} : \times E_\rho \rightarrow E_\rho$ by

$$\mathcal{S}(t, \boldsymbol{\phi}) = \mathbf{u}_t(\cdot; \boldsymbol{\phi}), \quad s \in [-\tau, 0]$$

where $\mathbf{u}(t; \boldsymbol{\phi})$ is the unique solution to (32) with $\mathbf{u}(s) = \boldsymbol{\phi}(s)$ for $s \in [-\tau, 0]$

Existence of a global attractor

Using Assumption 16 it is straightforward to derive the estimate

$$\|\mathbf{u}_t\|_{E_\rho}^2 \leq R_1 e^{-\alpha t} \|\boldsymbol{\phi}\|_{E_\rho}^2 + R_2,$$

where

$$R_1 := e^{\alpha \tau}, \quad R_2 := \frac{2}{\alpha} \left(\|\boldsymbol{\beta}\|_\rho^2 + \frac{1}{\alpha} (\rho_\Sigma \kappa^2 + \|\mathbf{g}\|_\rho^2) \right).$$

Therefore the closed and bounded set

$$\Lambda := \left\{ \boldsymbol{\eta} \in E_\rho : \|\boldsymbol{\phi}\|_{E_\rho} \leq \sqrt{1 + R_2} \right\}$$

is absorbing and positive invariant for the semigroup $\{\mathcal{S}(t)\}_{t \geq 0}$.

Finally, the asymptotic compactness of the semigroup $\{\mathcal{S}(t)\}_{t \geq 0}$ can be shown in a similar way to Bates, Lu & Wang [9] with an asymptotic tails estimates. It then follows that the semigroup $\{\mathcal{S}(t)\}_{t \geq 0}$ generated by the delay lattice system (32) has a global attractor \mathcal{A} in E_ρ .

Remark 2. The solutions of the lattice model with sigmoidal function approximate those of the model with the Heaviside function, and thus provide an alternative method of showing their existence [40].

Remark 3. If Assumption 20 guaranteeing uniqueness of solutions does not hold, then the lattice model (31) generates a set-valued semi-dynamical system, which can be shown to have a global attractor using essentially the same proof.

6 Closing remarks

Three types of lattice dynamical systems arising from the biological sciences are investigated; they are lattice reaction-diffusion systems, Hopfield neural lattice systems, and neural field lattice systems. The lattice reaction-diffusion system has the traditional discretized Laplacian operator that models the simplest tri-diagonal interconnection structure. But the delayed recovery brings discontinuity and thus the system has to be formulated as a differential inclusion on Banach spaces. The Hopfield neural lattice systems have finite neighborhood nonlinear interconnection structures, and the neural field lattice systems have global linear interconnection structures. The main tools to study all the systems are the theory of global, non-autonomous, or random attractors.

Though not included in this chapter, lattice systems modeled using the p -Laplacian $\operatorname{div}(|\nabla u|^{p-2}\nabla u)$ or the $p(x)$ -Laplacian $\operatorname{div}(|\nabla u|^{p(x)-2}\nabla u)$, also involve interesting nonlinear connections and are of potential interest in the biological sciences (see, e.g., [54, 60]). The central difference version of the (scalar) p -Laplacian operator is

$$(\Gamma \mathbf{u})_i := |Bu_i|^{p-2}Bu_i - |B^*u_i|^{p-2}B^*u_i, \quad i \in \mathbb{Z} \tag{33}$$

with $p \geq 2$. (The case $p = 2$ reduces to the usual Laplacian case.)

Based on a reaction-diffusion counterpart of (1) with the Laplacian replaced by the p -Laplacian, Gu & Kloeden [31] proposed and investigated non-autonomous p -Laplacian lattice system

$$\frac{du_i(t)}{dt} = v|u_{i+1} - u_i|^{p-2}(u_{i+1} - u_i) - v|u_i - u_{i-1}|^{p-2}(u_i - u_{i-1}) - \lambda u_i - f_i(t, u_i),$$

and established the existence and uniqueness of solutions and the existence of a nonautonomous pullback attractor [53] in ℓ^2 , under similar assumptions to those above. It is still an open problem to extend the results in [31] to the larger space ℓ^p .

The $p(x)$ -Laplacian operator $\operatorname{div}(|\nabla u|^{p(x)-2}\nabla u)$ has been used in the continuum context to model a wide range of nonlinear and state dependent diffusive structures. Partial differential equations with the $p(x)$ -Laplacian on a bounded smooth domain $\Omega \in \mathbb{R}^n$ are studied (see, e.g., [54]) in the Musielak-Orlicz space space

$$L^{p(\cdot)}(\Omega) := \left\{ \mathbf{u} : \Omega \rightarrow \mathbb{R} : \mathbf{u} \text{ is measurable, } \int_{\Omega} |\mathbf{u}(x)|^{p(x)} dx < \infty \right\},$$

with the exponent function $p(\cdot) \in \mathcal{C}(\bar{\Omega})$ satisfying $1 < \min_{x \in \bar{\Omega}} p(x) \leq \max_{x \in \bar{\Omega}} p(x)$.

The discretized version of the scalar $p(x)$ -Laplacian operator gives a generalized version of the operator Γ defined in (33) for variable exponents is given by

$$(\Gamma_{\mathbf{p}}\mathbf{u})_i := |B^* u_i|^{p_i-2} [(B p_i)(B^* u_i) \ln |B^* u_i| + (p_i - 1) B B^* u_i],$$

where the exponent function $p(\cdot)$ has also been discretized to a real valued bi-infinite sequences $\mathbf{p} = (p_i)_{i \in \mathbb{Z}}$, which is assumed to satisfy

$$1 < p^- := \inf_{i \in \mathbb{Z}} p_i \leq p^+ := \sup_{i \in \mathbb{Z}} p_i < \infty.$$

The corresponding discrete Musielak-Orlicz space $\ell_{\mathbf{p}}$ of real valued bi-infinite sequences is given by

$$\ell_{\mathbf{p}} := \left\{ \mathbf{u} = (u_i)_{i \in \mathbb{Z}} : \sum_{i \in \mathbb{Z}} |u_i|^{p_i} < \infty \right\},$$

which is a Banach space with the norm

$$\|\mathbf{u}\|_{\mathbf{p}} := \inf \left\{ \lambda > 0 : \rho \left(\frac{\mathbf{u}}{\lambda} \right) \leq 1 \right\}, \quad \rho(\mathbf{u}) := \sum_{i \in \mathbb{Z}} |u_i|^{p_i}.$$

See Han, Kloeden & Simsen [38]. The investigation of LDS on the space $\ell_{\mathbf{p}}$ is completely open.

Acknowledgements The work was partially supported by NSF of China (Grant No. 11571125) and Simons Foundation (Collaboration Grants for Mathematicians No. 429717).

References

1. Abdallah, A. Y.: Asymptotic behavior of the Klein-Gordon-Schrödinger lattice dynamical systems. *Commun. Pure Appl. Anal.* **5**, 55–69 (2006)
2. Abdallah, A. Y.: Uniform global attractors for first order non-autonomous lattice dynamical systems. *Proc. Amer. Math. Soc.* **138**, 3219–3228 (2010)
3. Afraimovich, V. S., Nekorkin, V. I.: Chaos of traveling waves in a discrete chain of diffusively coupled maps. *Int. J. Bifur. Chaos* **4**, 631–637 (1994)
4. Amari, S. I.: Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybernet.* **27**, 77–87 (1977)
5. Arnold, L.: Random dynamical systems. Springer-Verlag, Berlin (1998)
6. Aubin, J.P., Cellina, A. : Differential Inclusions, Set-Valued Maps and Viability Theory. Springer-Verlag, Berlin (1984)
7. Bates, P. W., Chen, X., Chmaj, A. J. J.: Traveling waves of bistable dynamics on a lattice. *SIAM J. Math. Anal.* **35**, 520–546 (2003)
8. Bates, P. W., Lisei, H., Lu, K.: Attractors for stochastic lattice dynamical systems. *Stochastics and Dynamics* **6**, 1–21 (2006)
9. Bates, P. W., Lu, K., Wang, B.: Attractors for lattice dynamical systems. *Inter. J. Bifur. & Chaos* **11**, 143–153 (2001)

10. Bell, J.: Some threshold results for models of myelinated nerves. *Math. Biosci.* **54**, 181-190 (1981)
11. Bell, J., Cosner, C.: Threshold behaviour and propagation for nonlinear differential-difference systems motivated by modeling myelinated axons. *Quarterly Appl. Math.* **42**, 1–14 (1984)
12. Cahn, J. W., Mallet-Paret, J., Van Vleck, E. S.: Traveling wave solutions for systems of ODEs on a two-dimensional spatial lattice. *SIAM J. Appl. Math.* **59**, 455–493 (1999)
13. Caraballo, T., Han, Xiaoying: Applied nonautonomous and random dynamical systems. Springer Briefs series, Springer–Verlag (2016)
14. Caraballo, T., Lu, K.: Attractors for stochastic lattice dynamical systems with a multiplicative noise. *Front. Math. China* **3**, 317-335 (2008)
15. Caraballo, T., Morillas, F., Valero, J.: Random attractors for stochastic lattice systems with non-Lipschitz nonlinearity. *J. Difference Equ. Appl.* **17**, 161–184 (2011)
16. Caraballo, T., Morillas, F., Valero, J.: Attractors of stochastic lattice dynamical systems with a multiplicative noise and non-Lipschitz nonlinearities. *J. Differential Equations* **253**, 667–693 (2012)
17. Caraballo, T., Morillas, F., Valero, J.: On differential equations with delay in Banach spaces and attractors for retarded lattice dynamical systems. *Discrete Contin. Dyn. Syst.* **34**, 51–77 (2014)
18. Chaplin, M.: Do we underestimate the importance of water in cell biology ?. *Nature Reviews Molecular Cell Biology* **7**, 861–866 (2006)
19. Chen, X., Guo, J.: Existence and asymptotic stability of traveling waves of discrete quasilinear monostable equations. *Journal of Differential Equations* **184** 549 – 569 (2002)
20. Chow, S.-N., Mallet-Paret, J., Shen, W.: Traveling waves in lattice dynamical systems. *J. Diff. Eq.* **149**, 248–291 (1998)
21. Chow, S.-N., Shen, W.: Dynamics in a discrete Nagumo equation: Spatial topological chaos. *SIAM J. Appl. Math.* **55**, 1764–1781 (1995)
22. Chua, L.O., Roska, T.: The CNN paradigm. *IEEE Trans.Circuits Syst.* **40**, 147–156 (1993)
23. Chua, L. O. , Yang, L.: Cellular neural networks: Theory. *IEEE Trans. Circuits Syst.* **35**, 1257–1272 (1988)
24. Chua, L. O. , Yang, L.: Cellular neural networks: Applications. *IEEE Trans. Circuits Syst.* **35**, 1273–1290 (1988)
25. Coombes, S., Graben, P. B., Potthast, R., Wright J. (Editors): *Neural Fields: Theory and Applications*. Springer, Heidelberg (2014)
26. Deimling, K.: *Differential Equations on Banach Spaces*. Springer-Verlag, Heielberg (1977)
27. Dogaru, R., Chua, L. O.: Edge of chaos and local activity domain of FitzHugh-Nagumo equation. *Internat. J. Bifur. Chaos Appl. Sci. Engrg.* **8**, 211–257 (1998)
28. Erneux, T., Nicolis, G.: Propagating waves in discrete bistable reaction diffusion systems. *Physica D* **67**, 237–244 (1993)
29. Fan, X., Wang, Y.: Attractors for a second order nonautonomous lattice dynamical system with nonlinear damping. *Physics Letters A* **365**, 17–27 (2007)
30. Faye, G: Traveling fronts for lattice neural field equations. *Physica D* **378–379**, 20–32 (2018)
31. Gu, A, Kloeden, P. E. : Asymptotic Behavior of a nonautonomous p-Laplacian lattice system. *International Journal of Bifurcation and Chaos* **26**, 1650174 (2016)
32. Guo, J., Wu, C.: Traveling wave front for a two-component lattice dynamical system arising in competition models. *Journal of Differential Equations* **252**, 4357 – 4391 (2012)
33. Han, Xiaoying: Asymptotic dynamics of stochastic lattice differential equations: a review. Continuous and distributed systems. II, *Stud. Syst. Decis. Control* **30** 121–136 (2015)
34. Han, Xiaoying, Kloeden, P. E.: Lattice systems with switching effects and delayed recovery. *J. Differential Eqns.* **261**, 2986–3009 (2016)
35. Han, Xiaoying, Kloeden, P. E.: *Random Ordinary Differential Equations and their Numerical Solution*. Springer Nature, Singapore (2017)
36. Han, Xiaoying, Kloeden, P. E.: Attractors under Discretisation. Springer Briefs series, Springer–Verlag (2017)
37. Han, Xiaoying, Kloeden, P. E.: Asymptotic behaviour of a neural field lattice model with a Heaviside operator. *Physica D* (doi.org/10.1016/j.physd.2018.09.004)

38. Han, Xiaoying, Kloeden, P. E., Simsen, J.: Sequence spaces with variable exponents for lattice models with nonlinear diffusion. *Modern Mathematics and Mechanics – Fundamentals, Problems, Challenges*, Springer-Verlag (2018)
39. Han, Xiaoying, Kloeden, P. E., Sonner, S.: Discretisation of the global attractor of a lattice system. (under review)
40. Han, Xiaoying, Kloeden, P.E., Wang, Xiaoli: Sigmoidal approximation of Heaviside functions in neural lattice models. (preprint)
41. Han, X., Shen, W., Zhou, S.: Random attractors for stochastic lattice dynamical systems in weighted spaces. *J. Differential Equations* **250**, 1235–1266 (2011)
42. Han, X, Kloeden, P. E., Usman, B: Long term behavior of a random Hopfield neural lattice model **18**, 809–824 (2019)
43. Han, X, Kloeden, P. E., Usman, B: Upper semi-continuous convergence of attractors for a Hopfield-type lattice model (preprint)
44. Hopfield, J. J. : Neurons with graded response have collective computational properties like those of two-stage neurons, *Proc. Nat. Acad. Sci. U.S.A* **81**, 3088–3092 (1984)
45. Hsu, C-H., Lin, S-S.: Existence and Multiplicity of Traveling Waves in a Lattice Dynamical System. *Journal of Differential Equations* **164**, 431 – 450 (2000)
46. Huang, J.: The random attractor of stochastic FitzHugh-Nagumo equations in an infinite lattice with white noises. *Phys. D* **233**, 83–94 (2007)
47. Joyner, R. W., Ramza, B. M., Osaka, T., Tan, R. C.: Cellular mechanisms of delayed recovery of excitability in ventricular tissue, *American Journal of Physiology* **260**, 225–233 (1991)
48. Kapral, R.: Discrete models for chemically reacting systems. *J. Math. Chem.* **6**, 113–163 (1991)
49. Karachaliou, N. I., Yannacopoulos, A. N.: Global existence and compact attractors for the discrete nonlinear Schrödinger equation. *J. Differential Equations* **217**, 88–123 (2005)
50. Keener, J. P.: Propagation and its failure in coupled systems of discrete excitable cells. *SIAM J. Appl. Math.* **47**, 556–572 (1987)
51. Keener, J. P.: The effects of discrete gap junction coupling on propagation in myocardium. *J. Theor. Biol.* **148**, 49–82 (1991)
52. Kloeden, P.E., Lorenz, J.: Stable attracting sets in dynamical systems and in their one-step discretizations. *SIAM J. Numer. Analysis* **23**, 986 – 995 (1986)
53. Kloeden, P. E., Rasmussen, M. :Nonautonomous dynamical systems. *Mathematical Surveys and Monographs* **176**, American Mathematical Society, Providence (2011)
54. Kloeden, P.E., Simsen, J.: Pullback attractors for non-autonomous evolution equations with spatially variable exponents. *Commun. Pure & Appl. Anal.* **13**, 2543–2557 (2014)
55. Laplante, J. P., Erneux, T.: Propagating failure in arrays of coupled bistable chemical reactors. *J. Phys. Chem.* **96**, 4931–4934 (1992)
56. Ma, S, Zhao, X.: Existence, uniqueness and stability of travelling waves in a discrete reaction diffusion monostable equation with delay: *Journal of Differential Equations* **217**, 54 – 87 (2005)
57. Mallet-Paret, J.: The Global Structure of Traveling Waves in Spatially Discrete Dynamical Systems. *J. Dyn. Diff. Eqs.* **11** 49–127 (1999)
58. Martsenyuk, V., Klos-Witkowska, A., Sverstiuk, A.: Stability, bifurcation and transition to chaos in a model of immunosensor based on lattice differential equations with delay. *Electronic Journal of Qualitative Theory of Differential Equations* **27** 1–31 (2018)
59. Pankov, A. A., Pflüger, K.: Travelling waves in lattice dynamical systems. *Math. Methods Appl. Sci.* **23**, 1223–1235 (2000)
60. Persson, E., Halle, B: Cell water dynamics on multiple time scales. *Proceedings of the National Academy of Sciences* **105**, 6266–6271 (2008)
61. Rashevsky, N.: *Mathematical Biophysics*. Dover Publications, New York (1960)
62. Scott, A.C. : Analysis of a myelinated nerve model. *Bull. Math. Biophys.* **26**, 247–254 (1964)
63. Shen, W. : Lifted lattices, hyperbolic structures, and topological disorders in coupled map lattices. *SIAM J. Appl. Math.* **56**, 1379–1399 (1996)
64. Shipston, M. J. : Alternative splicing of potassium channels: a dynamic switch of cellular excitability. *Trends in Cell Biology* **11**, 353–358 (2001)

65. Wang, B.: Dynamics of systems on infinite lattices. *J. Differential Equations* **221**, 224–245 (2006)
66. Wang, B.: Asymptotic behavior of non-autonomous lattice systems. *J. Math. Anal. Appl.* **331**, 121–136 (2007)
67. Wang, Xiaoli, Kloeden, P. E., Yang, Meihua: Asymptotic behaviour of a neural field lattice model with delays (under review)
68. Zhao, X., Zhou, S.: Kernel sections for processes and nonautonomous lattice systems. *Discrete Contin. Dyn. Syst. Ser. B* **9**, 763–785 (2008)
69. Zhou, S.: Attractors for second order lattice dynamical systems. *J. Differential Equations* **179**, 605–624 (2002)
70. Zhou, S.: Attractors for lattice systems corresponding to evolution equations. *Nonlinearity* **15**, 1079–1095 (2002)
71. Zhou, S.: Attractors for first order dissipative lattice dynamical systems. *Phys. D* **178**, 51–61 (2003)
72. Zhou, S.: Attractors and approximations for lattice dynamical systems. *J. Differential Equations* **200**, 342–368 (2004)
73. Zinner, B: Existence of traveling wavefront solutions for the discrete Nagumo equation. *J. Diff. Eq.* **96**, 1–27 (1992)



Balancing Prevention and Suppression of Forest Fires with Fuel Management as a Stock

Betsy Heines and Suzanne Lenhart and Charles Sims

Abstract To study the effects of prevention and suppression on the occurrence of large forest fires, we incorporate the stochasticity of the time of a forest fire into our model and corresponding optimal control problem. In our model, the effects of prevention management spending accumulate over time. Our goal is to determine the optimal combination of the prevention management spending rate over time and one-time suppression spending which would maximize the expected value of a forest. By choosing a hazard function for the random variable for the time of fire, we can convert our stochastic problem into a deterministic problem. We illustrate our results numerically using the 2011 Las Conchas Fire example. Overall, our results support the importance of prevention efforts.

1 Introduction

The number of acres being burned in U.S. forests each year is increasing [19]. Fires are larger and more severe, on average, and the cost to suppress and extinguish these large fires is rising [19, 6]. Recent fire suppression and exclusion policies have resulted in dense forests with more ladder fuels [3]. Additionally, many controlled wildland fires have not been allowed, leading to more continuous, dense forests and severe fires [9, 1]. In particular, fire-adapted ecosystems, where low-intensity

Betsy Heines

Name, University of Tennessee, Department of Mathematics, Knoxville, TN 37996 e-mail: betsy.heines@gmail.com

Suzanne Lenhart

University of Tennessee, Department of Mathematics, Knoxville, TN 37996 e-mail: slenhart@utk.edu

Charles Sims

University of Tennessee, Department of Economics and Baker Center for Public Policy, Knoxville, TN 37996 e-mail: cbsims@utk.edu

surface fires were a common occurrence and were regenerative, now experience high-severity, stand-replacing fires where most of the trees are killed [9].

Alternative fuels management methods to control flammability include mechanical, chemical, or biological means [15]. Roughly 67 millions acres of forest need fuels management [30]. Some current fuels management practices include mechanical thinning and prescribed burning[20]. There is evidence for the value of fuels management treatments to reduce fire hazard and the size of large fire events [1, 34]. Currently, fire suppression spending is higher than expenditures on hazardous fuels reduction [11]. Issues of smoke, conservation of species, and lack of societal acceptance negatively affect the implementation of fuels management [33, 31]. More economic analysis concerning the effectiveness of such management strategies is needed [10, 13, 15].

Economic considerations enter into fire management plans in a variety of ways [17]. Mercer *et al.* [16] use a dynamic stochastic programming model. Linear-integer optimization on a standard-response model examined combined features of fuels management alternatives and initial wildfire suppression attack resource deployment [15]. Minas *et al.* [18] used a deterministic integer programming model combining fuel treatment and fire suppression planning.

However, none of these studies consider how trade-offs between fire prevention and suppression are shaped by associated uncertainties. In particular, the timing of fires is unknown, which causes uncertainty in the benefits of fire prevention efforts. In a review paper of fire management plans, Milne *et al.* discussed challenges of including risk and uncertainty in fire management decisions [17]. In a recent paper [8], the authors considered the economic trade-offs between fire prevention management and fire suppression when the time of fire is stochastic, but the past history of prevention spending has no effect on the acres burned and subsequent damages from a fire that ignites today. To make the effects of the prevention efforts more **realistic**, we extend this work to allow the effects of prevention management spending to accumulate over time in a cumulative prevention management stock. The non-timber damages and the acres burned are a function of the cumulative prevention management stock over time.

Reed developed a method for management strategies of a resource vulnerable to random collapse [24, 25, 26, 27, 4]. His method [28] converted a stochastic problem, due to the random time of collapse, into a deterministic optimal control problem (with ordinary differential equations) and was applied to forestry [24, 25], invasive species [5], and infectious diseases [2, 12]. We use Reed's method to consider optimal prevention spending when the time of fire is stochastic and the effects of prevention management spending accumulate in a stock.

In the next section, we formulate the model and the corresponding optimal control problem with management of prevention efforts and suppression spending. Then, we illustrate numerical results for an example motivated by the 2011 Los Conchas Fire in New Mexico. We close with some conclusions and possible extensions.

2 Model and Control Formulation

While including the uncertainty of the time of fire [17], we want to allow for cumulative effects of prevention management efforts. Our goal is to determine strategies to maximize the expected net present value of the forest over a finite time horizon. We start by solving for the optimal *ex post* fire suppression spending at the time of the fire. Using that, we then solve for the optimal *ex ante* fire prevention spending.

The fire event itself is taken to be instantaneous, since the time for doing suppression action is short compared to our underlying time frame. The cumulative prevention management stock exactly at the time of fire will decrease the damages and the number of acres burned. We are concentrating our model to represent large, high-severity fires.

In a forest with \bar{A} acres, let $A(t)$ be the number of unburned acres in a forest at time t , where t is less than τ , the random time when a fire occurs. The non-timber net benefits B per unit time is a function of the number of unburned acres in the forest, $B = B(A(t))$. Non-timber benefits include supporting and cultural ecosystem services provided by the forest. Thus B captures the benefits of unburned acres net of lost ecosystem services from lack of fire. Assume the next fire in the forest occurs at random time τ with $0 < \tau < T$. Before time τ , the present value of the net benefit from the forest is given by

$$\int_0^\tau \left[B(A(t)) - (ah(t) + \frac{\epsilon}{2}h^2(t)) \right] e^{-rt} dt, \quad (1)$$

where $h(t)$ is the prevention management effort rate. Note that in the integrand of the objective functional we include a quadratic cost term. The terms $ah(t) + \frac{\epsilon}{2}h^2(t)$ represent the cost of the prevention efforts.

The number of unburned acres $A(t)$ before τ satisfies:

$$A'(t) = \delta(\bar{A} - A(t)) \text{ with } A(0) = A_0 \leq \bar{A}, \quad (2)$$

where δ represents the regeneration rate of the forest, and is given by:

$$A(t) = \bar{A} - (\bar{A} - A_0)e^{-\delta t}. \quad (3)$$

2.1 Ex Post Fire Suppression

The accumulation of benefits from prevention management h over time is the cumulative prevention management stock $z(t)$, which satisfies this differential equation

$$z'(t) = h(t) - \gamma z(t) \text{ with } z(0) = z_0. \quad (4)$$

The cumulative prevention management stock z increases with prevention management effort rate h and decays at of rate of γ proportional to the current level of z .

Decay in the prevention management stock captures regrowth of vegetation and re-accumulation of fuels that naturally occurs following fuel management. Thus, $z(t)$ reflects cumulative benefits of prevention management spending while also capturing the impermanence of prevention management treatments.

The number of acres destroyed in the fire, K , is represented by:

$$K = K(z(\tau), x(\tau)), \tag{5}$$

where the *ex post* fire suppression expenditures at the time of the fire is $x(\tau)$. We assume K is decreasing with respect to increases in cumulative prevention management stock and suppression spending; i.e. $\frac{\partial K}{\partial z} < 0$ and $\frac{\partial K}{\partial x} < 0$.

Let $\hat{A}(t)$, the number of unburned acres in the forest following a fire at time τ , be expressed as

$$\hat{A}(\tau) = A(\tau) - K(z(\tau), x(\tau)). \tag{6}$$

As in [8], we assume that another large, high-severity fire does not occur in our finite time horizon $[0, T]$. After the fire, the number of unburned acres \hat{A} in the forest increases according to the differential equation

$$\hat{A}'(t) = \delta(\bar{A} - \hat{A}(t)) \text{ with } \hat{A}(\tau) = A(\tau) - K(z(\tau), x(\tau)), \tag{7}$$

and is explicitly given by

$$\hat{A}(t) = \bar{A} - \left(\bar{A} - \left(A(\tau) - K(z(\tau), x(\tau)) \right) \right) e^{-\delta(t-\tau)}. \tag{8}$$

The damages are a function of the number of acres destroyed in the fire:

$$D = D\left(K(z(\tau), x(\tau))\right), \tag{9}$$

which includes impacts to surrounding buildings and roads. We assume that larger fires have more impact: $\frac{\partial D}{\partial K} > 0$. Damages can be decreased by prevention and suppression actions: $\frac{\partial D}{\partial z} < 0$ and $\frac{\partial D}{\partial x} < 0$.

The function describing the flow of benefits before and after the fire is the same, even though we distinguish between unburned acres before the fire and unburned acres after the fire, A and \hat{A} , respectively. The net present value of the forest following a fire is given by:

$$\int_{\tau}^T B(\hat{A}(t))e^{-rt} dt - \left[D\left(K(z(\tau), x(\tau))\right) + x(\tau) \right] e^{-r\tau}, \tag{10}$$

subject to (8) and $x(\tau) \geq 0$. Assuming at most a single fire event in our time horizon, there is no incentive to invest in prevention following a fire.

The value of the forest after the fire, with $e^{-r\tau}$ factored out, is given by

$$JW(\tau, A(\tau), z(\tau), x(\tau)) = \int_{\tau}^T B(\hat{A}(t))e^{-r(t-\tau)}dt - \left[D\left(K(z(\tau), x(\tau)) \right) + x(\tau) \right]. \tag{11}$$

Note that JW is a function of $A(\tau)$ and not $\hat{A}(\tau)$ because \hat{A} is determined by the boundary condition containing $A(\tau)$ and the differential equation (7) with a dependence on K , the number of acres burned. Given a time of fire τ , the optimal *ex post* value of the forest is the solution to

$$\sup_{x(\tau)} \int_{\tau}^T B(\hat{A}(t))e^{-r(t-\tau)}dt - \left[D\left(K(z(\tau), x(\tau)) \right) + x(\tau) \right]$$

subject to $x(\tau) \geq 0$, (12)

$$\text{where } \hat{A}(t) = \bar{A} - \left(\bar{A} - \left(A(\tau) - K(z(\tau), x(\tau)) \right) \right) e^{-\delta(t-\tau)}, \tag{13}$$

with $x(\tau)$ being a real-valued scalar representing suppression spending. With $x^*(\tau)$, the optimal suppression spending, the maximized *ex post* value of the forest for a given τ , $A(\tau)$, and $z(\tau)$ is henceforth denoted by

$$JW^*(\tau, A(\tau), z(\tau)) = JW(\tau, A(\tau), z(\tau), x^*(\tau)). \tag{14}$$

From our assumptions on D and K , the cumulative prevention management stock increases the value of the forest following a fire:

$$\frac{\partial JW^*(\tau, A(\tau), z(\tau))}{\partial z} > 0. \tag{15}$$

Once our functional forms are chosen we explicitly determine $x^*(\tau)$.

2.2 Ex Ante Fire Prevention

If the time of fire $\tau < T$, then the total value of the forest over $[0, T]$ is given by the sum of the net value of the forest before the fire and the net value of the forest after the fire up to time T ,

$$\int_0^{\tau} [B(A(t)) - (ah(t) + \frac{\epsilon}{2}h^2(t))]e^{-rt}dt + \int_{\tau}^T B(\hat{A}(t))e^{-rt}dt - \left[D\left(K(z(\tau), x(\tau)) \right) + x(\tau) \right]e^{-r\tau},$$

where $A(t)$ is given by (3) and $\hat{A}(t)$ is given by (8).

If the time of the first fire is $\tau \geq T$, then we take the time of fire to be $\tau = T$ and the value of the forest would be

$$\int_0^T \left[B(A(t)) - (ah(t) + \frac{\epsilon}{2}h^2(t)) \right] e^{-rt} dt, \tag{16}$$

where $A(t)$ is given by (3). In this case, we recognize that a fire will eventually occur, but because it does not occur within the time horizon $[0, T]$ we do not subtract the instantaneous suppression costs or cost of damages to built structures.

In summary, the value of the forest can be represented by the piecewise function

$$\mathcal{V}(A_0, \tau, h, z) = \begin{cases} \int_0^\tau \left[B(A(t)) - (ah(t) + \frac{\epsilon}{2}h^2(t)) \right] e^{-rt} dt + e^{-r\tau} JW^*(\tau, A(\tau), z(\tau)) & \text{if } \tau < T \\ \int_0^T \left[B(A(t)) - (ah(t) + \frac{\epsilon}{2}h^2(t)) \right] e^{-rt} dt & \text{if } \tau = T, \end{cases} \tag{17}$$

where $A(t)$ is given by (3). Note that \hat{A} is completely contained within JW^* .

When the large fire event will occur is unknown. The time of fire τ is a realization of the mixed-type random variable (RV) \mathcal{T} , which is characterized by the hazard function ψ ,

$$\psi = \lim_{\Delta t \rightarrow 0} \left\{ \frac{Pr(\text{fire in } [t, t + \Delta t] | \text{no fire up to } t)}{\Delta t} \right\}. \tag{18}$$

The hazard function gives the conditional probability that a fire will occur at a time t given that no fire has occurred up to that time and it relates to the ecological concept of a fire return interval [7]. Our hazard function is taken to be a function of the *ex ante* cumulative prevention management stock,

$$\psi = \psi(z(t)). \tag{19}$$

We also assume $\frac{\partial \psi}{\partial z} < 0$.

The survivor function $S(t)$, the probability of the forest surviving to time t with no fire, is:

$$S(t) = e^{-\int_0^t \psi(z(s)) ds}. \tag{20}$$

with $S(0) = 1$. We assume that the integral representing the cumulative hazard, $\int_0^\infty \psi(z(s)) ds$ diverges to positive ∞ and $S(\infty) = 0$. The corresponding cumulative distribution function for \mathcal{T} is

$$F_{\mathcal{T}}(\tau) = \begin{cases} 1 - S(\tau) & \text{if } \tau < T \\ 1 & \text{if } \tau = T. \end{cases} \tag{21}$$

with a possible discontinuity at time T . The probability density function for $\mathcal{T} \in [0, T)$ is

$$f_{\mathcal{F}}(t) = \psi(z(t))S(t). \tag{22}$$

For $\mathcal{F} = T$, we have

$$\begin{aligned} P(\mathcal{F} = T) &= F_{\mathcal{F}}(T) - F_{\mathcal{F}}(T^-) \\ &= 1 - (1 - S(T)) \\ &= S(T). \end{aligned} \tag{23}$$

If $\tau = T$, no costs other those resulting from prevention management efforts are considered. Our goal is to determine the prevention management effort rate $h(t) \geq 0$ which maximizes the net present value of the forest over $[0, T]$ using deterministic optimal control. Using Reed’s techniques, we convert this stochastic problem to deterministic by taking the expectation of (17) with respect to the random variable \mathcal{F} and introducing a state variable to represent cumulative hazard [28].

The expected net present value of the forest over $[0, T]$, is given by

$$\begin{aligned} J(h) &= E_{\mathcal{F}}\{\mathcal{V}(A_0, \tau, h, z)\} \\ &= \int_0^T \left[\int_0^\tau \left[B(A(t)) - ah(t) - \frac{\epsilon}{2}h^2(t) \right] e^{-rt} dt \right. \\ &\quad \left. + JW^*(\tau, A(\tau), z(\tau)) e^{-r\tau} \right] \psi(z(\tau))S(\tau) d\tau \end{aligned} \tag{24}$$

$$+ S(T) \int_0^T \left[B(A(t)) - (ah(t) + \frac{\epsilon}{2}h^2(t)) \right] e^{-rt} dt, \tag{25}$$

which becomes

$$J(h) = \int_0^T \left[B(A(t)) - (ah(t) + \frac{\epsilon}{2}h^2(t)) + \psi(z(t))JW^*(t, A(t), z(t)) \right] S(t)e^{-rt} dt. \tag{26}$$

By introducing a new state variable y to represent cumulative hazard, our problem becomes deterministic. The cumulative hazard y satisfies

$$y'(t) = \psi(z(t)) \text{ with } y(0) = 0, \tag{27}$$

with $y(0) = 0$ coming from $S(0) = 1$. Now we have this relationship:

$$S(t) = e^{-y(t)}, \tag{28}$$

and this allows us to rewrite (26) with our new state variable y .

Our deterministic optimal control problem can be expressed as

$$\sup_{h \in U} \int_0^T \left[B(A(t)) - (ah(t) + \frac{\epsilon}{2}h^2(t)) + \psi(z(t))JW^*(t, A(t), z(t)) \right] e^{-rt-y(t)} dt \tag{29}$$

$$\text{with } y'(t) = \psi(z(t)) \text{ and } y(0) = 0, \quad z'(t) = h(t) - \gamma z(t) \text{ and } z(0) = z_0, \tag{30}$$

where

$$U = \{h : [0, T] \rightarrow [0, M] | h \text{ is Lebesgue measurable}\}, \tag{31}$$

and

$$A(t) = \bar{A} - (\bar{A} - A_0)e^{-\delta t}. \tag{32}$$

Thus, our control problem with stochastic time of fire has been converted to a deterministic optimal control problem.

2.3 Choosing Functional Forms

We now choose explicit functional forms for B , K , D , and ψ . The benefits function B is chosen as:

$$B(A(t)) = B_1A(t), \tag{33}$$

where parameter $B_1 \geq 0$. The number of acres completely burned by the fire, K , is given by

$$K(z, x) = \frac{k}{(k_1 + z)(k_2 + x)}, \tag{34}$$

with parameters $k > 0$ and $k_1, k_2 \geq 1$, where k is related to the size of a fire. The cost of damaged structures is directly proportional to K :

$$D(K(z, x)) = cK(z, x) = \frac{ck}{(k_1 + z)(k_2 + x)}, \tag{35}$$

with parameter $c \geq 0$ as the cost of damages in millions of dollars per thousand acres burned.

Using [28, 24, 2, 5], the hazard function ψ , is chosen as

$$\psi(z(t)) = be^{-\nu z(t)}. \tag{36}$$

The parameter $0 < b < 1$ represents the constant hazard rate when there is no prevention management effort. The constant $\nu > 0$ reflects the effectiveness of $z(t)$ on reducing hazard.

With these functional forms, we optimize the value of the forest after the fire JW . Using the solution to the state differential equation for $\hat{A}(t)$ above, we integrate the

flow of benefits from the time of fire τ to the end of our time horizon T and obtain *ex post* value of the forest

$$\begin{aligned}
 JW(\tau, A(\tau), z(\tau), x(\tau)) &= \frac{B_1 \bar{A}}{r} \left(1 - e^{-r(T-\tau)}\right) - \frac{B_1(\bar{A} - A(\tau))}{\delta + r} \left(1 - e^{-(\delta+r)(T-\tau)}\right) \\
 &\quad - K(z(\tau), x(\tau)) \left[\frac{B_1}{\delta + r} \left(1 - e^{-(\delta+r)(T-\tau)}\right) + c \right] - x(\tau). \tag{37}
 \end{aligned}$$

We maximize $JW(\tau, A(\tau), z(\tau), x(\tau))$ with respect to the suppression cost $x(\tau)$. Scalar optimization gives

$$\begin{cases} x^*(\tau) = 0 & \text{if } \frac{\partial JW}{\partial x(\tau)} < 0 \\ x^*(\tau) \geq 0 & \text{if } \frac{\partial JW}{\partial x(\tau)} = 0. \end{cases} \tag{38}$$

Using K a function of x , we obtain

$$\frac{\partial JW}{\partial x} = \left[\frac{B_1}{\delta + r} \left(1 - e^{-(\delta+r)(T-\tau)}\right) + c \right] \frac{k}{(k_1 + z)(k_2 + x)^2} - 1. \tag{39}$$

Using one case with $\frac{\partial JW}{\partial x(\tau)} = 0$ and $x^*(\tau) \geq 0$, and another case with $\frac{\partial JW}{\partial x(\tau)} < 0$ and $x^*(\tau) = 0$, the optimal suppression spending becomes

$$x^*(\tau, z(\tau)) = \max \left\{ 0, \sqrt{\frac{k}{(k_1 + z(\tau)) \left[\frac{B_1}{\delta + r} \left(1 - e^{-(\delta+r)(T-\tau)}\right) + c \right]} - k_2} \right\}, \tag{40}$$

and the optimal value of the forest following a fire becomes

$$JW^*(\tau, A(\tau), z(\tau)) = JW\left(\tau, A(\tau), z(\tau), x^*(\tau, z(\tau))\right). \tag{41}$$

We note that $\frac{\partial^2 JW}{\partial x^2} \leq 0$ and so the JW^* value is a maximum of JW (37).

We present the optimality system for our new optimal control problem. Let H represent the Hamiltonian with adjoints, $\lambda_1(t)$ and $\lambda_2(t)$, corresponding to the state variables y and z respectively:

$$\begin{aligned}
 H &= \left[B(A(t)) - (ah(t) + \frac{\epsilon}{2}h^2(t)) + \psi(z(t))JW^*(t, A(t), z(t)) \right] e^{-rt-y(t)} \\
 &\quad + \lambda_1(t)\psi(z(t)) + \lambda_2(t)(h(t) - \gamma z(t)). \tag{42}
 \end{aligned}$$

The conditional current-value Hamiltonian is given by $\mathcal{H} = e^{rt+y(t)}H$. Thus,

$$\begin{aligned} \mathcal{H} = & B(A(t)) - (ah(t) + \frac{\epsilon}{2}h^2(t)) + \psi(z(t))JW^*(t, A(t), z(t)) \\ & + \rho_1(t)\psi(z(t)) + \rho_2(t)(h(t) - \gamma z(t)), \end{aligned} \tag{43}$$

and the corresponding conditional current-value adjoint equations by

$$\rho_i(t) = e^{rt+y(t)}\lambda_i(t), \tag{44}$$

for $i = 1, 2$.

The conditional current-value adjoint differential equations are

$$\begin{aligned} \rho'_1(t) = & \left(r + \psi(z(t)) \right) \rho_1(t) + B(A(t)) - (ah(t) + \frac{\epsilon}{2}h^2(t)) \\ & + \psi(z(t))JW^*(t, A(t), z(t)), \end{aligned} \tag{45}$$

and

$$\begin{aligned} \rho'_2(t) = & \left(r + \gamma + \psi(z(t)) \right) \rho_2(t) - \psi(z(t)) \frac{\partial JW^*}{\partial z} \\ & - JW^*(t, A(t), z(t)) \frac{\partial \psi}{\partial z} - \rho_1(t) \frac{\partial \psi}{\partial z}, \end{aligned} \tag{46}$$

with transversality conditions

$$\rho_1(T) = 0 \text{ and } \rho_2(T) = 0. \tag{47}$$

Using

$$\frac{\partial \mathcal{H}}{\partial h} = -a - \epsilon h(t) + \rho_2(t), \tag{48}$$

and the bounds on the control $0 \leq h(t) \leq M$, we have

$$h^*(t) = \min \left\{ M, \max \left\{ 0, \frac{\rho_2(t) - a}{\epsilon} \right\} \right\}. \tag{49}$$

The concavity condition for a maximization is valid:

$$\frac{\partial^2 \mathcal{H}}{\partial h^2} = -\epsilon < 0. \tag{50}$$

The hazard function ψ is nonlinear in z , as is the function $JW^*(t, A(t), z(t))$, which represents the optimal value of the forest following a forest fire. We utilize the fact that Pontryagin’s Maximum Principle (PMP) states that the optimal control maximizes the Hamiltonian with respect to the control h pointwise at each t to numerically determine the optimal control [14]. An iterative method is used, starting with an initial guess for h , which gives x from (40). Then the state y followed by the adjoint ρ are solved numerically. A new control h is obtained by maximizing the

Hamiltonian pointwise at each time step using the MATLAB function ‘fminbnd’. Using an updated h , we compare our current variables with the previous values and continue to iterate until convergence occurs. We justify the use of PMP for our maximization problem since the existence of an optimal control pair is standard for this system.

Table 1: The table below includes the parameter values chosen to reflect the 2011 Las Conchas Fire.

Parameter	Units	Value	Justification
\bar{A}	acres(1000)	1700	size of SFNF, BNM, VCNP
r	/time	0.04	standard discount rate
k	acres(1000) \times \$ ² /time	7000	$k \approx$ size of fire \times suppression \$
k_1	\$ (mil.)/time	1	assumed
k_2	\$ (mil.)/time	1	assumed
δ	/time	0.05	Pipo: 70-250 years to mature
b	————	0.2	high frequency of fires in region
c	\$ (mil.)/ acres(1000)	0.1	114 buildings destroyed, 156,000 acres burned
B_1	\$ (mil.)/time	0.02	calculated from x^* formula
v	————	1	assumed

3 Numerical Results for Las Conchas Fire

Now that we have formulated our optimal control problem and the associated optimality system, we solve it numerically for a specific example using data from the 2011 Las Conchas Fire. A fallen power line started this fire on June 26, 2011, and the fire burned over the summer through parts of Santa Fe National Forest, Bandelier National Monument, and Valles Caldera National Preserve near Los Alamos, New Mexico. The fire was contained at the beginning of August 2011 [22, 32]. Over 150,000 acres burned and over \$40 million were spent on fire suppression [22, 32, 37]. In addition to suppression costs, over 110 structures were damaged [22, 32]. Parameter choices are summarized in Table 1.

The parameter \bar{A} represents the “size of the forest” in units of thousands of acres, which was approximately 1,700 for this event. [21, 36, 38].

The parameter δ represents the regeneration rate of the forest following a fire. We choose δ based on the dominant tree type in the forest, which in the Santa Fe Na-

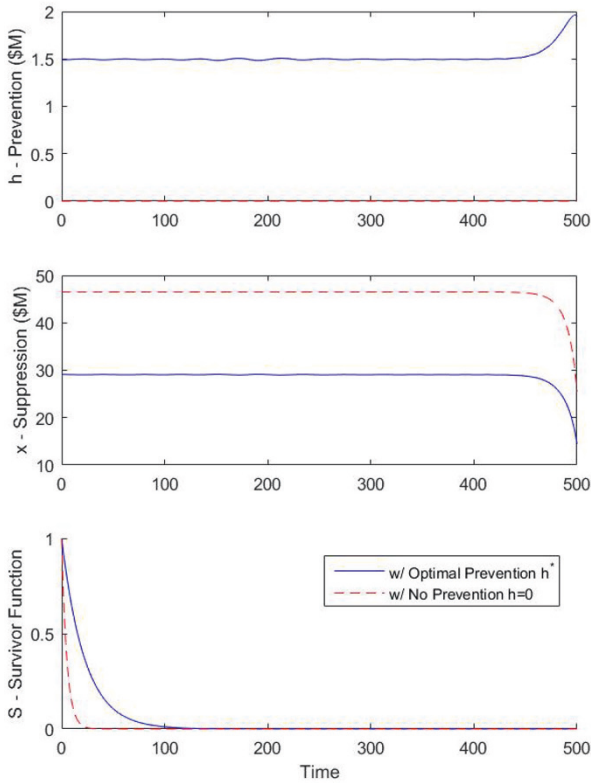


Fig. 1: The plots above contain the h^* , x^* , and S results of our optimal control problem using the Las Conchas Fire parameter set. For comparison, in each plot we include the case with optimal prevention management spending h^* and the case with no prevention spending $h = 0$.

tional Forest is Ponderosa Pine (Pipo) [35]. Assuming that the number of unburned acres was reduced by half, we choose a value for δ so that the number of unburned acres after 100 years has approximately returned to \bar{A} , giving $\delta = 0.05$. The discount rate is chosen to be $r = 0.04$ [29]. The parameter b represents the background fire hazard. To capture the probability that large, high-severity fires happen frequently in the region, we set $b = 0.2$ [39].

The parameters k , k_1 , and k_2 in the function K need to be set to approximate the \$40M spent on suppression and the 157 acres burned. For simplicity we choose $k_1 = k_2 = 1$ and $k = 7,000$.

From the literature, we choose $v = 1$ in the hazard function [25, 2]. The parameter c represents the cost of damages to structures in millions of dollars per thousand

acres burned. Since 114 buildings were damaged [32] our estimate for D is \$17.2 million. Using the number of acres destroyed in the fire, $K = 157$, then $c \approx \frac{D}{K}$ becomes $c = 0.1$.

To decide the parameter B_1 , representing the flow of non-timber benefits, [23]. We use our equation (40) to determine B_1 based on our other parameter choices and the amount of money spent on suppression. Using the amount of suppression spending as approximately optimal for x^* and using $z(\tau) = 0, \tau = 0$, and $T = 500$, we take $B_1 = 0.02$. For the quadratic cost parameter, we choose $\varepsilon = 2$ and $a = 1$ for our illustration.

The parameter M is the upper bound on prevention management spending h . We choose $M = B_1\bar{A} = 34$. That is, we stipulate that prevention management spending rate h is never greater than the flow of benefits when the forest is entirely unburned. We do not vary this parameter because in all cases tested, this upper bound is not reached by h^* and does not even come close to it.

The parameter γ gives the rate of decay for cumulative prevention management stock. We choose a few values to examine that represent a few different scenarios. Choosing values close to zero indicates a slow decline of stock, while larger values for γ indicate a quick decline of the benefits of prevention management efforts. Thus, we solve our optimal control problem using $\gamma = 0.5, 1, 5$ and let our baseline value for the parameter be $\gamma = 1$ when we vary some parameters.

The parameter z_0 is the initial condition for the cumulative prevention management stock. Its value can be used to reflect whether or not, or to what extent, there have been prevention management efforts in an area prior to the application of our optimal control problem. We compare a few values for z_0 , choosing $z_0 = 0, 1, 5$. Setting the initial condition to zero implies that no prevention management efforts have been recently made in the forest. We let $z_0 = 0$ be our baseline value.

For this problem, we consider a time horizon of $T = 5$.

The selection of these parameters is scenario driven. We perform a local sensitivity analysis where we consider several different parameter scenarios by varying one parameter and holding the others constant at their stated baseline values. In particular, we vary z_0 and γ because they directly relate to the cumulative prevention management stock z . We also vary the initial condition for the number of unburned acres A_0 .

First, we consider the results when our optimal control problem is solved at the baseline values for γ and z_0 and we vary the initial condition for the number of unburned acres with $A_0 = 0.5\bar{A}, \bar{A}$. The results can be found in Table 2 and Figure 2. In the case when $A_0 = 0.5\bar{A}$, the expected net present value of the forest for 5 years is \$60.85 M and in the case when $A_0 = \bar{A}$, the expected net present value of the forest is \$130.5 M. Of course, because there are fewer unburned acres A in the forest in the case when $A_0 = 0.5\bar{A}$ and benefits are a function of unburned acres, it is not surprising that the expected net present value of the forest $J(h^*)$ is less in the case where $A_0 = 0.5\bar{A}$. Also contributing to this difference in values for $J(h^*)$ is that, in the case of $A_0 = 0.5\bar{A}$, there is more spending on prevention management than in the case where $A_0 = \bar{A}$. Moreover, the higher level of prevention management spending in the case where $A_0 = 0.5\bar{A}$ leads to a greater accumulation of prevention

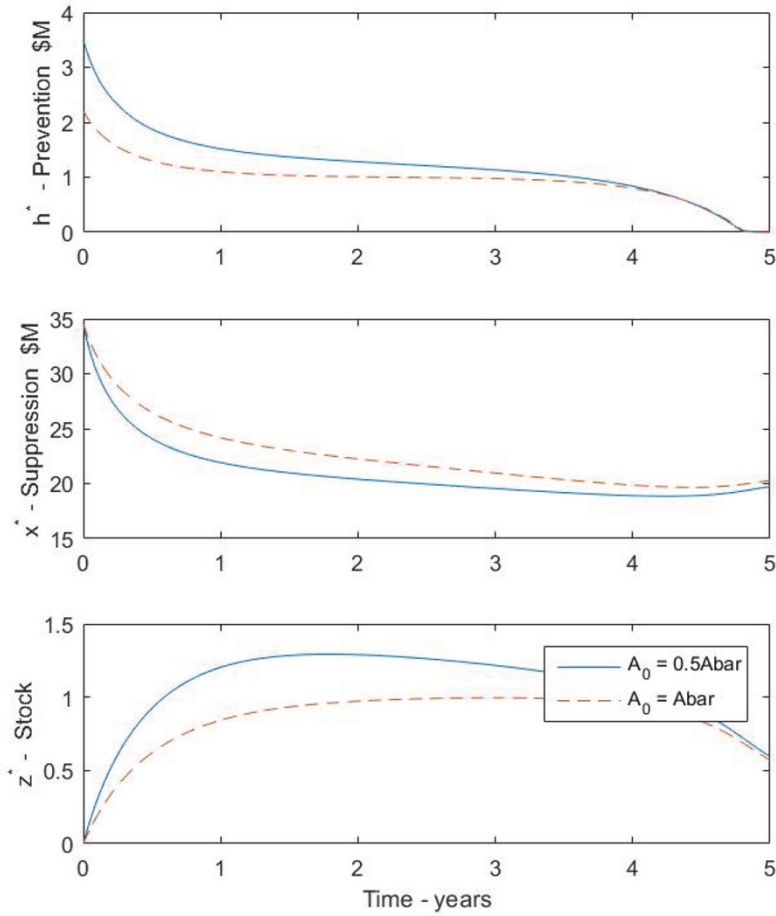


Fig. 2: The plots above show the results of our optimal control problem with two different initial conditions for the number of healthy acres in the forest A_0 . We use $A_0 = 0.5\bar{A}$ and $A_0 = \bar{A}$ and compare prevention management spending h^* , suppression spending x^* , and cumulative prevention management stock z^* .

Table 2: The table above gives the value of the objective functional $J(h^*)$ evaluated at the optimal control for the different parameter scenarios tested.

Parameter	$J(h^*)$ (\$M)
$A_0 = 0.5\bar{A}$	60.85
$A_0 = \bar{A}$	130.50
$\gamma = 0.5$	135.42
$\gamma = 1$	130.50
$\gamma = 5$	118.04
$z_0 = 0$	130.50
$z_0 = 1$	134.69
$z_0 = 5$	140.52

management stock z^* , which explains the lower optimal suppression spending x^* as $\frac{\partial x^*}{\partial z} < 0$.

Next, we consider the results when the decay rate γ is varied, with $A_0 = \bar{A}$ and $z_0 = 0$. See Figure 3 and the corresponding rows in Table 2. As the rate of decay γ of cumulative prevention management stock increases the expected net present value of the forest $J(h^*)$ decreases. In the case where γ is largest, prevention spending h^* is lowest, along with the value of the forest $J(h^*)$.

Let’s examine the solution to the state differential equation for z to try and gain a better understanding of this state variable:

$$z(t) = z_0 e^{-\gamma t} + \int_0^t e^{-\gamma(t-s)} h(s) ds. \tag{51}$$

Let’s suppose, for instance, that prevention management spending is constant with $h = C$, and let’s suppose that $z_0 = 0$. Then with a simple calculation we see that

$$z(t) = \frac{C}{\gamma} (1 - e^{-\gamma t}). \tag{52}$$

This suggests that prevention management spending needs to be relatively high in order for stock to accumulate in a meaningful way. For instance, in the case where $\gamma = 5$, even if $h = 2M$ per year, the cumulative prevention management stock would never rise above 0.4 and hence would not be very effective at reducing suppression costs or hazard. In this case a very high decay rate γ for cumulative stock is “worse” than if the the effects of prevention management spending were instantaneous. Thus, in cases where the rate of decay is very high, the utility of prevention management spending is greatly decreased. As we can see in our quick example with h constant, $\gamma \gg 1$ has a substantial effect on how prevention spending contributes to the stock.

In the large γ case when $\gamma = 5$ (see Figure 3), optimal prevention management spending h^* is approximately constant at \$0.6M for the first 4.5 years of the 5 year

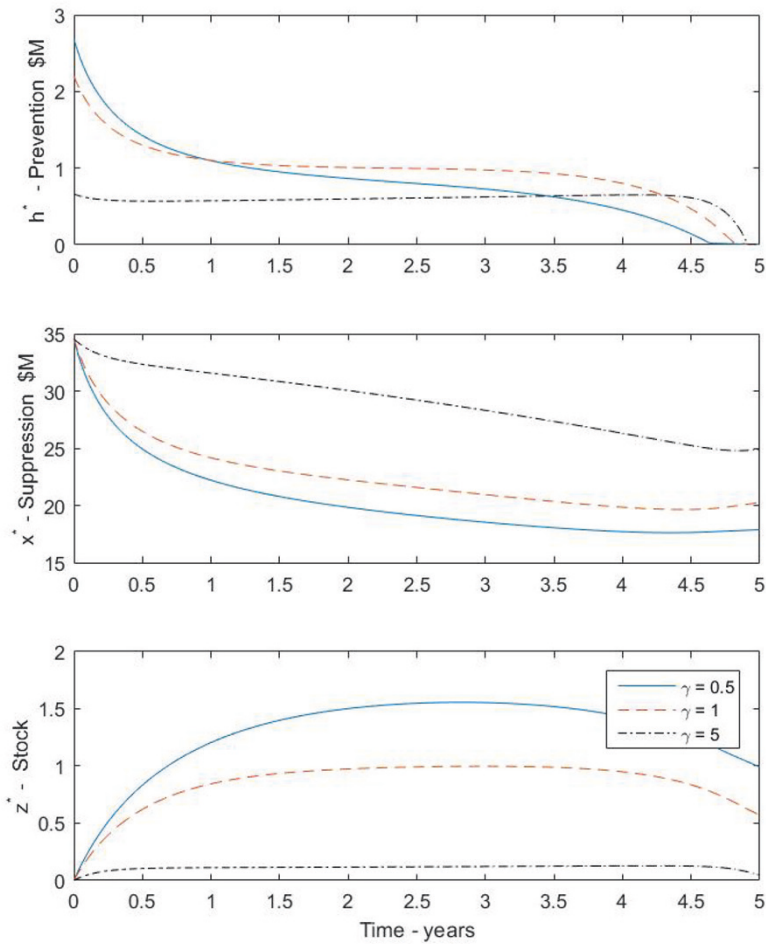


Fig. 3: The plots show the results of our optimal control problem with three different values for the parameter γ , which controls the rate of decay of cumulative prevention management stock z . We use $\gamma = 0.5$, $\gamma = 1$, and $\gamma = 5$ and compare prevention management spending h^* , suppression spending x^* , and cumulative prevention management stock z^* .

time horizon. The cumulative prevention management stock z^* is very low, around \$0.1 M, which is expected due to our previous analysis. In the case when $\gamma = 1$, optimal prevention spending begins near \$2 million and reduces to roughly \$1 million after one year. During this first year prevention management spending is less than in the case when $\gamma = 0.5$. After one year, h^* is larger in the case when $\gamma = 1$. As can be seen for the prevention stock z^* , a smaller rate of decay $\gamma = 0.5$ allows for a quick accumulation of stock, which stays relatively high, even with decreasing levels of optimal prevention management spending. As is seen, different values of γ have a varying effect on prevention management spending, meaning that there is not a strict monotonic relationship between the value of γ and the level of prevention management spending h^* .

Finally, we vary the parameter for the initial condition for cumulative prevention management stock z_0 . We use values $z_0 = 0, 1, 5$ and let $A_0 = \bar{A}$ and $\gamma = 1$. As seen in Figure 4, initially prevention management spending h^* is higher for lower values of initial cumulative prevention management stock z_0 . However, near $t = 3$ these three trajectories begin to coincide. Unsurprisingly, the three different x^* and z^* trajectories all come together near the same time. Thus, it appears that for this given set of parameters, there is an optimal stock level, and despite the value chosen for the initial stock level z_0 , prevention management h^* is chosen so that the optimal stock level is eventually reached. As we see in Table 2 the expected net present value of the forest $J(h^*)$ increases with increasing values for z_0 . This is likely due to a lower prevention management spending rate h^* in the cases for larger values of z_0 .

Table 3: In this table we list the value of the objective functional evaluated at the optimal control for three different cases.

Case	$J(h^*)$ \$M
w/ stock, w/ quad. cost: $z_0 = 1, \gamma = 1$	134.69
no stock, w/ quad. cost	134.05
no stock, no quad. cost	141.47

We wish to compare our optimal control problem with cumulative prevention management stock to an optimal control problem where the effects of prevention management spending are taken to be instantaneous as in [8]. This optimal control problem with instantaneous effects of prevention spending is given by:

$$\max_{h \in U} \int_0^T \left[B(\bar{A}) - (ah(t) + \frac{\epsilon}{2}h^2(t)) + \psi(h(t))JW^*(t, h(t)) \right] e^{-r-y(t)} dt \quad (53)$$

subject to $y'(t) = \psi(h(t))$ with $y(0) = 0$,

$$h(t) \geq 0, \quad (54)$$

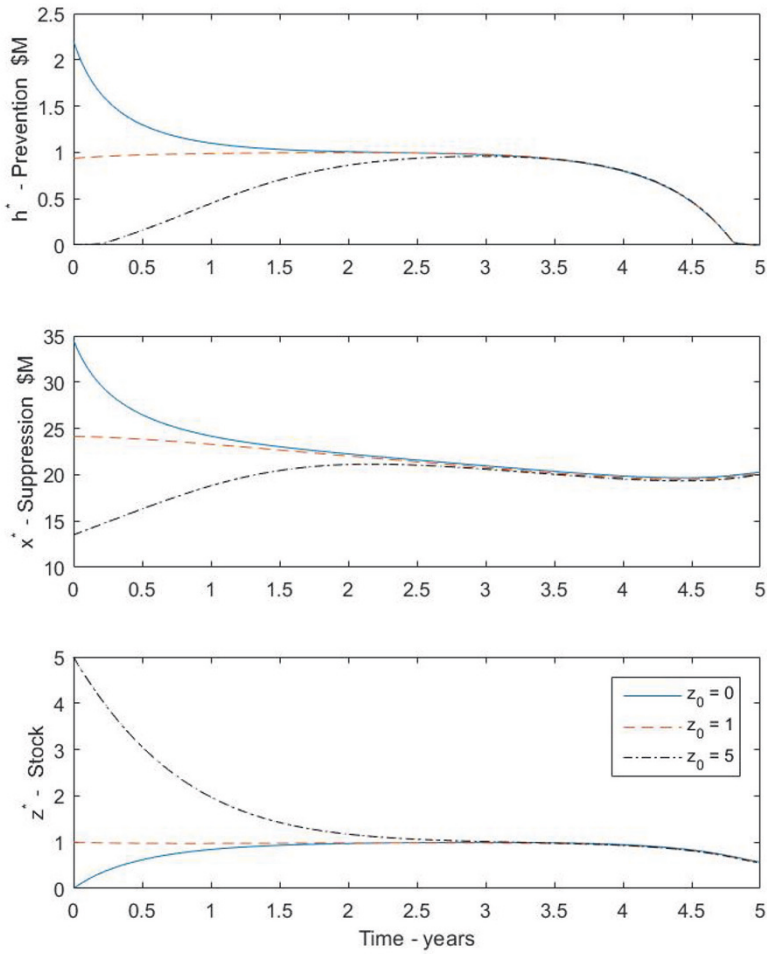


Fig. 4: The plots show the results of our optimal control problem with three different initial conditions for cumulative prevention management stock z_0 . We use $z_0 = 0$, $z_0 = 1$, and $z_0 = 5$ and compare prevention management spending h^* , suppression spending x^* , and cumulative prevention management stock z^* .

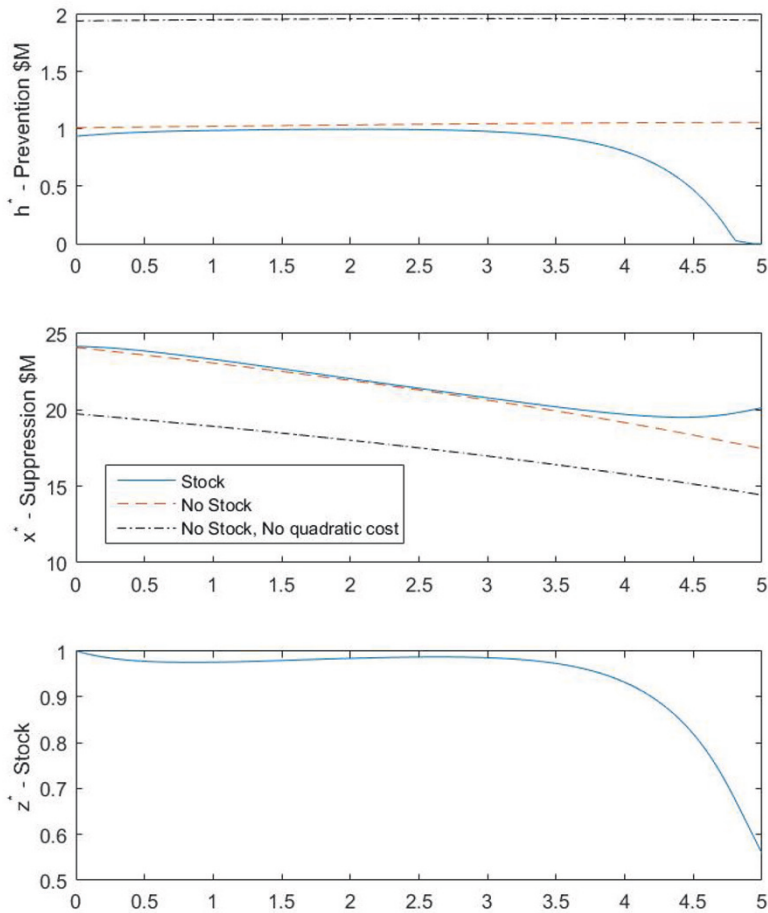


Fig. 5: Results from three different optimal control problems are displayed: with cumulative prevention management stock and quadratic cost term, no cumulative prevention management stock and quadratic cost term, and no cumulative prevention management stock and no quadratic cost term. We use $z_0 = 1$, $\gamma = 1$, and $A_0 = \bar{A}$.

where

$$U = \{h : [0, T] \rightarrow [0, \infty) | h \text{ is piecewise continuous}\}. \quad (55)$$

This optimal control problem was derived in a similar way as our cumulative prevention management stock optimal control problem.

The results for the case comparisons are contained in Table 3 and Figure 5. We also include comparisons to a optimal control problem with no stock and that does not include a quadratic cost term in the objective functional. For the case with cumulative stock, we choose $z_0 = 1$ and $\gamma = 1$. We choose $z_0 = 1$ because we want to closely match the situation in the optimal control problem without cumulative stock so that direct comparisons can be made. We choose a nonzero initial condition for cumulative stock z because in the instantaneous prevention effects optimal control problem, prevention management spending is effective immediately. In contrast, stock takes time to accumulate. Thus, if we choose $z_0 = 0$ it will take time for stock to accumulate and be effective; this is not reflective of the no stock situation. By choosing $z_0 = 1$, we allow the cumulative stock z to affect the hazard and number of acres burned in the fire early in the time horizon in a way similar to the problem without cumulative prevention stock. As is seen in Table 3, the values of the objective functional evaluated at the optimal control are nearly equal in the cases where the quadratic cost is considered. One significant difference between these cases is that in the cumulative stock case, optimal prevention management spending h^* decreases to zero as we approach the end of the time horizon. This is because h^* , given by (49), is determined by adjoint equation $\rho_2(t)$ which has transversality condition $\rho_2(T) = 0$. Hence, unless we were to include a salvage term to change this, h^* will always be pulled to zero at $t = T$ in the case including cumulative stock.

We also can compare the quadratic cost case without stock to the case without quadratic cost (also without stock). As seen in Figure 5, in the case when there is not a quadratic cost term, optimal prevention management spending is nearly double the case when there is a quadratic cost term. Moreover, the expected value of the forest $J(h^*)$ is greater in the case when the quadratic cost term is not incorporated. Thus, the inclusion of a quadratic cost term in the objective functional has a substantial impact on the solution to our optimal control problem with cumulative prevention management stock.

Furthermore, because we chose $\gamma = 1$, stock is not accumulating over time because prevention management spending h^* is slightly less than one. Thus, over the first three years of the time horizon, the stock z^* stays approximately constant near one. Let's compare this case to the case where $\gamma = 0.5$. In Figure 6 (with quadratic cost), we can see that with a lower rate of decay γ , cumulative prevention management stock is able to accumulate over time rather than remaining approximately constant even as prevention management spending h^* decreases. This in turn leads to a greater expected value of the forest. In particular, we see that $J(h^*) = \$140.5M$ in the case when $\gamma = 0.5$. Therefore, given that the rate of decay γ of cumulative prevention management stock is small enough to allow for meaningful accumulation of prevention management stock, it is possible to realize lower prevention management spending levels and an increased value of the forest.

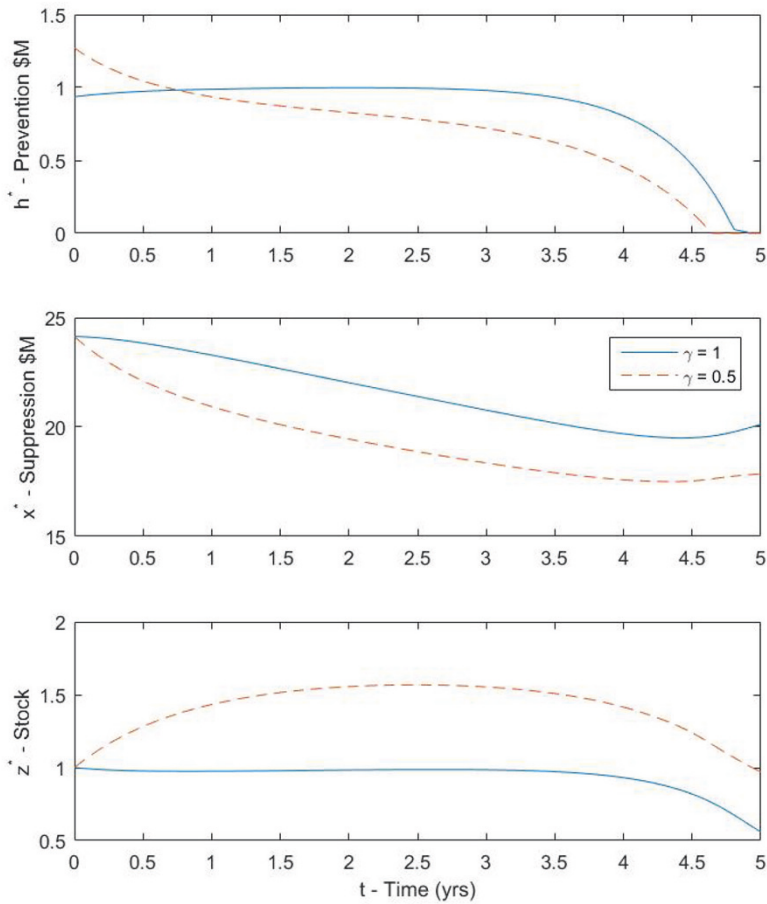


Fig. 6: In the plots above (using quadratic costs), we compare optimal prevention management stock h^* , optimal suppression spending x^* , and optimal cumulative prevention management stock z^* in the cumulative prevention management stock optimal control problem for two different values for the stock decay parameter: $\gamma = 0.5, 1$. Here, we take $z_0 = 1$ in both cases.

4 Conclusions

To investigate economic tradeoffs between fuels management spending and suppression spending with uncertain timing of large fire events, we formulate an optimal control problem with stochastic time of fire and convert it to a deterministic optimal control problem using Reed's method. We present numerical results from our optimal control problem applied to a parameter set based on a recent fire event in New Mexico with a local sensitivity analysis for three selected parameters.

Our goal is to examine the economic trade-offs between prevention management spending and suppression spending with a cumulative prevention management stock state variable. The inclusion of this state variable allowed for the effects of prevention management spending to accumulate over time or decay very rapidly, depending on the choice for γ .

Numerically, we solve this optimal control problem by maximizing the conditional current-value Hamiltonian point-wise to determine h^* . We varied the parameters A_0 , z_0 , and γ to examine their local effect on the expected value of the forest $J(h^*)$ and corresponding controls and stock z . Larger values for the initial condition A_0 for the number of unburned acres led to increased values for $J(h^*)$ and a decreased prevention management spending rate h^* . Larger values of initial cumulative prevention management stock z_0 also led to larger values of $J(h^*)$ as less prevention management spending h^* was required. Increased values for the rate of stock decay γ led to decreased values for $J(h^*)$ with mixed results for its effects on h^* . Also, when varying the initial cumulative prevention management stock z_0 , we see that all trajectories eventually merge to one trajectory over time.

In order to make comparisons between our optimal control problem with instantaneous prevention management effects and our optimal control problem with cumulative prevention management effects, we construct an intermediate optimal control problem assuming instantaneous prevention management effects with an additional quadratic cost term in the objective functional. We conclude from our work that given a small enough rate of decay γ for cumulative prevention management stock, less prevention management spending h^* is required and the expected net present value of the forest $J(h^*)$ is increased from the cases when γ is large and when stock is not considered. Thus, in cases when the cumulative prevention management stock decays too quickly, less is spent on prevention management as it is not worth the investment. Additionally, we see that the quadratic cost term in the objective functional has a substantial effect on the prevention management spending rate h^* .

There are extensions to be considered in this work, as well as other numerical methods may be applied in cases with longer time intervals and with removal of the quadratic cost term with cumulative prevention stock. We could consider the size of the fire to be stochastic or model explicitly the spread of the fire in space.

References

- [1] Agee, J. K. and Skinner, C. N. (2005). Basic principles of forest fuel reduction treatments. *Forest Ecology and Management*, 211(1):83–96.
- [2] Berry, K., Finnoff, D., Horan, R. D., and Shogren, J. F. (2015). Managing the endogenous risk of disease outbreaks with non-constant background risk. *Journal of Economic Dynamics and Control*, 51:166–179.
- [3] Calkin, D. E., Gebert, K. M., Jones, J. G., and Neilson, R. P. (2005). Forest service large fire area burned and suppression expenditure trends, 1970–2002. *Journal of Forestry*, 103(4):179–183.
- [4] Clarke, H. R. and Reed, W. J. (1994). Consumption/pollution tradeoffs in an environment vulnerable to pollution-related catastrophic collapse. *Journal of Economic Dynamics and Control*, 18(5):991–1010.
- [5] Finnoff, D., Shogren, J. F., Leung, B., and Lodge, D. (2007). Take a risk: preferring prevention over control of biological invaders. *Ecological Economics*, 62(2):216–222.
- [6] Gebert, K. M. and Black, A. E. (2012). Effect of suppression strategies on federal wildland fire expenditures. *Journal of Forestry*, 110(2):65–73.
- [7] Grissino-Myers, H. D. (1999). Modeling fire interval data from the american southwest with the weibull distribution. *Environmental Modelling and Software*, 9(1):37–50.
- [8] Heines, B., Lenhart, S., and Sims, C. (2018). Assessing the economics tradeoffs between prevention and suppression of forest fires. *Natural Resource Modelling*, 31:1–30.
- [9] Hessburg, P. F., Agee, J. K., and Franklin, J. F. (2005). Dry forests and wildland fires of the inland northwest usa: contrasting the landscape ecology of the pre-settlement and modern eras. *Forest Ecology and Management*, 211(1):117–139.
- [10] Hesseln, H. (2000). The economics of prescribed burning: a research review. *Forest Science*, 46(3):322–334.
- [11] Holmes, T. P., Prestemon, J. P., and Abt, K. L. (2008). *The economics of forest disturbances: Wildfires, storms, and invasive species*, volume 79. Springer Science & Business Media.
- [12] Horan, R. D. and Fenichel, E. P. (2007). Economics and ecology of managing emerging infectious animal diseases. *American Journal of Agricultural Economics*, 89(5):1232–1238.
- [13] Kline, J. D. (2011). *Issues in evaluating the costs and benefits of fuel treatments to reduce wildfire in the Nation's forests*. DIANE Publishing.
- [14] Lenhart, S. and Workman, J. T. (2007). *Optimal Control Applied to Biological Models*. Chapman & Hall/CRC.
- [15] Mercer, D. E., Haight, R. G., and Prestemon, J. P. (2008). Analyzing tradeoffs between fuels management, suppression, and damages from wildfire. In *The Economics of Forest Disturbances*, volume 1, pages 247–272. Springer.
- [16] Mercer, D. E., Prestemon, J. P., Butry, D. T., and Pye, J. M. (2007). Evaluating alternative prescribed burning policies to reduce net economic damages from wildfire. *American Journal of Agricultural Economics*, 89(1):63–77.

- [17] Milne, M., Clayton, H., Dovers, S., and Cary, G. J. (2014). Evaluating benefits and costs of wildland fires: critical review and future applications. *Environmental Hazards*, 13(2):114–132.
- [18] Minas, J., Hearne, J., and Martell, D. (2015). An integrated optimization model for fuel management and fire suppression preparedness planning. *Annals of Operations Research*, 232(1):201–215.
- [19] National Interagency Fire Center (2016). NIFC Fire Information: Statistics. https://www.nifc.gov/fireInfo/fireInfo_statistics.html.
- [20] National Park Service (2016). Wildland fire strategic planning. <https://www.nps.gov/fire/wildland-fire/about/plans.cfm>.
- [21] National Parks Service (2017). Bandelier National Monument. <https://www.nps.gov/band/index.htm>.
- [22] National Wildfire Coordinating Group (2013). Inciweb: Incident Information System: Las Conchas. <https://inciweb.nwcg.gov/incident/2385>.
- [23] Ninan, K. and Inoue, M. (2013). Valuing forest ecosystem services: what we know and what we don't. *Ecological Economics*, 93:137–149.
- [24] Reed, W. J. (1984). The effects of the risk of fire on the optimal rotation of a forest. *Journal of Environmental Economics and Management*, 11:180–190.
- [25] Reed, W. J. (1987). Protecting a forest against fire: Optimal protection patterns and harvest policies. *Natural Resource Modeling*, 2:23–54.
- [26] Reed, W. J. (1988). Optimal harvesting of a fishery subject to random catastrophic collapse. *Mathematical Medicine and Biology*, 5(3):215–235.
- [27] Reed, W. J. and Apaloo, J. (1991). Evaluating the effects of risk on the economics of juvenile spacing and commercial thinning. *Canadian Journal of Forest Research*, 21(9):1390–1400.
- [28] Reed, W. J. and Heras, H. E. (1992). The conservation and exploitation of vulnerable resources. *Bulletin of Mathematical Biology*, 54(2-3):185–207.
- [29] Row, C., Kaiser, H., and Sessions, J. (1981). Discount rate for long-term forest service. *Journal of Forestry*, 79(6):367–369.
- [30] Rummer, B., Prestemon, J., May, D., Miles, P., Vissage, J., McRoberts, R., Liknes, G., Shepperd, W. D., Ferguson, D., Elliot, W., et al. (2005). A strategic assessment of forest biomass and fuel reduction treatments in western states. <https://pdfs.semanticscholar.org/eabf/2f388d6e6202f6826f6341c1aa9d25254ac2.pdf>.
- [31] Ryan, K. C., Knapp, E. E., and Varner, J. M. (2013). Prescribed fire in North American forests and woodlands: history, current practice, and challenges. *Frontiers in Ecology and the Environment*, 11(s1):e15–e24.
- [32] Southwest Fire Science Consortium (2011). Las Conchas Fire: Jemez Mountains, NM. http://swfireconsortium.org/wp-content/uploads/2012/11/Las-Conchas-Factsheet_bsw.pdf.
- [33] Stephens, S. L. and Ruth, L. W. (2005). Federal forest-fire policy in the united states. *Ecological applications*, 15(2):532–542.

- [34] Thompson, M., Anderson, N., et al. (2015). Modeling fuel treatment impacts on fire suppression cost savings: A review. *California Agriculture*, 69(3):164–170.
- [35] United States Department of Agriculture (2002). Plant Fact Sheet: Ponderosa Pine. https://plants.usda.gov/factsheet/pdf/fs_pipo.pdf.
- [36] United States Department of Agriculture: Forest Service (2007). NFS Acreage by State, Congressional District and County. https://www.fs.fed.us/land/staff/lar/2007/TABLE_6.htm.
- [37] United States Department of Agriculture: Forest Service (2011). Southwest Jemez–CFLRP Annual Report 2011. <https://www.fs.usda.gov/detail/santafe/landmanagement/projects/?cid=stelprdb5416651>.
- [38] United States Department of Agriculture: Forest Service (2017a). Santa Fe National Forest. <https://www.fs.usda.gov/santafe/>.
- [39] United States Department of Agriculture: Forest Service (2017b). Santa Fe National Forest GIS Data. <https://www.fs.usda.gov/detail/r3/landmanagement/gis/?cid=stelprdb5203736>.



A Free-Model Characterization of the Asymptotic Certainty Equivalent by the Arrow-Pratt Index

Daniel Hernández-Hernández and Erick Treviño-Aguilar

Abstract This work concerns with the asymptotic behavior of the optimal wealth process, measured through the certainty equivalent of utility functions with convergent Arrow-Pratt risk aversion index, which we call regular. It is proved that, when the time horizon converges to infinity, the value function is independent of the initial capital. Moreover, when the performance is measured by another regular utility function with the same asymptotic Arrow-Pratt risk aversion index, the constant optimal value is the same, and the sets of optimal investment strategies coincide. Interestingly, these results do not depend on a model specification.

1 Introduction

A sector of financial industry provides advisement on investment decisions for retirement of an individual or at an institutional level for a pension fund. There are some rules of thumb. According to [1], a first rule of thumb (RT1) is to encourage young investors to take more risk than older investors. A second rule of thumb (RT2) is that conservative investors should take portfolios with higher shares of bonds than stocks, under the hypothesis that bonds are less risky than stocks. There is empirical evidence that both rules are common practice and mature households take less shares of stocks in comparison with younger households; see e.g., [2]. A third rule of thumb (RT3) dwells on the implementation of constant capital proportions investment strategies.

Daniel Hernández-Hernández

Department of Probability and Statistics, Research Center for Mathematics, Apartado Postal 402, Guanajuato, Gto., 36000, México, e-mail: dher@cimat.mx

Erick Treviño-Aguilar

Department of Economics and Finance, Universidad de Guanajuato, Guanajuato, Gto., 36000, México, e-mail: erick.trevino@ugto.mx

There are at least two theories on which the design of a portfolio is based. Leaving horizon effects aside for a moment, according to Markowitz's classic theory, the efficient portfolio maximizes returns for a fixed level of volatility and this generates the efficient frontier of return-volatility. The investor chooses the risk level, namely, the volatility and in this form expresses his preferences. There is an alternative formulation to Markowitz's criterion of maximizing returns subject to a volatility constraint which also gives an equivalent way to express preferences of risk. In fact, through the introduction of a Lagrange multiplier. The connection between these two formulations derives from the decreasing function connecting Lagrange multipliers with levels of risk. In this context, the Lagrange multiplier has the interpretation of an index of risk aversion: The higher the multiplier, the higher the penalization to risk, and thus, the lower the risk tolerance accepted.

Markowitz's theory notwithstanding its normative nature, incorporates risk preferences, and in its Lagrangian formulation has a "formal" connection with utility theory. Indeed, the Arrow-Pratt absolute risk aversion index scales the marginal utility gain of a marginal move from a completely riskless financial position to hold risky assets; see Section 3 for a classical result in this direction. Thus, up to error terms in a Taylor expansion, a utility maximizer also maximizes Markowitz's problem in his Lagrangian formulation. In a sense which we precise below, this is the case for the asymptotic utility maximization problem which we study here:

$$\sup \liminf_{T \rightarrow \infty} \frac{1}{r(T)} \mathcal{E}_u \left(X_T^{c, \xi} \right),$$

where \mathcal{E}_u denotes the certainty equivalent with respect to a utility function u and $X_T^{c, \xi}$ denotes the capital at time T generated by following an investment strategy ξ and savings L and r is a growth rate. The supremum is taken over a suitable class of admissible strategies. Portfolio wealth together with savings are seen as the founding sources for retirement. Savings can be seen as a cumulative proportion of income, with a fixed rate. Thus, savings are exogenously specified but are part of the criterion since utility is measured from both, savings and portfolio wealth. We will show for a class of 'regular' utility functions that the only characteristic determining an optimal allocation is the asymptotic behavior of the Arrow-Pratt absolute risk aversion index with no restrictions for the underlying model. Thus, two regular utility functions with the same asymptotical index have the same optimal portfolios. We will prove this for risk aversion parameters in the interval $(-\infty, 0)$. In order to establish our main results, we study the family of exponential utilities parameterized by its risk aversion index and establish continuity properties with respect to this index.

After this introduction, the paper is organized as follows. In Section 2, we recall some results about the stability of the problem of utility maximization in which different elements of the problem's data are subject to a perturbation, such as time horizon or the utility function. This background puts in perspective the analysis in this paper, since our results can be seen as related to stability when a given utility

function is approximating, as measured by an asymptotic index, a benchmark exponential utility function. Section 3 is about the Arrow-Pratt absolute risk aversion index. Section 4 presents our asymptotic utility maximization problem. In Section 5, our main results are presented, in which a complete characterization of the solution to our utility problem through the Arrow-Pratt index is given. To this end, the exponential utility with different risk aversions will be instrumental. In Section 6, we explore the rules of thumb in diffusion models. We will see that there is certain evidence supporting them but strongly depending on the specification of the model.

2 Stability of expected utility.

We are interested in an asymptotic utility problem in which time horizon converges to infinity. However, for the sake of discussion let us start with a fixed finite horizon. Larsen and Žitković [3] motivates with Hadamard [4] well-posedness criteria, meaning that when a new problem is presented, the existence and uniqueness of a solution, together with the analysis of the sensitivity with respect to changes in input data, should be studied. In order to be more specific, for the problem of expected utility maximization, existence and uniqueness results have been obtained under rather general conditions, while a deeper understanding of the sensitivity problem is still needed. Accordingly, [3] study stability of expected utility with respect to small market price of risk deviations. In their Theorem 2.12, they show under appropriate conditions that expected utility and its optimal trading strategy are jointly continuous with respect to initial wealth and market price of risk. In this direction Hernández and Schied [5] study the problem of robustness when the drift term of price is changed for the logarithmic utility function. Stability with respect to market price of risk is again taken up by Mostovyi and Sirbu [6] in which Taylor expansions of second order for value functions around optimal trading strategies are developed. Interestingly, their analysis is based on Taylor expansions where the (relative) coefficient of risk aversion is an important element.

Take a family of utility functions $U^\delta : \mathbb{R} \rightarrow \mathbb{R}$, with $\delta \geq 0$. Xing [7] considers a finite horizon T and the family of expected utilities with finite horizon: $\delta \rightarrow \sup E[U_\delta(X_T)]$, where the supremum is taken over a suitable family of investment strategies. He shows that pointwise continuity of the family with respect to δ , yields continuity of expected utility and convergence of optimal strategies in a suitable sense.

The previously cited papers study, in a finite horizon, stability of expected utility with respect to perturbations in initial wealth, the utility function and market price of risk. Let us now focus on stability with respect to forward movements of the horizon. In this regard, the best known results are on ‘Turnpike Theory’. McKenzie [8] recognizes three different classes of results in this theory. The prototype theorems in the first and second classes compare strategies under optimality for finite versus

infinite horizons and aim to establish convergence. McKenzie [8] traces back the label “Turnpike” to Dorfman et al. [9, Chapter 12] and attributes to them the first such theorems in the first class. It should be mentioned that the main motivation is that of capital accumulation on von Neumann [10] model which yields the difference between the first and the second class through the concept of “support prices”. The third class compares optimal strategies in infinite horizons with respect to different starting wealths. Notice that in the three classes, theorems concern stability of expected utility with respect to perturbations on horizon and initial wealth. A relative to Turnpike Theory is known as ‘portfolio turnpike theory’ and can be traced back to Mossin [11]. He shows that isoelastic utility functions (also known as power utilities or constant relative risk aversion utilities) allow for optimal myopic strategies: Optimal strategies independent of the horizon and initial wealth. The converse is also true. Thus, if optimal strategies are myopic then utilities must be isoelastic. An asymptotic version of such result is given by Huberman and Ross [12]. Thus, if the relative risk aversion index converges to a constant (so that the absolute risk aversion index converges to zero), then the optimal strategy will also converge to that of an isoelastic utility. Extensions in continuous time include Cox and Huang [13], Huang and Zariphopoulou [14], Dybvig et al. [15].

Let \tilde{U} be an isoelastic utility function so that $\tilde{U}(x) := x^p/p$ with $p \in (-\infty, 1)/\{0\}$ (for $p = 0$ let $\tilde{U}(x) = \log(x)$). Let U be a utility function with

$$\lim_{x \rightarrow \infty} \frac{U'(x)}{\tilde{U}'(x)} = 1.$$

Let X^T (resp. \tilde{X}^T) denote the optimal portfolio with respect to U (resp. \tilde{U}) for a finite horizon T . Let $r_u^T = X_u^T(\tilde{X}_u^T)^{-1}$ and $\Pi_z^T = \int_0^z (r_u^T)^{-1} dr_u^T$. Under general conditions, Guasoni et al. [16] show that

1. $\lim_{T \rightarrow \infty} \mathbb{P}^T(\sup_{u \in [0, T]} |r_u^T - 1| \geq \varepsilon) = 0.$
2. $\lim_{T \rightarrow \infty} \mathbb{P}^T([\Pi, \Pi]_T \geq \varepsilon) = 0,$

for a suitable “myopic” family of probability measures $\{\mathbb{P}^T\}_T$ which under further conditions can be taken as restrictions of a unique probability measure \mathbb{P} on a given stochastic basis. For the certainty equivalent Guasoni et al. [17, Theorem 2.4] show the asymptotic result

$$\lim_{T \rightarrow \infty} \frac{U^{-1}(E_{\mathbb{P}}[U(\tilde{X}_T^T)])}{U^{-1}(E_{\mathbb{P}}[U(X_T^T)])} = 1.$$

We will show that a regular utility function converges in a suitable sense to an exponential utility function and this last function determines the expected asymptotic certainty equivalent together with the family of optimal investment strategies. Thus, optimal strategies of regular utility functions are completely determined by its asymptotical Arrow-Pratt risk aversion index. In particular, for any regular utility function the asymptotic certainty equivalent and the optimal trading strategies are independent of the initial wealth. In this sense, we can see the results in this paper

as turnpike theorems in which we move out of isoelastic utility and thus, of zero asymptotic absolute risk aversion.

3 Arrow-Pratt index

We start this section with a few definitions.

Definition 1. A utility function is defined as a non-decreasing concave function $u : \mathbb{R} \rightarrow \mathbb{R}$. When a utility function u is of class C^2 we define the function

$$A_u(x) := \frac{u''(x)}{u'(x)}. \tag{1}$$

The well-known Arrow-Pratt absolute risk aversion function is in our notation equal to $-A_u$. Here it will be more convenient to work with minus the Arrow-Pratt function, thus our function A . If the limit

$$A[u] := \lim_{x \rightarrow \infty} A_u(x), \tag{2}$$

exists, and is a negative real number, then we say that u is a regular utility function with asymptotic absolute risk aversion $A[u]$.

Remark 1. A utility function u with Hyperbolic Absolute Risk Aversion (HARA) index is regular. It satisfies for constants a and b

$$A_u(x) = \frac{1}{ax + b}.$$

If $a = 0$ then the utility function has Constant Absolute Risk Aversion (CARA) with $A[u] = b^{-1}$. In this class, we find the exponential utility $1 - e^{\lambda x}$, with $\lambda < 0$. If $a \neq 0$ then $A[u] = 0$. Such is the case of power utilities $\gamma^{-1}x^\gamma$ and logarithmic utility function $\log(x)$.

Consider an agent taking decisions based on a utility function u . Assume he starts with an initial capital x and must compare between a loss without uncertainty so its capital reduces to $x - \pi$ or a risky situation in which its capital reduces to $x + Z$ where Z is random with $E[Z] = 0$. In this situation, the indifference price of Z given the initial capital x , denoted $\pi = \pi(x, Z)$, satisfies

$$u(x - \pi) = E[u(x + Z)]. \tag{3}$$

In this case, he will be indifferent between holding the capital $x - \pi$ or the risky asset $x + Z$. Taylor approximations on both sides of equation (3) result in:

$$u(x) - \pi u'(x) + o(\pi^2) = E[u(x) + Zu'(x) + \frac{1}{2}Z^2u''(x) + o(Z^3)]. \tag{4}$$

Then, up-to second and third order errors:

$$\pi = -\frac{1}{2} \frac{u''(x)}{u'(x)} \sigma^2(Z), \tag{5}$$

where $\sigma^2(Z)$ denotes the variance of Z . Thus, the indifference price is proportional to the variance of Z and the function A_u .

The argument based on Taylor polynomials is well known in the seminal works of Arrow [19] and Pratt [18]. We give a version and formalize the intuition of (4) in the next result. The proof follows a well-known path and it is included in 6.4 in order to illustrate the relevance of our class of regular utility functions.

Definition 2. The certainty equivalent of a random variable Y with respect to the utility function u is given by

$$\mathcal{E}_u(Y) := u^{-1}(E[u(Y)]). \tag{6}$$

Proposition 1. Let Z be a random variable uniformly bounded from below with $E[Z] = 0$ and $E[|Z|^3] < \infty$. Let u be a regular utility function of class C^3 with

$$\left| A'_u(x) u'(x) \right| \leq k, \tag{7}$$

for a constant k . Then

$$\lim_{t \searrow 0} \frac{\mathcal{E}_u(x + tZ) - x}{t^2 E[Z^2]} = \frac{1}{2} A_u(x).$$

The next result is an extension for non-bounded random variables of Cavazos and Hernández [20, Theorem 4.1].

Theorem 1. Let u and v be two regular utility functions. Take $x_0 \in (0, \infty)$. Assume that for all $x \in (x_0, \infty)$ we have $A_u(x) \geq A_v(x)$. Then, for Y a random variable with $Y \geq x_0$ a.s. we have

$$\mathcal{E}_u(Y) \geq \mathcal{E}_v(Y). \tag{8}$$

Proof. For Y as in the statement of the theorem and $n \in \mathbb{N}$ we have

$$\mathcal{E}_u(Y \wedge n) \geq \mathcal{E}_v(Y \wedge n),$$

due to [20, Theorem 4.1]. Now we can just take the limit as $n \nearrow \infty$ to see that

$$\mathcal{E}_u(Y) \geq \mathcal{E}_v(Y),$$

by monotone convergence theorem.

4 Maximization of asymptotic utility

In this section we introduce our asymptotic utility problem. To this end, we fix a continuous function $r : [0, \infty) \rightarrow [0, \infty)$ with $\lim_{T \rightarrow \infty} r(T) = \infty$ and a non-decreasing function $L : [0, \infty) \rightarrow [0, \infty)$ with

$$\lim_{T \rightarrow \infty} L(T) = \infty. \tag{9}$$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space where a filtration \mathbb{F} is defined. Let S be a \mathbb{R}^{d+1} -valued semimartingale price process defined on this probability space; see e.g., Protter [21] for details on semimartingales and its integration theory. Let \mathcal{A} denote the class of admissible strategies defined as predictable stochastic processes ξ taking values in \mathbb{R}^{d+1} that are integrable with respect to S and the integral is uniformly bounded from below by a constant; ξ_t should be interpreted as the number of shares hold at time t of the underlying asset S . For $\xi \in \mathcal{A}$ and $c > 0$ we will use the notation

$$X_T^{c, \xi} := c + \int_0^T \xi_s \cdot dS_s + L(T); \tag{10}$$

here “ \cdot ” represents the inner product on \mathbb{R}^{d+1} . Note that L is going to be fixed throughout the paper and therefore, there is no need to include it in the notation of $X^{c, \xi}$. For a regular utility function u , an admissible strategy $\xi \in \mathcal{A}$ and $c > 0$, let

$$K^u(c, \xi) := \liminf_{T \rightarrow \infty} \frac{1}{r(T)} \mathcal{E}_u \left(X_T^{c, \xi} \right). \tag{11}$$

Our main goal is to characterize the value function:

$$K^u(c) := \sup_{\xi \in \mathcal{A}} K^u(c, \xi). \tag{12}$$

A few remarks are in order.

Remark 2. It is also possible to consider \limsup in the definition of $K^u(c, \xi)$. However, \liminf will be better for our approach here. See the proof of Theorem 3. It also has a financial motivation as it expresses a conservative point of view: It considers the worst fluctuations on windows of time $[0, T]$.

Other papers studying related problems to our criterion (12) show that we can take \limsup without changing the result; see e.g., [20]. However, we expect this robustness to depend on the model and we want to keep our discussion free of a model specification.

Remark 3. The first part $c + \int \xi \cdot dS$ in the definition (10) represents the gains generated by following a self-financing investment strategy ξ and is well common in finance. The second half given by the function L represents a cumulative fraction of income from the point of view of a small investor who is planning investment for retirement. We will show that the choice of the optimal strategy for regular utility

functions is independent of L , as long as it diverges as in our assumption (9), but not too fast:

$$\lim_{T \rightarrow \infty} \frac{L(T)}{r(T)} = 0. \tag{13}$$

Remark 4. If we insist on quantifying utility only from the wealth generated by a portfolio, then it is necessary to restrict to strategies able to generate wealths converging (in some sense) to infinity. Note that an arbitrage-free model does not exclude the possibility of portfolios with exponential growth. This is the case of a model driven by an Ornstein-Uhlenbeck process see e.g., Föllmer and Schachermayer [22]. In discrete time this has been proved for stationary ergodic processes; see Dempster et al. [23]. But then again, this will depend on the model. The study of utility maximization under models driven by Ornstein-Uhlenbeck process in [22] motivates our consideration of a general rate function r . Indeed, they show that the certainty equivalent growth rate can be faster than just taking the identity function $r(T) = T$.

Alternatively, we could assume that one of the assets, say S^0 , is of the form e^{at} and thus diverges to ∞ . This is a common assumption in portfolio turnpike literature; see e.g. [12] and [16]. This however introduces a constraint in the model and requires to isolate portfolios from \mathcal{A} . Thus, we prefer our more simple specification.

5 Parametric families of utilities.

In this section we present our main results, which are summarized as follows: For regular utility functions, the value function K is constant, thus, independent of the initial capital and characterized by the asymptotical Arrow-Pratt index. Moreover, given an arbitrary regular utility function and an exponential utility function with the same asymptotical Arrow-Pratt index, it can be concluded that they have the same set of optimal investment strategies.

5.1 Continuous families of utilities.

Definition 3. A collection of utility functions $\{f_\alpha\}_{\alpha \in (-\infty, 0)}$ with

$$A[f_\alpha] = \alpha, \tag{14}$$

will be called a parametric family of utilities. A parametric family of utilities is continuous with respect to asymptotic risk aversion if

- For any sequence $\{\alpha_n\}_{n \in \mathbb{N}} \subset (-\infty, 0)$ converging to $\alpha \in (-\infty, 0)$ and $c > 0$ we have

$$\lim_{n \rightarrow \infty} K^{f_{\alpha_n}}(c) = K^{f_\alpha}(c). \tag{15}$$

In this case, we will simply say that the family is continuous.

The class of continuous families of utilities is interesting in regard to the following result.

Theorem 2. *Let $\{f_\alpha\}_{\alpha \in (-\infty, 0)}$ be a continuous parametric family of utilities. Take a regular utility function u and $c > 0$. Then, for $\alpha = A[u]$ we have*

$$K^{f_\alpha}(c) = K^u(c). \tag{16}$$

Proof. Take an element $\xi \in \mathcal{A}$. Then, for $\varepsilon \in (0, -\alpha)$, we have

$$\begin{aligned} \liminf_{T \rightarrow \infty} \frac{1}{r(T)} \mathcal{E}_{f_{\alpha-\varepsilon}}(X_T^{c, \xi}) &\leq \liminf_{T \rightarrow \infty} \frac{1}{r(T)} \mathcal{E}_u(X_T^{c, \xi}) \\ &\leq \liminf_{T \rightarrow \infty} \frac{1}{r(T)} \mathcal{E}_{f_{\alpha+\varepsilon}}(X_T^{c, \xi}), \end{aligned} \tag{17}$$

due to the monotonicity property of Theorem 1, the property (9) of L , and our definition of \mathcal{A} . We now take supremum in (17) over $\xi \in \mathcal{A}$ on all sides of the inequalities to obtain

$$K^{f_{\alpha-\varepsilon}}(c) \leq K^u(c) \leq K^{f_{\alpha+\varepsilon}}(c). \tag{18}$$

We take the limit as $\varepsilon \searrow 0$ in (18) to see that (16) holds true, due to the continuity property (15) of the family $\{f_\alpha\}_{\alpha \in (-\infty, 0)}$.

5.2 Exponential utilities

For $\lambda \in (-\infty, 0)$, let

$$\begin{aligned} f_\lambda(x) &:= -e^{\lambda x} \\ g_\lambda(y) &:= \frac{1}{\lambda} \log(-y). \end{aligned} \tag{19}$$

Note that g_λ is the inverse function of f_λ . Let $\{f_\lambda\}_{-\infty < \lambda < 0}$ be the family determined by (19) with family of inverse functions $\{g_\lambda\}_{-\infty < \lambda < 0}$. Note also that for f_λ the absolute risk aversion function is constant: $A_{f_\lambda}(x) = \lambda$. Thus, $A[f_\lambda] = \lambda$.

Recall the notation $X_T^{c, \xi}$ in equation (10) and the value function K^u defined in equation (12). For our parametric family of exponential utilities we will write K^λ instead of K^{f_λ} . Thus, for $c > 0$

$$\begin{aligned} K^\lambda(c) &= \sup_{\xi \in \mathcal{A}} \liminf_{T \rightarrow \infty} \frac{1}{r(T)} \left\{ \frac{1}{\lambda} \log \left(E \left[e^{\lambda X_T^{c, \xi}} \right] \right) \right\} \\ &= \frac{1}{\lambda} \inf_{\xi \in \mathcal{A}} \limsup_{T \rightarrow \infty} \frac{1}{r(T)} \left\{ \log \left(E \left[e^{\lambda X_T^{c, \xi}} \right] \right) \right\}. \end{aligned} \tag{20}$$

We note for future reference that

$$\begin{aligned} K^\lambda(c) &= \frac{1}{\lambda} \inf_{\xi \in \mathcal{A}} \limsup_{T \rightarrow \infty} \frac{1}{r(T)} \left\{ \log \left(e^{\lambda c} E \left[e^{\lambda X_T^{0,\xi}} \right] \right) \right\} \\ &= K^\lambda(0). \end{aligned} \tag{21}$$

Lemma 1. For $c > 0$ fixed, the function

$$\lambda \rightarrow \lambda K^\lambda(c), \tag{22}$$

is convex on $(-\infty, 0)$.

Proof. Take $\xi^1, \xi^2 \in \mathcal{A}$, $\alpha \in (0, 1)$ and $\lambda^1, \lambda^2 \in (-\infty, 0)$. Let $\lambda^3 := \lambda^1 \alpha + \lambda^2 (1 - \alpha)$. We define ξ^3 as the convex combination

$$\xi^3 := \xi^1 \frac{\lambda^1 \alpha}{\lambda^3} + \xi^2 \frac{\lambda^2 (1 - \alpha)}{\lambda^3}.$$

It is clear that $\xi^3 \in \mathcal{A}$. Now we proceed as in the proof of Lemma 2.2.5 in Dembo and Zeitouni [24]. We have

$$\begin{aligned} E \left[e^{\lambda^3 X_T^{c,\xi^3}} \right] &= E \left[e^{\alpha \lambda^1 X_T^{c,\xi^1}} e^{(1-\alpha) \lambda^2 X_T^{c,\xi^2}} \right] \\ &\leq E \left[e^{\lambda^1 X_T^{c,\xi^1}} \right]^\alpha E \left[e^{\lambda^2 X_T^{c,\xi^2}} \right]^{1-\alpha}. \end{aligned} \tag{23}$$

Now we take logarithm on the first and the last terms of (23) to see that

$$\log E \left[e^{\lambda^3 X_T^{c,\xi^3}} \right] \leq \alpha \log E \left[e^{\lambda^1 X_T^{c,\xi^1}} \right] + (1 - \alpha) \log E \left[e^{\lambda^2 X_T^{c,\xi^2}} \right]. \tag{24}$$

In view of (20), the inequality (24) yields the convexity of (22).

Theorem 3. The exponential family of utilities is continuous and for any regular utility function u we have

$$K^u(c) = K^\lambda(c), \text{ for } c > 0, \tag{25}$$

with $\lambda = A[u]$. Moreover, $K^u(\cdot)$ is a constant function.

Proof. For $c > 0$ fixed, the function

$$\lambda \rightarrow \lambda K^\lambda(c),$$

is convex in $(-\infty, 0)$ due to Lemma 1. Thus, the function

$$\lambda \rightarrow K^\lambda(c),$$

is continuous. As a consequence, the exponential family of utilities is continuous as in Definition 3 and by Theorem 2 we see that for the regular utility function u

$$K^u(c) = K^\lambda(c),$$

with $\lambda = A[u]$. The equality $K^\lambda(c) = K^\lambda(0)$ is just (21) and shows that $K^u(\cdot)$ is a constant function.

Theorem 4. *Let u be a regular utility function. Take $c > 0$. Let $\lambda = A[u]$. An investment strategy $\xi^* \in \mathcal{A}$ is optimal for $K^\lambda(c)$ so that*

$$K^\lambda(c) = \liminf_{T \rightarrow \infty} \frac{1}{r(T)} \left\{ \frac{1}{\lambda} \log \left(E \left[e^{\lambda X_T^{c, \xi^*}} \right] \right) \right\}$$

if and only if it is also optimal for $K^u(c)$:

$$K^u(c) = \liminf_{T \rightarrow \infty} \frac{1}{r(T)} \left\{ u^{-1} \left(E \left[u \left(X_T^{c, \xi^*} \right) \right] \right) \right\}.$$

Proof. Let $\xi^* \in \mathcal{A}$ be an optimal strategy for $K^u(c)$. The function

$$\lambda \rightarrow \liminf_{T \rightarrow \infty} \frac{1}{r(T)} \mathcal{E}_\lambda(X_T^{c, \xi^*}) \tag{26}$$

is convex in the interval $(-\infty, 0)$. Indeed, this follows easily from the inequality (23) by taking $\xi^1 = \xi^2 = \xi^*$. As a consequence, the function (26) is continuous on $(-\infty, 0)$.

Put $\xi = \xi^*$ in the inequality (17) and then take the limit $\varepsilon \searrow 0$ to see that

$$\liminf_{T \rightarrow \infty} \frac{1}{r(T)} \left\{ u^{-1} \left(E \left[u \left(X_T^{c, \xi^*} \right) \right] \right) \right\} = \liminf_{T \rightarrow \infty} \frac{1}{r(T)} \left\{ \frac{1}{\lambda} \log \left(E \left[e^{\lambda X_T^{c, \xi^*}} \right] \right) \right\},$$

due to the continuity of the function (26). The optimality of ξ^* for $K^\lambda(c)$ is now consequence of Theorem 3. The other direction follows similarly.

Remark 5. Let us see Theorems 3 and 4 in the perspective of Turnpike Theorems. A regular utility function u must satisfy that $A[u] = \lim_{x \rightarrow \infty} A_u(x)$ exists as a negative real number. In this case for $\varepsilon > 0$ with $A[u] + \varepsilon < 0$, there exists $x_\varepsilon > 0$ such that for $x \geq x_\varepsilon$ it holds that $|A_u(x) - A[u]| < \varepsilon$. But then, the equality $u'' = A_u u'$ yields the estimation

$$u'(x_\varepsilon) e^{(A[u] - \varepsilon)(x - x_\varepsilon)} \leq u'(x) \leq u'(x_\varepsilon) e^{(A[u] + \varepsilon)(x - x_\varepsilon)}, \tag{27}$$

due to Gronwall's inequality; see Ethier and Kurtz [25, Theorem A.5.1]. In particular we see that u' converges to zero and as fast as $e^{A[u]x}$ and the same is true for u'' . Thus, our definition of a regular utility function incorporates the usual condition of marginal utility convergence towards the benchmark utility function (in Turnpike

Theory, the isoelastic utility, in the present paper, the exponential utility). Our condition involves also the second derivative and allows for a stronger characterization of optimal investment strategies.

6 Common practices under the perspective of diffusion models

In addition to the value function K^λ , it will be useful to introduce the finite horizon exponential utility maximization problem

$$u_T(x) := \sup_{\xi \in \mathcal{A}} E \left[-\exp \left\{ \lambda \left(x + \int_0^T \xi dS \right) \right\} \right],$$

with $\lambda \in (-\infty, 0)$, together with the limit function

$$K^{\lambda,+}(x) := \frac{1}{\lambda} \limsup_{T \rightarrow \infty} \frac{1}{r(T)} \log(-u_T(x)). \quad (28)$$

It is clear that $K^{\lambda,+}$ dominates from above the function K^λ defined above in (20).

Throughout this section we shall be restricted to the case when there is only one risky asset. It should be noted that in the definition of u_T the function L defined in Section 4 (see (9)) does not appear. The reason for removing this term from the definition of the wealth process $X_T^{x,\xi}$ in (10) is that we shall be working first with finite horizon problems and then, when the limit as $T \rightarrow \infty$ is taken, we can use hypothesis (13).

6.1 Rules of thumb under geometric Brownian motion

Let $\{S_t\}_{0 \leq t \leq T}$ be the price process in the Black-Scholes model without drift. Thus $S_t = e^{\sigma W_t}$ with $\sigma > 0$ and $\{W_t\}_{0 \leq t \leq T}$ a Brownian motion. The unique martingale probability measure Q of S has density

$$Z_T = \exp \left\{ -\frac{\sigma}{2} W_T - \frac{1}{8} \sigma^2 T \right\}$$

and is a crucial element in the solution of exponential utility maximization $u_T(x)$. Let us recall briefly the solution provided by Föllmer and Schachermayer [22] for this problem. A warning to the reader about notation is the following. In this paper we take λ to be negative whereas [22] consider λ to be positive and exponential utility is then $-e^{-\lambda t}$. We keep our notation with the appropriate adjustments. The optimal wealth is given by

$$X_T^\lambda = x - \frac{1}{\lambda} [H_T(Q | P) - \log Z_T],$$

where

$$H_T(Q | P) = E_Q[\log Z_T]$$

denotes the relative entropy of Q with respect to P on \mathcal{F}_T . The maximal expected utility at time T is given by

$$u_T(x) = -\exp\{\lambda x - H_T(Q | P)\},$$

while the certainty equivalent is equal to $x - \frac{1}{\lambda} H_T(Q | P)$. In this case there is an explicit expression for the entropy

$$H_T(Q | P) = \frac{1}{8} \sigma^2 T, \tag{29}$$

see [22, Proposition 5.4].

The investment strategy ξ^λ which represents X_T^λ as $x + \int_0^T \xi_t^\lambda dS_t$ is explicitly given by $\xi^\lambda = \frac{1}{|\lambda|} \xi^*$, with

$$\xi_t^* := \frac{1}{2S_t}; \tag{30}$$

see [22, Proposition 5.5]. Thus, the optimal strategy is *myopic* in that it does not depend on the horizon T . Let us see the consequences of this property in regard to our problem K^λ and the rules of thumb. First of all, note that we have a process $\{\xi_t^\lambda\}_{t \in [0, \infty)}$ which is defined on the whole interval $[0, \infty)$ and defines an optimal investment strategy in each period $[0, T]$. It is not an element of \mathcal{A} but it does not increase the value under appropriate conditions; see [26, Theorem 2.2 and Lemma 5.1]. Thus, $K^\lambda = K^{\lambda,+}$ and

$$K^{\lambda,+}(x) = \frac{1}{\lambda} \lim_{T \rightarrow \infty} \frac{1}{T} \log[-u_T(x)] = \frac{1}{8|\lambda|} \sigma^2. \tag{31}$$

From the explicit form of the optimal strategy we see that more risk aversion from the investor indeed yields less from the risky asset in the portfolio. Thus, making out sense of the second rule of thumb RT2: less shares of stocks for conservative investors, if conservative means smaller absolute risk aversion index. How about RT1 in which youngsters should take more risk, tantamount of more stock shares?. There are two possible formulations. The first is just considering that young people will perceive a safer world, interpreting this as smaller $|\lambda|$, and then, RT2 justifies RT1. The second possibility is in which the horizon T plays a role, since it is longer for youngsters. However, for our asymptotic problem and under geometric Brownian motion, the horizon does not play a role since the optimal investment strategy is myopic. Hence, RT1 requires to start with the prior assumption that young people decide under lower λ 's, and thus, in the language of Samuelson [27] are willing to take more "businessmen risk" in comparison to older people.

6.2 Constant proportions under geometric Brownian motion

Another common practice is to implement investment strategies defined by constant proportions of capital. So we want to have an estimation of the certainty equivalent for this class. Results of the previous section demonstrate that such strategies are non-optimal. Yet, its performance is unclear and is what we study in this section.

A portfolio value $X^{x_0, \varpi}$ associated with a constant proportion of capital type of strategy ϖ satisfies

$$\begin{aligned} X^{x_0, \varpi} &= x_0 \exp \left\{ \int \frac{\varpi}{S} dS - \frac{1}{2} \int \frac{\varpi^2}{S^2} d\langle S \rangle \right\} \\ &= x_0 \exp \left\{ \varpi \left(\int \sigma dW + \frac{1}{2} d\langle \sigma W \rangle \right) - \frac{1}{2} \int \varpi^2 d\langle \sigma W \rangle \right\} \\ &= x_0 \exp \left\{ \varpi \sigma W + \frac{1}{2} \langle \sigma W \rangle (\varpi - \varpi^2) \right\}, \end{aligned}$$

the first equality is just equation (37) in the appendix. In particular, the growth rate of $X^{x_0, \varpi}$ is given by

$$\begin{aligned} \lim_{T \rightarrow \infty} \frac{1}{T} \log X_T^{x_0, \varpi} &= \varpi \lim_{T \rightarrow \infty} \frac{1}{T} \log(S_T) + \lim_{T \rightarrow \infty} \frac{1}{T} \frac{1}{2} \langle \sigma W \rangle_T (\varpi - \varpi^2) \\ &= \frac{1}{2} \sigma^2 (\varpi - \varpi^2). \end{aligned}$$

Thus, the optimal constant proportion is given by $\varpi = \frac{1}{2}$ which maximizes logarithmic utility.

In the next result, the asymptotic certainty equivalent of constant proportions is estimated from below.

Proposition 2. For $\varepsilon > 0$, let

$$\varpi^\varepsilon = \frac{\varepsilon}{8 + 2\varepsilon}. \quad (32)$$

Let X^ε be the capital generated by a strategy of the constant proportion of capital ϖ^ε . Then

$$\lim_{T \rightarrow \infty} \frac{1}{\lambda T} \log E[e^{\lambda X_T^\varepsilon}] \geq \frac{1}{|\lambda|} \frac{1}{2} \frac{1}{4 + \varepsilon} \sigma^2.$$

Proof. For a fixed constant proportion ϖ , let X^ϖ be the portfolio generated by the strategy of constant proportion ϖ . We use the notation $\beta := \frac{1}{2} T \sigma^2 (\varpi - \varpi^2)$, $r := |\lambda| x_0 e^\beta$. For $a > 0$ we let $A := a \log a - a \log(|\lambda| x_0) - a$ and $\psi = \frac{a}{r}$. We have

$$\begin{aligned}
 E[e^{\lambda X_T^\varpi}] &= E \left[\exp \{ -r\psi\varpi\sigma W_T \} \exp \left\{ \sup_{z \in \mathbb{R}} (r\psi\varpi\sigma z - re^{\varpi\sigma z}) \right\} \right] \\
 &= \exp \{ r(\psi \log \psi - \psi) \} E [\exp \{ -r\psi\varpi\sigma W_T \}] \\
 &= \exp \{ r(\psi \log \psi - \psi) \} \exp \left\{ \frac{1}{2} T (r\psi\varpi\sigma)^2 \right\} \\
 &= \exp \left\{ A - a\beta + \frac{1}{2} T (a\varpi\sigma)^2 \right\} \\
 &= \exp \left\{ A + \frac{1}{2} a\varpi\sigma^2 T (a\varpi + \varpi - 1) \right\}.
 \end{aligned}
 \tag{33}$$

As a consequence

$$\begin{aligned}
 \lim_{T \rightarrow \infty} \frac{1}{\lambda T} \log E[e^{\lambda X_T^\varpi}] &\geq \lim_{T \rightarrow \infty} \frac{1}{\lambda T} \left(A + \frac{1}{2} a\varpi\sigma^2 T (a\varpi + \varpi - 1) \right) \\
 &= \frac{1}{2\lambda} \sigma^2 a\varpi (a\varpi + \varpi - 1).
 \end{aligned}$$

A simple substitution shows that for $\varepsilon > 0$, $a = \frac{4}{\varepsilon}$ and $\varpi = \frac{\varepsilon}{8+2\varepsilon}$, we indeed have

$$a\varpi (a\varpi + \varpi - 1) = \frac{-1}{4 + \varepsilon}.$$

Remark 6. Strategies holding constant proportions are non-optimal but Proposition 2 shows that they are close to the optimal growth of certainty equivalent (31). They are commonly used. Implementation costs of a more complex strategy together with ‘near-optimality’ might motivate a better perspective of such strategies.

6.3 Rules of thumb under a factor model

Fix a two dimensional Brownian motion (W^1, W^2) . For $\theta \in [0, 1]$, the process Y is defined by

$$dY = g(Y)dt + \beta(Y)[\theta dW^1 + \sqrt{1 - \theta^2}dW^2].$$

The process Y is an exogenous factor driving the coefficients of the price process S with dynamic

$$dS = \mu(S, Y)dt + \gamma(S, Y)dW^1.
 \tag{34}$$

The dynamic of a portfolio takes the form

$$dX = \xi dS.$$

For a given ξ , let Q^ξ be the probability measure with density $\frac{dQ^\xi}{d\mathbb{P}} = M$ where

$$M := \mathcal{E} \left\{ \lambda \int \xi \gamma(S, Y) dW^1 \right\},$$

and \mathcal{E} denotes the Dooleans-Dade exponential. We want to minimize $E[e^{\lambda X_T}]$. In order to give a dynamical programming approach to this minimization problem we follow Fleming and Soner [28, Section VI.8], and to this end, the notation $\rho = -\lambda$ will be more convenient. We have

$$\begin{aligned} & E[e^{\lambda X_T}] \\ &= E_{Q^\xi} \left[\frac{1}{M_T} e^{\lambda X_T} \right] \\ &= E_{Q^\xi} \left[\exp \left\{ -\lambda \int_0^T \xi \gamma dW^1 + \frac{1}{2} \lambda^2 \int_0^T (\xi \gamma)^2 dt \right\} \exp \left\{ \lambda \int_0^T \xi \gamma dW^1 + \lambda \int_0^T \xi \mu dt \right\} \right] \\ &= E_{Q^\xi} \left[\exp \left\{ \rho \int_0^T \ell(\xi, S, Y) dt \right\} \right], \end{aligned}$$

where

$$\ell(\xi, s, y) := -\xi \mu(s, y) + \frac{\rho}{2} (\xi \gamma(s, y))^2.$$

Under Q^ξ the processes S and Y follow the dynamic

$$\begin{pmatrix} dS \\ dY \end{pmatrix} = f(\xi, S, Y) dt + \rho^{-\frac{1}{2}} \sigma \cdot \begin{pmatrix} d\tilde{W}^1 \\ dW^2 \end{pmatrix},$$

where $\tilde{W}^1 = W^1 + \rho \int \xi \gamma(S, Y) dt$ is a Q^ξ -Brownian motion and

$$\begin{aligned} f &= \begin{pmatrix} \mu - \rho \gamma^2 \xi \\ g - \theta \rho \beta \gamma \xi \end{pmatrix} \\ \sigma &= \rho^{\frac{1}{2}} \begin{pmatrix} \gamma & 0 \\ \theta \beta & \sqrt{1 - \theta^2} \beta \end{pmatrix}. \end{aligned}$$

We also put

$$a = \sigma \sigma^{\text{tr}} = \rho \begin{pmatrix} \gamma^2 & \theta \beta \gamma \\ \theta \beta \gamma & \beta^2 \end{pmatrix}.$$

Let Φ denote the value function of our minimization problem. The logarithmic transformed function $V = \frac{1}{\rho} \log \Phi$ satisfies under suitable conditions a dynamic programming equation: $-V_t + \bar{H}(V) = 0$, where the Hamiltonian \bar{H} is as in [28, Section VI.8, Equation (8.11)]. It takes the form

$$\bar{H}[V] = -L[V] - \tilde{H}[V],$$

where the operators \tilde{H} and L are, respectively,

$$\begin{aligned} \tilde{H}[V] &:= \inf_{\xi} [-\rho\gamma^2\xi V_s - \theta\rho\beta\gamma\xi V_y + \ell] \\ L[V] &:= \mu V_s + gV_y + \frac{1}{2}\gamma^2 V_{ss} + \theta\beta\gamma V_{ys} + \frac{1}{2}\beta^2 V_{yy} + \frac{1}{2}DV^{\text{tr}}(a)^{\text{tr}}DV. \end{aligned}$$

An optimal feedback control takes the form

$$\xi^* = V_s + \theta \frac{\beta}{\gamma} V_y + \frac{1}{\rho} \frac{\mu}{\gamma^2}. \tag{35}$$

Here we do not pursue to prove the existence of a solution to the dynamic equation, which can be done; see e.g., Nagai [29]. Instead, we compare the form of the optimal strategy with the rules of thumb.

It should be clear, due to the form (35) of the optimal strategy, that under the dynamic (34), the rules of thumb will be confirmed or rejected according to different choices of the coefficients. So let us choose some options. Take for concreteness $\mu \geq 0$. Then, RT2 in which conservative investors should hold less shares of stocks makes sense in the region where the partial derivatives V_s and V_y decrease with ρ . Paradoxically, this relationship is inverted as soon as $\mu < 0$. If $\theta = 0$, then there is no interaction between the randomness (W^1, W^2) in Y and S , and the optimal strategy is pretty much determined by the market price of risk and volatility. Thus, the second term in the right hand side of (35) appears as a correction factor due to this interaction. In the special case in which μ only depends on Y , as in Bachelier’s model, we will have that V does not depend on S and then the optimal solution is again myopic and therefore $K^\lambda = K^{\lambda,+}$.

6.4 Concluding remarks

In this paper we studied an asymptotic utility problem for a wide class of utility functions and showed that the solution is characterized through the behavior at infinity of the Arrow-Pratt absolute risk aversion index. Taking into account our arguments’ simplicity, the interest of the results here presented lies on the one hand on its free-model scope, and on the other, on highlighting, from a new point of view, the role of the Arrow-Pratt index on portfolio decisions.

Acknowledgements The work of the first author was partially supported by Conacyt under grant 254166.

Appendix

Stochastic Logarithm and portfolio proportions.

For a continuous semimartingale Z , the stochastic logarithm is defined by

$$\mathcal{L}(Z) := \int \frac{dZ}{Z}. \quad (36)$$

We will characterize investment strategies ξ that yield a constant proportion ϖ of the capital. This means that

$$\varpi = \frac{\xi S}{x + \int \xi dS},$$

remains constant along the time. Then, for $X_t = x + \int_0^t \xi_z dS_z$ we have

$$X_t = x + \int \frac{\varpi X_z}{S_z} dS_z.$$

Therefore, X solves the following dynamic

$$d\mathcal{L}(X) = \varpi d\mathcal{L}(S).$$

Thus

$$X = x \mathcal{E} \left(\int \varpi \frac{dS}{S} \right), \quad (37)$$

where

$$\mathcal{E}(Z) = e^{Z - \frac{1}{2}\langle Z \rangle},$$

denotes the Dooleans-Dade exponential of the continuous semimartingale Z .

Proof of Proposition 1.

Proof. 1. Let

$$J(t) := E[u(x + tZ)] - u(x)$$

and

$$R(t) := \int_x^{x+tZ} (x+tZ-s)^2 u'''(s) ds.$$

Note that for t fixed, $R(t)$ is clearly a measurable function. A Taylor expansion of second order shows

$$\begin{aligned}
 J(t) &= tu'(x)E[Z] + \frac{t^2}{2}u''(x)E[Z^2] + \frac{1}{2}E[R(t)] \\
 &= \frac{t^2}{2}u''(x)E[Z^2] + \frac{1}{2}E[R(t)].
 \end{aligned}
 \tag{38}$$

Let

$$R'(t) := \int_{u(x)}^{u(x)+J(t)} (u(x) + J(t) - s)u''(s)ds.$$

A first order Taylor expansion shows that

$$u^{-1}(E[u(x+tZ)]) = x + \frac{1}{u'(x)}J(t) + R'(t).$$

Then

$$\begin{aligned}
 u^{-1}(E[u(x+tZ)]) &= x + \frac{u''(x)}{u'(x)} \frac{t^2}{2}E[Z^2] + \frac{1}{2u'(x)}E[R(t)] + R'(t) \\
 &= x + t^2E[Z^2] \frac{A_u(x)}{2} + \frac{1}{2u'(x)}E[R(t)] + R'(t).
 \end{aligned}$$

2. There exists a random variable η depending on t with $\eta^+ \leq Z^+$ and $\eta^- \leq Z^-$ and

$$R(t) = tZ(tZ - t\eta)^2 u'''(x + t\eta). \tag{39}$$

Let us prove the claim. Assume without loss of generality that Z is non negative. For t and x fixed, the functions $h, k : \Omega \times \mathbb{R}$ defined by

$$\begin{aligned}
 h(\omega, y) &= \int_x^{x+ty} (x + tZ(\omega) - s)^2 u'''(s)ds \\
 k(\omega, y) &= tZ(\omega)(tZ(\omega) - ty)^2 u'''(x + ty)
 \end{aligned}$$

are clearly measurable in ω and continuous in y . Thus, they are Caratheodory functions. In particular, $\alpha(\omega) = h(\omega, Z(\omega))$ is measurable; see Rockafellar and Wets [30][Example 14.29, Corollary 14.34]. The correspondence

$$\begin{aligned}
 F(\omega) &= \{r \in [0, Z(\omega)]\} \cap \{r \mid k(\omega, r) \leq \alpha(\omega)\} \cap \{r \mid -k(\omega, r) \leq -\alpha(\omega)\} \\
 &= \{r \in [0, Z(\omega)]\} \cap \{r \mid k(\omega, r) = \alpha(\omega)\},
 \end{aligned}$$

is clearly closed-valued. Let us see that it is measurable. Indeed, F is defined as the intersection of correspondences, which are level-set mappings and then, are measurable; see [30][Proposition 14.33]. The intersection of closed-valued measurable correspondences is closed-valued and measurable; see [30][Proposition 14.11]. Thus, F is measurable.

For each ω the set $F(\omega)$ is non-empty (this is just the integral mean value theorem). As a consequence, there exists a measurable selection of F ; see [30][Corollary 14.6]. This is the required random variable η .

A simple computation shows that

$$u''' = (A'_u + A_u^2)u'. \quad (40)$$

Then u''' is a bounded function, due to our assumption (7) that $A'_u u'$ is a bounded function, that u is a regular function and thus A_u is a bounded function and u' converges to zero; see (27). The boundedness of u''' together with (39) yields the existence of a constant k with

$$E[|R(t)|] \leq kt^3 E[|Z|^3].$$

Thus $\lim_{t \searrow 0} R(t)/t^2 = 0$. The same is true for R' : $\lim_{t \searrow 0} R'(t)/t^2 = 0$, which can be proved similarly.

References

1. J. Y. Campbell, L. M. Viceira, *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors*, Oxford: Oxford University Press, 2002.
2. A. Fagereng, C. Gottlieb, L. Guiso, Asset market participation and portfolio choice over the life-cycle, *The Journal of Finance* 72 (2) (2017) 705–750.
3. K. Larsen, G. Žitković, Stability of utility-maximization in incomplete markets, *Stochastic Processes and their Applications* 117 (11) (2007) 1642–1662.
4. J. Hadamard, Sur les problèmes aux dérivées partielles et leur signification physique, *Princeton University Bulletin* 13 (1902) 49–52.
5. D. Hernández-Hernández, A. Schied, A control approach to robust utility maximization with logarithmic utility and time consistent penalties, *Stochastic Processes and their Applications* 117 (2007) 980–1000.
6. O. Mostovyi, M. Sirbu, Sensitivity analysis of the utility maximization problem with respect to model perturbations, Preprint.
7. H. Xing, Stability of the exponential utility maximization problem with respect to preferences, *Mathematical Finance* 27 (1) (2017) 38–67.
8. L. W. McKenzie, Turnpike theory, *Econometrica* 44 (5) (1976) 841–865.
9. R. Dorfman, P. Samuelson, R. Solow, *Linear Programming and Economic Analysis*. New York, McGraw-Hill, 1958.
10. J. von Neumann, Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes, *Ergeb. eine Math. Koll. Vienna ed. Karl Menger* 8 (1937) 73–83.
11. J. Mossin, Optimal multiperiod portfolio policies, *The Journal of Business* 41 (2) (1968) 215–229.
12. G. Huberman, S. Ross, Portfolio turnpike theorems, risk aversion, and regularly varying utility functions, *Econometrica* 51 (5) (1983) 1345–1361.
13. J. C. Cox, C. Huang, A continuous-time portfolio turnpike theorem, *Journal of Economic Dynamics and Control* 16 (3) (1992) 491 – 507.
14. C. Huang, T. Zariphopoulou, Turnpike behavior of long-term investments, *Finance and Stochastics* 3 (1) (1999) 15–34.

15. P. H. Dybvig, L. C. G. Rogers, K. B., Portfolio turnpikes, *The Review of Financial Studies* 12 (1) (1999) 165.
16. P. Guasoni, C. Kardaras, S. Robertson, H. Xing, Abstract, classic, and explicit turnpikes, *Finance and Stochastics* 18 (1) (2014) 75–114.
17. P. Guasoni, J. Muhle-Karbe, H. Xing, Robust portfolios and weak incentives in long-run investments, *Mathematical Finance* 27 (1) (2017) 3–37.
18. J. W. Pratt, Risk aversion in the small and in the large, *Econometrica* 32 (1/2) (1964) 122–136.
19. K. J. Arrow, *Aspects of the Theory of Risk Bearing*, Yrjo Jahnssonin Saatio, Helsinki, 1965.
20. Cavazos-Cadena, R. and Hernández-Hernández, D., A characterization of the optimal average cost via the arrow-pratt sensitivity function, *Mathematics of Operations Research* 41 (1) (2015) 224–235.
21. P. Protter, *Stochastic integration and differential equations*, in: *Stochastic modelling and applied probability*, version 2.1, second Edition, Vol. 21, Springer, Berlin Heidelberg New York, 2005.
22. H. Föllmer, W. Schachermayer, Asymptotic arbitrage and large deviations, *Mathematics and Financial Economics* 1 (3) (2008) 213–249.
23. M. A. H. Dempster, I. V. Evstigneev, K. R. Schenk-Hoppé, Exponential growth of fixed-mix strategies in stationary asset markets, *Finance and Stochastics* 7 (2) (2003) 263–276.
24. A. Dembo, O. Zeitouni, *Large deviations techniques and applications*, 2nd Edition, Springer, Berlin, Heidelberg, New York, 1998.
25. S. N. Ethier, T. G. Kurtz, *Markov Processes, characterization and convergence*, Wiley series in probability and mathematical statistics, John Wiley and Sons, 1986.
26. Y. Kabanov, C. Stricker, On the optimal portfolio for the exponential utility maximization: Remarks to the six-author paper, *Mathematical finance* 12 (2) (2002) 125–134.
27. P. A. Samuelson, Lifetime Portfolio Selection By Dynamic Stochastic Programming, *The Review of Economics and Statistics* 51 (3) (1969) 239–246.
28. W. Fleming, H. Soner, *Controlled Markov Processes and Viscosity Solutions*, 2nd Edition, Springer-Verlag New York, 2006.
29. H. Nagai, Downside risk minimization via a large deviations approach, *Ann. Appl. Probab.* 22 (2) (2012) 608–669.
30. R. T. Rockafellar, J. B. R. Wets, *Variational Analysis*, Springer, 1997.



Binary Mean Field Stochastic Games: Stationary Equilibria and Comparative Statics

Minyi Huang and Yan Ma

Abstract This paper considers mean field games in a multi-agent Markov decision process (MDP) framework. Each player has a continuum state and binary action, and benefits from the improvement of the condition of the overall population. Based on an infinite horizon discounted individual cost, we show existence of a stationary equilibrium, and prove its uniqueness under a positive externality condition. We further analyze comparative statics of the stationary equilibrium by quantitatively determining the impact of the effort cost.

1 Introduction

Mean field game theory provides a powerful methodology for reducing complexity in the analysis and design of strategies in large population dynamic games [25, 30, 37]. Following ideas in statistical physics, it takes a continuum approach to specify the aggregate impact of many individually insignificant players and solves a special stochastic optimal control problem from the point of view of a representative player. By this methodology, one may construct a set of decentralized strategies for the original large but finite population model and show its ε -Nash equilibrium property [25, 26, 30]. A related solution notion in Markov decision models is the oblivious equilibrium [55]. The readers are referred to [12, 16, 17, 18, 19] for an overview on

Minyi Huang

School of Mathematics and Statistics, Carleton University, Ottawa, ON K1S 5B6, Canada e-mail: mhuang@math.carleton.ca. This author was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada

Yan Ma

School of Mathematics and Statistics, Zhengzhou University, 450001, Henan, China e-mail: mayan203@zzu.edu.cn. This author was supported by the National Science Foundation of China (No. 11601489)

mean field game theory and further references. For mean field type optimal control, see [12, 56], but the analysis in these models only involves a single decision maker.

Dynamic games within an MDP setting originated from the work of Shapley and are called stochastic games [21, 50]. Their mean field game extension has been studied in the literature; see e.g. [3, 13, 46, 55]. Continuous time mean field games with finite state space can be found in [22, 35]. Our previous work [27, 28] studied a class of mean field games in a multi-agent Markov decision process (MDP) framework. The players in [27] have continuum state spaces and binary action spaces, and have coupling through their costs. The state of each player is used to model its risk (or unfitness) level, which has random increase if no active control is taken. Naturally, the one-stage cost of a player is an increasing function of its own state apart from coupling with others. The motivation of this modeling framework comes from applications including network security investment games and flu vaccination games [34, 38, 40]; when the one-stage cost is an increasing function of the population average state, it reflects positive externalities. Markov decision processes with binary action spaces also arise in control of queues and machine replacement problems [4, 10]. Binary choice models have formed a subject of significant interest [8, 15, 48, 49, 54]. Our game model has connection with anonymous sequential games [33], which combine stochastic game modeling with a continuum of players. In anonymous sequential games one determines the equilibrium as a joint state-action distribution of the population and leaves the individual strategies unspecified [33, Sec. 4], although there is an interpretation of randomized actions for players sharing a given state.

For both anonymous games and MDP based mean field games, stationary solutions with discount have been studied in the literature [3, 33]. These works give more focus on fixed point analysis to prove the existence of a stationary distribution. This approach does not address ergodic behavior of individuals or the population while assuming the population starts from the steady-state distribution at the initial time. Thus, there is a need to examine whether the individuals collectively have the ability to move into that distribution at all when they have a general initial distribution. Our ergodic analysis based approach will provide justification of the stationary solution regarding the population's ability to settle down around the limiting distribution.

The previous work [27, 28] studied the finite horizon mean field game by showing existence of a solution with threshold policies, and under an infinite horizon discounted cost further proved there is at most one stationary equilibrium for which existence was not established. A similar continuous time modeling is introduced in [57], which addresses Poisson state jumps and impulse control. It should be noted that except for linear-quadratic models [9, 26, 31, 39, 43], mean field games rarely have closed-form solutions and often rely on heavy numerical computations. Within this context, the consideration of structured solutions, such as threshold policies, is of particular interest from the point of view of efficient computation and simple implementation. Under such a policy, the individual states evolve as regenerative processes [6, 51].

By exploiting stochastic monotonicity, this paper adopts more general state transition assumptions than in [27, 28] and continues the analysis on the stationary equa-

tion system. The first contribution of the present paper is the proof of the existence of a stationary equilibrium. Our analysis depends on checking the continuous dependence of the limiting state distribution on the threshold parameter in the best response. The existence and uniqueness analysis in this paper has appeared in a preliminary form in the conference paper [29].

A key parameter in our game model is the effort cost. Intuitively, this parameter is a disincentive indicator of an individual for taking active efforts, and in turn will further impact the mean field forming the ambient environment of that agent. This suggests that we can study a family of mean field games parametrized by the effort costs and compare their solution behaviors. We address this in the setup of comparative statics, which have a long history in the economic literature [24, 42, 47] and operations research [53] and provide the primary means to analyze the effect of model parameter variations. For dynamic models, such as economic growth models, the analysis follows similar ideas and is sometimes called comparative dynamics [5, 11, 45, 47] by comparing two dynamic equilibria. In control and optimization, such studies are usually called sensitivity analysis [14, 20, 32]. For comparative statics in large static games and mean field games, see [1, 2]. Our analysis is accomplished by performing perturbation analysis around the equilibrium of the mean field game.

The paper is organized as follows. Section 2 introduces the mean field stochastic game. The best response is analyzed in Section 3. Section 4 proves existence and uniqueness of stationary equilibria. Comparative statics are analyzed in Section 5. Section 6 concludes the paper.

2 The Markov Decision Process Model

2.1 Dynamics and Costs

The system consists of N players denoted by \mathcal{A}_i , $1 \leq i \leq N$. At time $t \in \mathbb{Z}_+ = \{0, 1, 2, \dots\}$, the state of \mathcal{A}_i is denoted by x_t^i , and its action by a_t^i . For simplicity, we consider a population of homogeneous (or symmetric) players. Each player has state space $\mathbf{S} = [0, 1]$ and action space $\mathbf{A} = \{a_0, a_1\}$. A value of \mathbf{S} may be interpreted as a risk or unfitness level. A player can either take inaction (as a_0) or make an active effort (as a_1). For an interval I , let $\mathcal{B}(I)$ denote the Borel σ -algebra of I .

The state of each player evolves as a controlled Markov process, which is affected only by its own action. For $t \geq 0$ and $x \in \mathbf{S}$, the state has a transition kernel specified by

$$P(x_{t+1}^i \in B | x_t^i = x, a_t^i = a_0) = Q_0(B|x), \quad (1)$$

$$P(x_{t+1}^i = 0 | x_t^i = x, a_t^i = a_1) = 1, \quad (2)$$

where $Q_0(\cdot|x)$ is a stochastic kernel defined for $B \in \mathcal{B}(\mathbf{S})$ and $Q_0([x, 1]|x) = 1$. By the structure of Q_0 , the state of the player deteriorates if no active control is taken. The vector process (x_t^1, \dots, x_t^N) constitutes a controlled Markov process in higher dimension with its transition kernel defining a product measure on $(\mathcal{B}(\mathbf{S}))^N$ for given $(x_t^1, \dots, x_t^N, a_t^1, \dots, a_t^N)$.

Define the population average state $x_t^{(N)} = \frac{1}{N} \sum_{i=1}^N x_t^i$. The one stage cost of \mathcal{A}_i is

$$c(x_t^i, x_t^{(N)}, a_t^i) = R(x_t^i, x_t^{(N)}) + \gamma 1_{\{a_t^i = a_1\}},$$

where $\gamma > 0$ and $\gamma 1_{\{a_t^i = a_1\}}$ is the effort cost. The function $R \geq 0$ is defined on $\mathbf{S} \times \mathbf{S}$ and models the risk-related cost. Let v^i denote the strategy of \mathcal{A}_i . We introduce the infinite horizon discounted cost

$$J_i(x_0^1, \dots, x_0^N, v^1, \dots, v^N) = E \sum_{t=0}^{\infty} \beta^t c(x_t^i, x_t^{(N)}, a_t^i), \quad 1 \leq i \leq N. \quad (3)$$

The standard methodology of mean field games may be applied by approximating $\{x_t^{(N)}, t \geq 0\}$ by a deterministic sequence $\{z_t, t \geq 0\}$ which depends on the initial condition of the system. One may solve the limiting optimal control problem of \mathcal{A}_i and derive a dynamic programming equation for its value function denoted by $v_i(t, x, (z_k)_{k=0}^{\infty})$, whose dependence on t is due to the time-varying sequence $\{z_t, t \geq 0\}$. Subsequently one derives another equation for the mean field $\{z_t, t \geq 0\}$ by averaging the individual states across the population. This approach, however, has the drawback of heavy computational load.

2.2 Stationary Equilibrium

We are interested in a steady-state form of the solution of the mean field game starting with $\{z_t, t \geq 0\}$. Such steady state equations provide information on the long time behavior of the solution and are of interest in their own right. They may also be used for approximation purposes to compute strategies efficiently. We introduce the system

$$v(x) = \min \left[\beta \int_0^1 v(y) Q_0(dy|x) + R(x, z), \quad \beta v(0) + R(x, z) + \gamma \right], \quad (4)$$

$$z = \int_0^1 x \mu(dx), \quad (5)$$

where μ is a probability measure on \mathbf{S} . We say $(v, z, \mu, a^i(\cdot))$ is a *stationary equilibrium* to (4)-(5) if i) the feedback policy $a^i(\cdot)$, as a mapping from \mathbf{S} to $\{a_0, a_1\}$, is the best response with respect to z in (4), ii) given an initial distribution of $x_0^i, \{x_t^i, t \geq 0\}$ under the policy a^i has its distribution converging (under a total variation norm or only weakly) to the stationary distribution (or called limiting distribution) μ .

We may interpret v as the value function of an MDP with cost $\bar{J}_i(x_0^i, z, v^i) = E \sum_{t=0}^\infty \beta^t c(x_t^i, z, a_t^i)$. An alternative way to interpret (4)-(5) is that the initial state of \mathcal{A}_i has been sampled according to the “right” distribution μ , and that z is obtained by averaging an infinite number of such initial values by the law of large numbers [52]. A similar solution notion is adopted in [2, 3] but ergodicity is not part of their solution specification.

Let the probability measure μ_k be the distribution of \mathbb{R} -valued random variable Z_k , $k = 1, 2$. We say μ_2 stochastically dominates μ_1 , and denote $\mu_1 \leq_{st} \mu_2$, if $\mu_2((y, \infty)) \geq \mu_1((y, \infty))$ (or equivalently, $P(Z_2 > y) \geq P(Z_1 > y)$) for all y . It is well known [44] that $\mu_1 \leq_{st} \mu_2$ if and only if

$$\int \psi(y) \mu_1(dy) \leq \int \psi(y) \mu_2(dy) \tag{6}$$

for all increasing function ψ (not necessarily strictly increasing) for which the two integrals are finite. A stochastic kernel $\mathcal{Q}(B|x)$, $0 \leq x \leq 1$, $B \in \mathcal{B}(\mathbf{S})$, is said to be strictly stochastically increasing if $\varphi(x) := \int_{\mathbf{S}} \psi(y) \mathcal{Q}(dy|x)$ is strictly increasing in $x \in \mathbf{S}$ for any strictly increasing function $\psi : [0, 1] \rightarrow \mathbb{R}$ for which the integral is necessarily finite. $\mathcal{Q}(\cdot|x)$ is said to be weakly continuous if φ is continuous whenever ψ is continuous.

Let $\{Y_t, t \geq 0\}$ be a Markov process with state space $[0, 1]$, transition kernel $Q_0(\cdot|x)$ and initial state $Y_0 = 0$. So each of its trajectories is monotonically increasing. Define $\tau_{Q_0}^\theta = \inf\{t|Y_t \geq \theta\}$ for $\theta \in (0, 1)$. It is clear that $\tau_{Q_0}^{\theta_1} \leq \tau_{Q_0}^{\theta_2}$ for $0 < \theta_1 < \theta_2 < 1$.

The following assumptions are introduced.

- (A1) $\{x_i^j, i \geq 1\}$ are i.i.d. random variables taking values in \mathbf{S} .
- (A2) $R(x, z)$ is a continuous function on $\mathbf{S} \times \mathbf{S}$. For each fixed z , $R(\cdot, z)$ is strictly increasing.
- (A3) i) $Q_0(\cdot|x)$ satisfies $Q_0([x, 1]|x) = 1$ for any x , and is strictly stochastically increasing; ii) $Q_0(dy|x)$ is weakly continuous and has a positive probability density $q(y|x)$ for each fixed $x < 1$; iii) for any small $0 < \delta < 1$, $\inf_x Q_0([1 - \delta, 1]|x) > 0$.
- (A4) $R(x, \cdot)$ is increasing for each fixed x .
- (A5) $\lim_{\theta \uparrow 1} E \tau_{Q_0}^\theta = \infty$.

(A3)-iii) will be used to ensure the uniform ergodicity of the controlled Markov process. In fact, under (A3) we can show $E \tau_{Q_0}^\theta < \infty$. The following condition is a special case of (A3).

- (A3') There exists a random variable such that $Q_0(\cdot|x)$ is equal to the law of $x + (x - 1)\xi$ for some random variable ξ with probability density $f_\xi(x) > 0$, a.e. $x \in \mathbf{S}$.

When (A3') holds, we can verify (A5) by analyzing the stopping time $\tau_\xi = \inf\{t|\prod_{s=1}^t \xi_s \leq 1 - \theta\}$, where $\{\xi_s, s \geq 1\}$ is a sequence of i.i.d. random variables with probability density f_ξ . For existence analysis of the mean field game, (A5) will be used to ensure continuity of the mean field when the threshold θ approaches 1.

Proposition 1 *The two conditions are equivalent:*

- i) $\mu_1 \leq_{st} \mu_2$, and $\mu_1 \neq \mu_2$;
- ii) $\int_{\mathbb{R}} \phi(y)\mu_1(dy) < \int_{\mathbb{R}} \phi(y)\mu_2(dy)$ for all strictly increasing function ϕ for which both integrals are finite.

Proof. Assume i) holds. By [44, Theorem 1.2.16], we have

$$\phi(Z_1) \leq_{st} \phi(Z_2), \tag{7}$$

and so $E\phi(Z_1) \leq E\phi(Z_2)$. Since $\mu_1 \neq \mu_2$, there exists y_0 such that $P(Z_1 > y_0) \neq P(Z_2 > y_0)$. Take r such that $\phi(y_0) = r$. Then

$$P(\phi(Z_1) > r) \neq P(\phi(Z_2) > r). \tag{8}$$

If $E\phi(Z_1) = E\phi(Z_2)$ were true, by (7) and [44, Theorem 1.2.9], $\phi(Z_1)$ and $\phi(Z_2)$ would have the same distribution, which contradicts (8). We conclude $E\phi(Z_1) < E\phi(Z_2)$, which is equivalent to ii).

Next we show ii) implies i). Let ψ be any increasing function satisfying (6) with two finite integrals. When ii) holds, we take $\phi_\varepsilon = \psi + \frac{\varepsilon y}{1+|y|}$, $\varepsilon > 0$. Then $\int \phi_\varepsilon \mu_1(dy) < \int \phi_\varepsilon \mu_2(dy)$ holds for all $\varepsilon > 0$. Letting $\varepsilon \rightarrow 0$, then (6) follows and $\mu_1 \leq_{st} \mu_2$. It is clear $\mu_1 \neq \mu_2$. \square

3 Best Response

For this section we assume (A1)-(A3). We take any fixed $z \in [0, 1]$ and consider (4) as a separate equation, which is rewritten below:

$$v(x) = \min \left\{ \beta \int_0^1 v(y)Q_0(dy|x) + R(x,z), \quad \beta v(0) + R(x,z) + \gamma \right\}. \tag{9}$$

Here z is not required to satisfy (5). In relation to the mean field game, the resulting optimal policy will be called the best response with respect to z . Denote $G(x) = \int_0^1 v(y)Q_0(dy|x)$.

Lemma 1. *i) Equation (9) has a unique solution $v \in C([0, 1], \mathbb{R})$.*

ii) v is strictly increasing.

iii) The optimal policy is determined as follows:

- a) If $\beta G(1) < \beta v(0) + \gamma$, $a^i(x) \equiv a_0$.
- b) If $\beta G(1) = \beta v(0) + \gamma$, $a^i(1) = a_1$ and $a^i(x) = a_0$ for $x < 1$.
- c) If $\beta G(0) \geq \beta v(0) + \gamma$, $a^i(x) \equiv a_1$.
- d) If $\beta G(0) < \beta v(0) + \gamma < \rho G(1)$, there exists a unique $x^* \in (0, 1)$ and a^i is a threshold policy with parameter x^* , i.e., $a^i(x) = a_1$ if $x \geq x^*$ and $a^i(x) = a_0$ if $x < x^*$.

Proof. Define the dynamic programming operator

$$(\mathcal{L}g)(x) = \min \left\{ \beta \int_0^1 g(y) Q_0(dy|x) + R(x, z), \quad \beta g(0) + R(x, z) + \gamma \right\}, \quad (10)$$

which is from $C([0, 1], \mathbb{R})$ to itself. The proving method in [27], [28, Lemma 6], which assumed (A3'), can be extended to the present equation (9) in a straightforward manner.

In particular, for the proof of ii) and iii), we obtain progressively stronger properties of v and G . First, denoting $g_0 = 0$ and $g_{k+1} = \mathcal{L}g_k$ for $k \geq 0$, we use a successive approximation procedure to show that v is increasing, which implies that G is continuous and increasing by weak continuity and monotonicity of Q_0 . Since R is strictly increasing in x , by the right hand side of (9), we show that v is strictly increasing, which implies the same property for G by strict monotonicity of Q_0 . \square

For the optimal policy specified in part iii) of Lemma 1, we can formally denote the threshold parameters for the corresponding cases: a) $\theta = 1^+$, b) $\theta = 1$, c) $\theta = 0$, and d) $\theta = x^*$. Such a policy will be called a θ -threshold policy. We give the condition for $\theta = 0$ in the best response.

Lemma 2. For $\gamma > 0$ and v solving (9),

$$\beta G(0) \geq \beta v(0) + \gamma \quad (11)$$

holds if and only if

$$\gamma \leq \beta \int_0^1 R(y, z) Q_0(dy|0) - \beta R(0, z). \quad (12)$$

Proof. We show necessity first. Suppose (11) holds. Note that $G(x)$ is strictly increasing on $[0, 1]$. Equation (9) reduces to

$$v(x) = \beta v(0) + R(x, z) + \gamma, \quad (13)$$

$$\beta G(x) \geq \beta v(0) + \gamma, \quad \forall x. \quad (14)$$

From (13), we uniquely solve

$$v(0) = \frac{1}{1-\beta} [R(0, z) + \gamma], \quad v(x) = \frac{\beta}{1-\beta} [R(0, z) + \gamma] + R(x, z) + \gamma, \quad (15)$$

which combined with (14) implies (12).

We continue to show sufficiency. If $\gamma > 0$ satisfies (12), we use (15) to construct v and verify (13) and (14). So v is the unique solution of (9) satisfying (11). \square

The next lemma gives the condition for $\theta = 1^+$ in the best response.

Lemma 3. For $\gamma > 0$ and v solving (9), we have

$$\beta G(1) < \beta v(0) + \gamma \quad (16)$$

if and only if

$$\gamma > \beta[V_\beta(1) - V_\beta(0)], \tag{17}$$

where $V_\beta(x) \in C([0, 1], \mathbb{R})$ is the unique solution of

$$V_\beta(x) = \beta \int_0^1 V_\beta(y)Q_0(dy|x) + R(x, z). \tag{18}$$

Proof. By Banach’s fixed point theorem, we can show that (18) has a unique solution. Next, by a successive approximation $\{V_\beta^{(k)}, k \geq 0\}$ with $V_\beta^{(0)} = 0$ in the fixed point equation, we can further show that V_β is strictly increasing. Moreover, $\int_0^1 V_\beta(y)Q_0(dy|x)$ is increasing in x by monotonicity of Q_0 .

We show necessity. Since G is strictly increasing, (16) implies that the right hand side of (9) now reduces to the first term within the parentheses and that $v = V_\beta$. So (17) follows.

To show sufficiency, suppose (17) holds. We have

$$\beta \int_0^1 V_\beta(y)Q_0(dy|x) \leq \beta V_\beta(1) < \beta V_\beta(0) + \gamma, \quad \forall x.$$

Therefore, $v := V_\beta$ gives the unique solution of (9) and $\beta G(1) < \beta v(0) + \gamma$. \square

Example 1. Let $R(x, z) = x(c + z)$, where $c > 0$. Take $Q_0(\cdot|x)$ as uniform distribution on $[x, 1]$. Then (18) reduces to

$$V_\beta(x) = \frac{\beta}{1-x} \int_x^1 V_\beta(y)dy + R(x, z).$$

Define $\phi(x) = \int_x^1 V_\beta(y)dy$, $x \in [0, 1]$. Then $\phi'(x) = -\frac{\beta}{1-x}\phi(x) - R(x, z)$ holds and we solve

$$\phi(x) = (1-x)^\beta \int_x^1 \frac{R(s, z)}{(1-s)^\beta} ds,$$

where the right hand side converges to 0 as $x \rightarrow 1^-$. We further obtain

$$V_\beta(x) = \beta(1-x)^{\beta-1} \int_x^1 \frac{R(s, z)}{(1-s)^\beta} ds + R(x, z)$$

for $x \in [0, 1)$, and the right hand side has the limit $\frac{R(1, z)}{1-\beta}$ as $x \rightarrow 1^-$. This gives a well defined $V_\beta \in C([0, 1], \mathbb{R})$. Therefore, $V_\beta(0) = \frac{\beta(c+z)}{(1-\beta)(2-\beta)}$. Then (17) reduces to $\gamma > \frac{2\beta(c+z)}{2-\beta}$.

4 Existence of Stationary Equilibria

Assume (A1)-(A5) for this section. Define the class \mathcal{P}_0 of probability measures on \mathbf{S} as follows: $\nu \in \mathcal{P}_0$ if there exist a constant $c_\nu \geq 0$ and a Borel measurable function $g(x) \geq 0$ defined on $[0, 1]$ such that

$$\nu(B) = \int_B g(x)dx + c_\nu 1_B(0),$$

where $B \in \mathcal{B}(\mathbf{S})$ and 1_B is the indicator function of B . When restricted to $(0, 1]$, ν is absolutely continuous with respect to the Lebesgue measure μ^{Leb} .

Let X be a random variable with distribution $\nu \in \mathcal{P}_0$. Set $x_t^i = X$. Define $Y_0 = x_{t+1}^i$ by applying $a_t^i \equiv a_0$. Further define $Y_1 = x_{t+1}^i$ by applying the r -threshold policy a_t^i with $r \in (0, 1)$.

Lemma 4. *The distribution ν_i of Y_i is in \mathcal{P}_0 for $i = 0, 1$.*

Proof. Let $q(y|x)$ denote the density function of $Q_0(\cdot|x)$ for $x \in [0, 1)$, where $q(y|x) = 0$ for $y < x$. Denote

$$g_0(y) = \int_{0 \leq x < y} q(y|x)\nu(dx), \quad y \in (0, 1),$$

and

$$g_1(y) = \int_{0 \leq x < y \wedge r} q(y|x)\nu(dx), \quad y \in (0, 1).$$

Then it can be checked that

$$P(Y_0 \in B) = \int_B g_0(y)dy, \quad P(Y_1 \in B) = \int_B g_1(y)dy + P(X \geq r)1_B(0).$$

This completes the lemma. \square

In order to show that (4)-(5) has a solution, we define a mapping $\Gamma: \mathbf{S} \rightarrow \mathbf{S}$ by the following rule. For $z \in [0, 1]$, we solve (4) to obtain a well defined threshold $\theta(z) \in [0, 1] \cup \{1^+\}$, which in turn determines a limiting distribution $\mu_{\theta(z)}$ of the closed-loop state process x_t^i by Lemma A.1. Define

$$\Gamma(z) = \int_0^1 x\mu_{\theta(z)}(dx).$$

If Γ has a fixed point, we obtain a solution to (4)-(5).

We analyze the case where the best response gives a strictly positive threshold. Assume

$$\gamma > \beta \max_{z \in [0, 1]} \int_0^1 [R(y, z) - R(0, z)]Q_0(dy|0). \tag{19}$$

Note that under a zero threshold policy, the behavior of the state process is sensitive to a positive perturbation of the threshold. The above condition ensures that the zero threshold will not occur, and this will ensure continuity of Γ to facilitate the fixed point analysis.

Lemma 5. *Assume (19). Then $\Gamma(z)$ is continuous on $[0, 1]$.*

Proof. Let $z_0 \in [0, 1]$ be fixed, giving a corresponding threshold parameter θ_0 when (9) is solved using z_0 . We check continuity at z_0 and consider 3 cases.

Case i) $\theta_0 \in (0, 1)$. Let π_0 be the stationary distribution with the θ_0 -threshold policy. Consider any fixed $\varepsilon > 0$. There exists ε_1 such that for all $\theta \in (\theta_0 - \varepsilon_1, \theta_0 + \varepsilon_1) \subset (0, 1)$, $|\int_0^1 x\pi(dx) - \int_0^1 x\pi_0(dx)| < \varepsilon$, where π is the stationary distribution associated with θ . This follows since $\lim_{\theta \rightarrow \theta_0} \|\pi - \pi_0\|_{TV} = 0$ by Lemma A.3. Now by the continuous dependence of the solution of the dynamic programming equation on z , we can select a sufficiently small $\delta > 0$ such that for all $|z - z_0| < \delta$, z generates a threshold parameter $\theta \in (\theta_0 - \varepsilon_1, \theta_0 + \varepsilon_1)$, which implies $|\Gamma(z) - \Gamma(z_0)| \leq \varepsilon$.

Case ii) z_0 gives $\theta_0 = 1$. Then $\Gamma(z_0) = 1$. Fix any $\varepsilon > 0$. Then we can show there exists ε_1 such that for all $\theta \in (1 - \varepsilon_1, 1)$, the associated stationary distribution π_θ gives $|\Gamma(z_0) - \int_0^1 x\pi_\theta(dx)| < \varepsilon$, where we use (A5) and the right hand side of (C.1) to estimate a lower bound for $\int_0^1 x\pi_\theta(dx)$. Now, there exists $\delta > 0$ such that any z satisfying $|z - z_0| < \delta$ gives a threshold θ either in $(1 - \varepsilon_1, 1)$ or equal to 1 or 1^+ ; for each case, we have $|\Gamma(z_0) - \int_0^1 x\pi_\theta(dx)| < \varepsilon$.

Case iii) z_0 gives $\theta_0 = 1^+$. Then there exists $\delta > 0$ such that any z satisfying $|z - z_0| < \delta$ gives a threshold parameter $\theta = 1^+$. Then $\Gamma(z) = \Gamma(z_0) = 1$. \square

Theorem 1. *Assume (19). There exists a stationary equilibrium to (4)-(5).*

Proof. Since Γ is a continuous function from $[0, 1]$ to $[0, 1]$ by Lemma 5, the theorem follows from Brouwer’s fixed point theorem. \square

Let $x_t^{i,\theta}$ and π_θ denote the state process and its stationary distribution, respectively, under a θ -threshold policy. Denote $z(\theta) = \int_0^1 x\pi_\theta(dx)$. We have the first comparison theorem on monotonicity.

Lemma 6. $z(\theta_1) \leq z(\theta_2)$ for $0 < \theta_1 < \theta_2 < 1$.

Proof. By the ergodicity of $\{x_t^{i,\theta}, t \geq 0\}$ in Lemma A.2, we have the representation $z(\theta_1) = \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{t=0}^{k-1} x_t^{i,\theta_1}$ w.p.1. Lemma C.2 implies $z(\theta_1) \leq z(\theta_2)$. \square

To establish uniqueness, we consider $R(x, z) = R_1(x)R_2(z)$, where $R_1 \geq 0$ and $R_2 \geq 0$, and which satisfies (A1)-(A5). We further make the following assumption.

(A6) $R_2 > 0$ is strictly increasing on \mathbf{S} .

This assumption indicates positive externalities since an individual benefits from the decrease of the population average state. This condition has a crucial role in the uniqueness analysis.

Given the product form of R , now (9) takes the form:

$$V(x) = \min \left[\beta \int_0^1 V(y) \mathcal{Q}_0(dy|x) + R_1(x)R_2(z), \quad \beta V(0) + R_1(x)R_2(z) + \gamma \right].$$

Consider $0 \leq z_2 < z_1 \leq 1$ and

$$V_l(x) = \min \left[\beta \int_0^1 V_l(y) \mathcal{Q}_0(dy|x) + R_1(x)R_2(z_l), \quad \beta V_l(0) + R_1(x)R_2(z_l) + \gamma \right]. \tag{20}$$

Denote the optimal policy as a threshold policy with parameter θ_l in $[0, 1]$ or equal to 1^+ , where we follow the interpretation in Section 3 if $\theta_l = 1^+$. We state the second comparison theorem about the threshold parameters under different mean field parameters z_l .

Theorem 2. θ_1 and θ_2 in (20) are specified according to the following scenarios:

- i) If $\theta_1 = 0$, then we have either $\theta_2 \in [0, 1]$ or $\theta_2 = 1^+$.
- ii) If $\theta_1 \in (0, 1)$, we have either a) $\theta_2 \in (\theta_1, 1)$, or b) $\theta_2 = 1$, or c) $\theta_2 = 1^+$.
- iii) If $\theta_1 = 1$, $\theta_2 = 1^+$.
- iv) If $\theta_1 = 1^+$, $\theta_2 = 1^+$.

Proof. Since $R_2(z_1) > R_2(z_2) > 0$, we divide both sides of (20) by $R_2(z_l)$ and define $\gamma_l = \frac{\gamma}{R_2(z_l)}$. Then $0 < \gamma_1 < \gamma_2$. The dynamic programming equation reduces to (D.2). Subsequently, the optimal policy is determined according to Lemma D.4. \square

Corollary 1. Assume (A6) in addition to the assumptions in Theorem 1. Then the system (4)-(5) has a unique stationary equilibrium.

Proof. The proof is similar to [27, 28], which assumed (A3'). \square

5 Comparative Statics

This section assumes (A1)-A(6). Consider the two solution systems

$$\begin{cases} \bar{v}(x) = \min \left[\beta \int_0^1 \bar{v}(y) \mathcal{Q}_0(dy|x) + R_1(x)R_2(\bar{z}), \quad \beta \bar{v}(0) + R_1(x)R_2(\bar{z}) + \bar{\gamma} \right], \\ \bar{z} = \int_0^1 x \bar{\mu}(dx), \end{cases} \tag{21}$$

and

$$\begin{cases} v(x) = \min \left[\beta \int_0^1 v(y) \mathcal{Q}_0(dy|x) + R_1(x)R_2(z), \quad \beta v(0) + R_1(x)R_2(z) + \gamma \right], \\ z = \int_0^1 x \mu(dx). \end{cases} \tag{22}$$

Suppose $\bar{\gamma}$ satisfies (19). By Corollary 1, (21) has a unique solution denoted by $(\bar{v}, \bar{z}, \bar{\mu}, \bar{\theta})$, where $\bar{\theta}$ is the threshold parameter. We further assume $\bar{\theta} \in (0, 1)$. Suppose $\gamma > \bar{\gamma}$. Then we can uniquely solve (v, z, μ, θ) . The next theorem presents a result on monotone comparative statics [53].

Theorem 3. *If $\gamma > \bar{\gamma}$, we have*

$$\theta > \bar{\theta}, \quad z > \bar{z}, \quad v > \bar{v}.$$

Proof. We prove by contraction. Assume $\theta \leq \bar{\theta}$. Then by Lemma 6, $z \leq \bar{z}$, and therefore, $\frac{\gamma}{R_2(z)} > \frac{\bar{\gamma}}{R_2(\bar{z})}$. By the method of proving Theorem 2, we would establish $\theta > \bar{\theta}$, which contradicts the assumption $\theta \leq \bar{\theta}$. We conclude $\theta > \bar{\theta}$. By Lemma 6 and Remark B.1, we have $z > \bar{z}$. For (21), we use value iteration to approximate \bar{v} by an increasing sequence of functions \bar{v}_k with $\bar{v}_0 = 0$. Similarly, v is approximated by v_k with $v_0 = 0$. By induction, we have $v_k \geq \bar{v}_k$ for all k . This proves $v \geq \bar{v}$.

Next, we have $\beta v(0) + R_1(x)R_2(z) + \gamma > \beta \bar{v}(0) + R_1(x)R_2(\bar{z}) + \bar{\gamma}$ on $[0, 1]$, and $\beta \int_0^1 v(y)Q_0(dy|x) + R_1(x)R_2(z) > \beta \int_0^1 \bar{v}(y)Q_0(dy|x) + R_1(x)R_2(\bar{z})$ on $(0, 1]$. By the method in [27, Lemma 2], we have $v > \bar{v}$ on $(0, 1]$. Then $\int_0^1 v(y)Q_0(dy|0) > \int_0^1 \bar{v}(y)Q_0(dy|0)$. This further implies $v(0) > \bar{v}(0)$. \square

Remark 1. It is possible to have $\theta = 1^+$ in Theorem 3.

By a continuity argument, we can further show $\lim_{\gamma \rightarrow \bar{\gamma}} (|\theta - \bar{\theta}| + |z - \bar{z}| + \sup_x |v(x) - \bar{v}(x)|) = 0$. In the analysis below, we take $\gamma = \bar{\gamma} + \varepsilon$ for some small $\varepsilon > 0$. For this section, we further introduce the following assumption.

(A7) For $\gamma > \bar{\gamma}$, (v, z, θ) has the representation

$$v(x) = \bar{v}(x) + \varepsilon w(x) + o(\varepsilon), \quad 0 \leq x \leq 1, \tag{23}$$

$$z = \bar{z} + \varepsilon z_\gamma + o(\varepsilon), \tag{24}$$

$$\theta = \bar{\theta} + \varepsilon \theta_\gamma + o(\varepsilon), \tag{25}$$

where v, z, θ are solved depending on the parameter γ and w is a function defined on $[0, 1]$. The derivatives z_γ and θ_γ at $\bar{\gamma}$ exist, and $R_2(z)$ is differentiable on $[0, 1]$. For $0 \leq x < 1$, the probability density function $q(y|x)$, $y \in [x, 1]$, for $Q_0(dy|x)$ is continuous on $\{(x, y) | 0 \leq x \leq y < 1\}$. Moreover, $\frac{\partial q(y|x)}{\partial x}$ exists and is continuous in (x, y) .

We aim to provide a characterization of $w, z_\gamma, \theta_\gamma$.

Theorem 4. *The function w satisfies*

$$w(x) = \begin{cases} \beta \int_0^1 w(y)Q_0(dy|x) + R_1(x)R_2'(\bar{z})z_\gamma, & 0 \leq x \leq \bar{\theta}, \\ \beta w(0) + R_1(x)R_2'(\bar{z})z_\gamma + 1, & \bar{\theta} < x \leq 1. \end{cases} \tag{26}$$

Proof. We have

$$\bar{v}(x) = \beta \int_0^1 \bar{v}(y)Q_0(dy|x) + R_1(x)R_2(\bar{z}), \quad x \in [0, \bar{\theta}]$$

and

$$v(x) = \beta \int_0^1 v(y)Q_0(dy|x) + R_1(x)R_2(z), \quad x \in [0, \theta].$$

Note that $\theta > \bar{\theta}$. For any fixed $x \in [0, \bar{\theta}]$, we have

$$v(x) - \bar{v}(x) = \beta \int_0^1 (v(y) - \bar{v}(y))Q_0(dy|x) + R_1(x)(R_2(z) - R_2(\bar{z})).$$

Then the equation of $w(x)$ for $x \in [0, \bar{\theta}]$ is derived. We similarly treat the case $x \in (\bar{\theta}, 1]$. \square

Remark 2. In general w has discontinuity at $x = \bar{\theta}$, so that $\beta \int_0^1 w(y)Q_0(dy|\bar{\theta}) \neq \beta w(0) + 1$. We give some interpretation. Let the value function be written as $v(x, \gamma)$ to explicitly indicate γ . Let the rectangle $[0, 1] \times [\gamma_a, \gamma_b]$ be a region of interest in which (x, γ) varies so that the value function defines a continuous surface. Then (θ, γ) starts at $(\bar{\theta}, \bar{\gamma})$ and traces out the curve of an increasing function along which the expression of the value function has a switch, and the value function surface may be visualized as two pieces glued together along the curve in a non-smooth way. The value of w amounts to finding on the surface the directional derivative in the direction of γ ; and therefore, discontinuity may occur at $x = \bar{\theta}$.

To better understand the solution of (26), we consider the general equation

$$W(x) = \begin{cases} \beta \int_0^1 W(y)Q_0(dy|x) + R_1(x)R'_2(z_0)c_0, & 0 \leq x \leq \theta_0, \\ \beta W(0) + R_1(x)R'_2(z_0)c_0 + 1, & \theta_0 < x \leq 1, \end{cases} \quad (27)$$

where $c_0, z_0 \in [0, 1]$ and $\theta_0 \in (0, 1)$ are arbitrarily chosen and fixed. Let $B([0, 1], \mathbb{R})$ be the Banach space of bounded Borel measurable functions with norm $\|g\| = \sup_x |g(x)|$. By a contraction mapping, we can show (27) has a unique solution $W \in B([0, 1], \mathbb{R})$.

We continue to characterize the sensitivity θ_γ of the threshold. Recall the partial derivative $\frac{\partial q(y|x)}{\partial x}$.

Lemma 7. *We have*

$$\beta \left[\int_{\bar{\theta}}^1 \bar{v}(y) \frac{\partial q(y|\bar{\theta})}{\partial x} dy - \bar{v}(\bar{\theta})q(\bar{\theta}|\bar{\theta}) \right] \theta_\gamma = 1 + \beta w(0) - \beta \int_{\bar{\theta}}^1 w(y)Q_0(dy|\bar{\theta}). \quad (28)$$

Proof. Write $\gamma = \bar{\gamma} + \varepsilon$. By the property of the threshold, we have

$$\beta \int_{\bar{\theta}}^1 \bar{v}(y)Q_0(dy|\bar{\theta}) = \beta \bar{v}(0) + \bar{\gamma}, \quad \beta \int_{\theta}^1 v(y)Q_0(dy|\theta) = \beta v(0) + \bar{\gamma} + \varepsilon.$$

Note that $\theta > \bar{\theta}$. We check

$$\begin{aligned} \Delta &:= \int_{\theta}^1 v(y)Q_0(dy|\theta) - \int_{\bar{\theta}}^1 \bar{v}(y)Q_0(dy|\bar{\theta}) \\ &= \int_{\theta}^1 v(y)Q_0(dy|\theta) - \int_{\theta}^1 \bar{v}(y)Q_0(dy|\bar{\theta}) - \int_{\bar{\theta}}^{\theta} \bar{v}(y)Q_0(dy|\bar{\theta}) \\ &= \int_{\theta}^1 v(y)Q_0(dy|\theta) - \int_{\theta}^1 \bar{v}(y)Q_0(dy|\theta) \\ &\quad + \int_{\theta}^1 \bar{v}(y)Q_0(dy|\theta) - \int_{\theta}^1 \bar{v}(y)Q_0(dy|\bar{\theta}) - \int_{\bar{\theta}}^{\theta} \bar{v}(y)Q_0(dy|\bar{\theta}) \\ &= \varepsilon \int_{\theta}^1 w(y)q(y|\theta)dy + (\theta - \bar{\theta}) \int_{\theta}^1 \bar{v}(y)[\partial q(y|\theta)/\partial x]dy - (\theta - \bar{\theta})\bar{v}(\bar{\theta})q(\bar{\theta}|\bar{\theta}) \\ &\quad + o(\varepsilon + |\theta - \bar{\theta}|) \\ &= \varepsilon \int_{\bar{\theta}}^1 w(y)q(y|\bar{\theta})dy + (\theta - \bar{\theta}) \int_{\bar{\theta}}^1 \bar{v}(y)[\partial q(y|\bar{\theta})/\partial x]dy - (\theta - \bar{\theta})\bar{v}(\bar{\theta})q(\bar{\theta}|\bar{\theta}) \\ &\quad + o(\varepsilon + |\theta - \bar{\theta}|). \end{aligned}$$

Note that

$$\beta\Delta = \beta[v(0) - \bar{v}(0)] + \varepsilon.$$

We derive

$$\beta \int_{\bar{\theta}}^1 w(y)Q_0(dy|\bar{\theta}) + \beta\theta_{\gamma} \int_{\bar{\theta}}^1 \bar{v}(y) \frac{\partial q(y|\bar{\theta})}{\partial x} dy - \beta\bar{v}(\bar{\theta})q(\bar{\theta}|\bar{\theta})\theta_{\gamma} = \beta w(0) + 1.$$

This completes the proof. \square

Lemma 8. *Given the threshold $\bar{\theta} \in (0, 1)$, the stationary distribution $\bar{\mu}$ has a probability density function (p.d.f.) $p(x)$ on $(0, 1]$, and $\bar{\mu}(\{0\}) = \pi_0$, where (p, π_0) is determined by*

$$\pi_0 = \int_{\bar{\theta}}^1 p(x)dx, \tag{29}$$

$$p(x) = \begin{cases} \int_0^x q(x|y)p(y)dy + \pi_0q(x|0), & 0 \leq x < \bar{\theta}, \\ \int_0^{\bar{\theta}} q(x|y)p(y)dy + \pi_0q(x|0), & \bar{\theta} \leq x \leq 1. \end{cases} \tag{30}$$

Proof. Let δ_0 be the dirac measure at $x = 0$. For any Borel subset $B \subset [0, 1]$, we have $\bar{\mu}(B) = \int_0^1 [Q_0(B|y)1_{(y < \bar{\theta})} + \delta_0(B)1_{(y \geq \bar{\theta})}] \bar{\mu}(dy)$. Then it can be checked that (p, π_0) satisfying the above equations determines the stationary distribution. Now we show there exists a unique solution. Let $\pi_0 > 0$ be a constant to be determined. Consider the Volterra integral equation

$$p(x) = \int_0^x q(x|y)p(y)dy + \pi_0q(x|0), \quad 0 \leq x \leq \bar{\theta}, \tag{31}$$

and we obtain a unique solution p in $C([0, \bar{\theta}], \mathbb{R})$ (see e.g. [36, p.33]). In fact p is a nonnegative function with $\int_0^{\bar{\theta}} p(x)dx > 0$. Subsequently, we further determine $p \geq 0$ on $[\bar{\theta}, 1]$ by (30). The solution p on $[0, 1]$ depends linearly on π_0 and so there exists a unique π_0 such that $\int_0^1 p(x)dx + \pi_0 = 1$. After we uniquely solve p for (30), we integrate both sides of this equation on $[0, 1]$ and obtain $\int_0^1 p(x)dx = \int_0^{\bar{\theta}} p(x)dx + \pi_0$, which implies that (29) is satisfied. \square

5.1 Special Case

Now we suppose $Q_0(dy|x)$ has uniform distribution on $[x, 1]$ for all fixed $0 \leq x < 1$, and $R(x, z) = R_1(x)R_2(z) = x(c+z)$, where $R_1(x) = x$, $R_2(z) = c+z$ and $c > 0$. In this case, (A2)-(A6) are satisfied. For (21), we have

$$\bar{v}(x) = \begin{cases} \frac{\beta}{1-x} \int_x^1 \bar{v}(y)dy + R_1(x)R_2(\bar{z}), & 0 \leq x \leq \bar{\theta}, \\ \beta \bar{v}(0) + R_1(x)R_2(\bar{z}) + \bar{\gamma}, & \bar{\theta} \leq x \leq 1. \end{cases} \quad (32)$$

Denote $\varphi(x) = \int_x^1 \bar{v}(y)dy$. Then

$$\dot{\varphi}(x) = -\frac{\beta}{1-x}\varphi - R_1(x)R_2(\bar{z}), \quad 0 \leq x \leq \bar{\theta}.$$

Taking the initial condition $\varphi(0)$, we have

$$\varphi(x) = \varphi(0)(1-x)^\beta - (1-x)^\beta \int_0^x \frac{R_1(\tau)R_2(\bar{z})}{(1-\tau)^\beta} d\tau.$$

On $[0, \bar{\theta}]$,

$$\begin{aligned} \bar{v}(x) &= (1-x)^{\beta-1}\bar{v}(0) - \beta(1-x)^{\beta-1} \int_0^x \frac{R_1(\tau)R_2(\bar{z})}{(1-\tau)^\beta} d\tau + R_1(x)R_2(\bar{z}) \\ &= (1-x)^{\beta-1} \left[\bar{v}(0) - \frac{\beta(c+\bar{z})}{(1-\beta)(2-\beta)} \right] + (c+\bar{z}) \left[\frac{\beta}{(1-\beta)(2-\beta)} + \frac{2x}{2-\beta} \right]. \end{aligned}$$

By the continuity of \bar{v} and its form on $[\bar{\theta}, 1]$, we have

$$\bar{v}(\bar{\theta}) = \beta \bar{v}(0) + \bar{\theta}(\bar{z} + c) + \bar{\gamma}. \quad (33)$$

Hence,

$$[(1-\bar{\theta})^{\beta-1} - \beta]\bar{v}(0) = \frac{\beta(c+\bar{z})[(1-\bar{\theta})^{\beta-1} - 1]}{(1-\beta)(2-\beta)} - \frac{\beta(c+\bar{z})\bar{\theta}}{2-\beta} + \bar{\gamma}. \quad (34)$$

On the other hand, since \bar{v} is increasing and $\bar{\theta}$ is the threshold, we have

$$\begin{aligned} \bar{v}(\bar{\theta}) &= \beta \int_{\bar{\theta}}^1 [\beta \bar{v}(0) + (c+z)y + \bar{\gamma}] \frac{1}{1-\bar{\theta}} dy + (c+\bar{z})\bar{\theta} \\ &= \beta^2 \bar{v}(0) + \beta \bar{\gamma} + \frac{\beta(c+\bar{z})}{2} + \left(\frac{\beta}{2} + 1\right)(c+\bar{z})\bar{\theta}, \end{aligned}$$

which combined with (33) gives

$$\frac{\beta}{2}(c+\bar{z})(1+\bar{\theta}) = (\beta \bar{v}(0) + \bar{\gamma})(1-\beta). \tag{35}$$

Given the special form of $Q_0(dy|x)$, (26) becomes

$$w(x) = \begin{cases} \frac{\beta}{1-x} \int_x^1 w(y) dy + R_1(x)R_2'(\bar{z})z_\gamma, & 0 \leq x \leq \bar{\theta}, \\ \beta w(0) + R_1(x)R_2'(\bar{z})z_\gamma + 1, & \bar{\theta} < x \leq 1. \end{cases} \tag{36}$$

The computation of w now reduces to uniquely solving $w(0)$. By the expression of w on $[0, \bar{\theta}]$, we have

$$\begin{aligned} w(\bar{\theta}) &= \beta \int_{\bar{\theta}}^1 w(y) Q_0(dy|\bar{\theta}) + R_1(\bar{\theta})R_2'(\bar{z})z_\gamma \\ &= \beta^2 w(0) + \beta + R_1(\bar{\theta})R_2'(\bar{z})z_\gamma + \frac{\beta R_2'(\bar{z})z_\gamma}{1-\bar{\theta}} \int_{\bar{\theta}}^1 R_1(y) dy \\ &= \beta^2 w(0) + \beta + \bar{\theta} z_\gamma + \beta z_\gamma \frac{1+\bar{\theta}}{2}. \end{aligned} \tag{37}$$

For $x \in [0, \bar{\theta}]$, we further write

$$w(x) = \frac{\beta}{1-x} \int_x^1 w(y) dy + R_1(x)R_2'(\bar{z})z_\gamma,$$

and solve

$$w(x) = (1-x)^{\beta-1} w(0) + z_\gamma x - \beta z_\gamma \left[\frac{(1-x)^{\beta-1}}{(1-\beta)(2-\beta)} - \frac{1}{1-\beta} + \frac{1-x}{2-\beta} \right],$$

which further gives

$$w(\bar{\theta}) = (1-\bar{\theta})^{\beta-1} w(0) + z_\gamma \bar{\theta} - \beta z_\gamma \left[\frac{(1-\bar{\theta})^{\beta-1}}{(1-\beta)(2-\beta)} - \frac{1}{1-\beta} + \frac{1-\bar{\theta}}{2-\beta} \right]. \tag{38}$$

By (37)–(38), we have

$$[\beta^{-1}(1-\bar{\theta})^{\beta-1} - \beta]w(0) = 1 + z_\gamma \left(\frac{1+\bar{\theta}}{2} + \frac{(1-\bar{\theta})^{\beta-1}}{(1-\beta)(2-\beta)} + \frac{1-\bar{\theta}}{2-\beta} - \frac{1}{1-\beta} \right). \tag{39}$$

Now from (30) we have

$$p(x) = \begin{cases} \int_0^x \frac{1}{1-y} p(y) dy + \pi_0, & 0 \leq x < \bar{\theta}, \\ \int_0^{\bar{\theta}} \frac{1}{1-y} p(y) dy + \pi_0, & \bar{\theta} \leq x \leq 1, \end{cases}$$

which determines

$$p(x) = \begin{cases} \frac{\pi_0}{1-x}, & 0 \leq x < \bar{\theta}, \\ \frac{\pi_0}{1-\bar{\theta}}, & \bar{\theta} \leq x \leq 1, \end{cases}$$

where $\pi_0 = \frac{1}{2-\ln(1-\bar{\theta})}$. We determine the mean field

$$\bar{z} = \int_0^{\bar{\theta}} xp(x)dx + \int_{\bar{\theta}}^1 xp(x)dx = \pi_0 \left(\frac{1-\bar{\theta}}{2} - \ln(1-\bar{\theta}) \right). \tag{40}$$

We further obtain $\frac{dz}{d\gamma}$ at $\bar{\gamma}$ as

$$z_\gamma = \frac{\ln(1-\bar{\theta}) - 3 + \frac{4}{1-\bar{\theta}}}{2[2-\ln(1-\bar{\theta})]^2} \theta_\gamma. \tag{41}$$

We note that a perturbation analysis directly based on the general case (30) is more complicated.

Now (28) reduces to

$$\left[\frac{\beta}{1-\bar{\theta}} \int_{\bar{\theta}}^1 \frac{\bar{v}(y)}{1-\bar{\theta}} dy - \frac{\beta \bar{v}(\bar{\theta})}{1-\bar{\theta}} \right] \theta_\gamma = 1 + \beta w(0) - \beta \int_{\bar{\theta}}^1 \frac{w(y)}{1-\bar{\theta}} dy.$$

By the expression of \bar{v} in (32) and w in (36) at $\theta = \bar{\theta}$, we obtain

$$\frac{(1-\beta)\bar{v}(\bar{\theta}) - \bar{\theta}(c+\bar{z})}{1-\bar{\theta}} \theta_\gamma = 1 + \beta w(0) - w(\bar{\theta}) + \bar{\theta} z_\gamma.$$

Recalling (33) and (37), we have

$$\frac{(1-\beta)[\beta \bar{v}(0) + \bar{\gamma}] - \beta \bar{\theta}(\bar{z} + c)}{1-\bar{\theta}} \theta_\gamma - \beta(1-\beta)w(0) + \frac{1+\bar{\theta}}{2} \beta z_\gamma = 1 - \beta. \tag{42}$$

By combining (34), (35) and (40), we have

$$\bar{v}(0) = [(1-\bar{\theta})^{\beta-1} - \beta]^{-1} \left[\frac{\beta(c+\bar{z})[(1-\bar{\theta})^{\beta-1} - 1]}{(1-\beta)(2-\beta)} - \frac{\beta(c+\bar{z})\bar{\theta}}{2-\beta} + \bar{\gamma} \right], \tag{43}$$

$$\bar{\theta} = \frac{2(1-\beta)(\beta \bar{v}(0) + \bar{\gamma})}{\beta(c+\bar{z})} - 1, \tag{44}$$

$$\bar{z} = \frac{1}{2-\ln(1-\bar{\theta})} \left(\frac{1-\bar{\theta}}{2} - \ln(1-\bar{\theta}) \right). \tag{45}$$

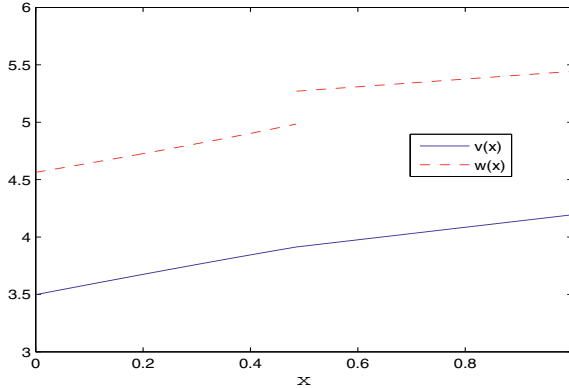


Fig. 1 Value function v and perturbation function w

Next, combining (39), (41) and (42), we obtain

$$\frac{(1 - \beta)[\beta\bar{v}(0) + \bar{\gamma}] - \beta\bar{\theta}(\bar{z} + c)}{1 - \bar{\theta}}\theta_\gamma - \beta(1 - \beta)w(0) + \frac{1 + \bar{\theta}}{2}\beta z_\gamma = 1 - \beta, \quad (46)$$

$$[\beta^{-1}(1 - \bar{\theta})^{\beta-1} - \beta]w(0) = 1 + z_\gamma\left(\frac{1 + \bar{\theta}}{2} + \frac{(1 - \bar{\theta})^{\beta-1}}{(1 - \beta)(2 - \beta)} + \frac{1 - \bar{\theta}}{2 - \beta} - \frac{1}{1 - \beta}\right), \quad (47)$$

$$z_\gamma = \frac{\ln(1 - \bar{\theta}) - 3 + \frac{4}{1 - \bar{\theta}}}{2[2 - \ln(1 - \bar{\theta})]^2}\theta_\gamma. \quad (48)$$

After $(\bar{v}(0), \bar{z}, \bar{\theta})$ has been determined from (43)-(45), the above gives a linear equation system with unknowns $w(0)$, θ_γ and z_γ .

Example 2. We take $R_1(x) = x$ and $R_2(z) = 0.5 + z$, $\bar{\gamma} = 0.5$. We numerically solve (43)-(45) to obtain $\bar{v}(0) = 3.497854$, $\bar{\theta} = 0.485162$, $\bar{z} = 0.345854$, and (46)-(48) to obtain $w(0) = 4.563055$, $\theta_\gamma = 1.162861$, $z_\gamma = 0.336380$. The curves of $v(x)$ and $w(x)$ are displayed in Fig. 1, where w has a discontinuity at $x = \bar{\theta}$ as discussed in Remark 2. The positive value of θ_γ implies the threshold increases with γ , as asserted in Theorem 3.

6 Conclusion

This paper considers mean field games in a framework of binary Markov decision processes (MDP) and establishes existence and uniqueness of stationary equilib-

ria. The resulting policy has a threshold structure. We further analyze comparative statics to address the impact of parameter variations in the model.

For future research, there are some potentially interesting extensions. One may consider a heterogenous population and study the emergence of free-riders who care more about their own effort costs and have less incentive to contribute to the common benefit of the population. Another modelling of a quite different nature involves negative externalities where other players' improvement brings more pressure on the player in question. For instance, this arises in competitions for market share. The modelling and analysis of the agent behavior will be of interest.

Appendix A: Preliminaries on Ergodicity

Assume (A3). The next two lemmas determine the limiting distribution of the state process under threshold policies.

Lemma A.1. *i) If $\theta = 0$, then the distribution of x_t^i remains to be the dirac measure δ_0 for all $t \geq 1$, for any x_0^i .*

ii) If $\theta = 1$ or $\theta = 1^+$, the distribution of x_t^i converges to the dirac measure δ_1 weakly.

Proof. Part i) is obvious and part ii) follows from (A3). \square

Let $x_t^{i,\theta}$ denote the state process generated by the θ -threshold policy with $\theta \in (0, 1)$, and let $P_\theta^t(x, \cdot)$ be the distribution of $x_t^{i,\theta}$ given $x_0^{i,\theta} = x$.

Lemma A.2. *For $\theta \in (0, 1)$, $\{x_t^{i,\theta}, t \geq 0\}$ is uniformly ergodic with stationary probability distribution π_θ , i.e.,*

$$\sup_{x \in S} \|P_\theta^t(x, \cdot) - \pi_\theta\|_{TV} \leq Kr^t, \tag{A.1}$$

for some constants $K > 0$ and $r \in (0, 1)$, where $\|\cdot\|_{TV}$ is the total variation norm of signed measures.

Proof. The proof is similar to that of the ergodicity theorem in [27], which assumed (A3'). We use (A3)-iii) to estimate r . \square

We take $C_s = \{0\}$ as a small set and $\theta \in (0, 1)$. The θ -threshold policy gives

$$P(x_2^{i,\theta} = 0 | x_0^{i,\theta} = 0) \geq \int_0^1 q(y|0)dy =: \varepsilon_0. \tag{A.2}$$

So for any Borel set B , $P(x_2^{i,\theta} \in B | x_0^{i,\theta} = 0) \geq \varepsilon_0 \delta_0(B)$, where δ_0 is the dirac measure. For θ' in a small neighborhood of θ , we can ensure that the θ' -threshold policy gives

$$P(x_2^{i,\theta'} \in B | x_0^{i,\theta'} = 0) \geq \frac{\varepsilon_0}{2} \delta_0(B). \tag{A.3}$$

Lemma A.3. *Suppose $\theta, \theta' \in (0, 1)$ for two threshold policies. Let the corresponding stationary distributions of the state process by π and π' . Then*

$$\lim_{\theta' \rightarrow \theta} \|\pi' - \pi\|_{\text{TV}} = 0.$$

Proof. Fix $\theta \in (0, 1)$. By (A.3) and [41], there exist a neighborhood $I_0 = (\theta - \kappa_0, \theta + \kappa_0) \subset (0, 1)$ and two constants $C, r \in (0, 1)$ such that for all $\theta' \in I_0$,

$$\|P_\theta^t(x, \cdot) - \pi\|_{\text{TV}} \leq Cr^t, \quad \|P_{\theta'}^t(x, \cdot) - \pi'\|_{\text{TV}} \leq Cr^t, \quad \forall x \in [0, 1].$$

Subsequently,

$$\|\pi' - \pi\|_{\text{TV}} \leq \|P_{\theta'}^t(0, \cdot) - P_\theta^t(0, \cdot)\|_{\text{TV}} + 2Cr^t.$$

For any given $\varepsilon > 0$, fix a large k_0 such that $2Cr^{k_0} \leq \varepsilon/2$. We show for all θ' sufficiently close to θ ,

$$\|P_{\theta'}^{k_0}(0, \cdot) - P_\theta^{k_0}(0, \cdot)\|_{\text{TV}} \leq \varepsilon/2.$$

Given two probability measures μ_t, μ'_t , define the probability measures μ_{t+1} and μ'_{t+1} ,

$$\mu_{t+1}(B) = \int_{\mathbf{S}} P_\theta(y, B) \mu_t(dy), \quad \mu'_{t+1}(B) = \int_{\mathbf{S}} P_{\theta'}(y, B) \mu'_t(dy),$$

for Borel set $B \subset [0, 1]$. Then

$$\begin{aligned} |\mu_{t+1}(B) - \mu'_{t+1}(B)| &\leq \left| \int_{\mathbf{S}} P_\theta(y, B) \mu_t(dy) - \int_{\mathbf{S}} P_{\theta'}(y, B) \mu_t(dy) \right| \\ &\quad + \left| \int_{\mathbf{S}} P_{\theta'}(y, B) \mu_t(dy) - \int_{\mathbf{S}} P_{\theta'}(y, B) \mu'_t(dy) \right| \\ &=: D_1 + D_2. \end{aligned}$$

We have

$$D_2 = \left| \int_{\mathbf{S}} P_{\theta'}(y, B) \mu_t(dy) - \int_{\mathbf{S}} P_{\theta'}(y, B) \mu'_t(dy) \right| \leq 2\|\mu_t - \mu'_t\|_{\text{TV}}.$$

Denote $\underline{\theta} = \min\{\theta, \theta'\}$ and $\bar{\theta} = \max\{\theta, \theta'\}$. Then

$$D_1 = \left| - \int_{[\underline{\theta}, \bar{\theta}]} Q_0(B|y) \mu_t(dy) + 1_B(0) \mu_t([\underline{\theta}, \bar{\theta}]) \right| \leq \mu_t([\underline{\theta}, \bar{\theta}]).$$

Setting $\mu_0 = \mu'_0 = \delta_0$, then $\mu_t = P_\theta^t(0, \cdot)$, $\mu'_t = P_{\theta'}^t(0, \cdot)$. Hence,

$$|P_{\theta'}^{t+1}(0, B) - P_\theta^{t+1}(0, B)| \leq 2\|P_{\theta'}^t(0, \cdot) - P_\theta^t(0, \cdot)\|_{\text{TV}} + P_\theta^t(0, [\underline{\theta}, \bar{\theta}]), \quad (\text{A.4})$$

which implies

$$\|P_{\theta'}^{t+1}(0, \cdot) - P_\theta^{t+1}(0, \cdot)\|_{\text{TV}} \leq 4\|P_{\theta'}^t(0, \cdot) - P_\theta^t(0, \cdot)\|_{\text{TV}} + 2P_\theta^t(0, [\theta, \theta']). \quad (\text{A.5})$$

For $\mu_0 = \mu'_0 = \delta_0$, we have $P_\theta^1(0, \cdot) = P_{\theta'}^1(0, \cdot)$. It is clear from (A.5) and Lemma 4 that for each $t \geq 1$,

$$\lim_{\theta' \rightarrow \theta} \|P_{\theta'}^t(0, \cdot) - P_\theta^t(0, \cdot)\|_{TV} = 0, \quad \lim_{\theta' \rightarrow \theta} P_\theta^t(0, [\underline{\theta}, \bar{\theta}]) = 0.$$

Therefore, for the fixed k_0 , there exists $\delta > 0$ such that for all θ' satisfying $|\theta' - \theta| < \delta$, $\|P_{\theta'}^{k_0}(0, \cdot) - P_\theta^{k_0}(0, \cdot)\|_{TV} < \frac{\varepsilon}{2}$ and $\|\pi' - \pi\|_{TV} \leq \varepsilon$. The lemma follows. \square

Appendix B: Cycle Average of A Regenerative Process

Let $0 < r < r' < 1$. Consider a Markov process $\{Y_t, t \geq 0\}$ with state space $[0, 1]$ and transition kernel $Q_Y(\cdot|y)$ which satisfies $Q_Y([y, 1]|y) = 1$ for any $y \in [0, 1]$ and is stochastically increasing. Suppose $Y_0 \equiv y_0 < r$. Define the stopping times

$$\tau = \inf\{t|Y_t \geq r\}, \quad \tau' = \inf\{t|Y_t \geq r'\}.$$

Lemma B.1. *If $E\tau < \infty$, then $E \sum_{t=0}^{\tau} Y_t < \infty$ and*

$$\frac{E \sum_{t=0}^{\tau} Y_t}{1 + E\tau} = \frac{EY_0 + EY_1 + \sum_{k=1}^{\infty} E(Y_{k+1} 1_{\{Y_k < r\}})}{2 + \sum_{k=1}^{\infty} P(Y_k < r)}. \tag{B.1}$$

Proof. Since $0 \leq Y_t \leq 1$ w.p. 1, $E \sum_{t=0}^{\tau} Y_t \leq 1 + E\tau$. It is clear that $\{\tau \geq k\} = \{Y_{k-1} < r\}$ for $k \geq 1$. We have

$$E\tau = \sum_{k=1}^{\infty} P(\tau \geq k) = 1 + \sum_{k=1}^{\infty} P(Y_k < r), \tag{B.2}$$

and

$$\begin{aligned} E \sum_{t=0}^{\tau} Y_t &= E \sum_{k=1}^{\infty} \left(\sum_{t=0}^k Y_t \right) 1_{\{\tau \geq k\}} \\ &= EY_0 + EY_1 + \sum_{k=2}^{\infty} E(Y_k 1_{\{\tau \geq k\}}) \\ &= EY_0 + EY_1 + \sum_{k=1}^{\infty} E(Y_{k+1} 1_{\{Y_k < r\}}). \end{aligned}$$

The lemma follows. \square

Lemma B.2. *Assume $E\tau' < \infty$. We have*

$$\frac{E \sum_{t=0}^{\tau} Y_t}{1 + E\tau} \leq \frac{E \sum_{t=0}^{\tau'} Y_t}{1 + E\tau'}. \tag{B.3}$$

Proof. $E\tau < \infty$ since $\tau \leq \tau'$ w.p.1. For $k \geq 1$, denote

$$p_k = P(Y_k < r), \quad \eta_k = P(r \leq Y_k < r'),$$

$$m_k = E(Y_{k+1} 1_{\{Y_k < r\}}), \quad \Delta_k = E(Y_{k+1} 1_{\{r \leq Y_k < r'\}}).$$

By Lemma B.1,

$$\frac{E \sum_{t=0}^{\tau} Y_t}{1 + E \tau} = \frac{EY_0 + EY_1 + \sum_{k=1}^{\infty} m_k}{2 + \sum_{k=1}^{\infty} p_k},$$

$$\frac{E \sum_{t=0}^{\tau'} Y_t}{1 + E \tau'} = \frac{EY_0 + EY_1 + \sum_{k=1}^{\infty} (m_k + \Delta_k)}{2 + \sum_{k=1}^{\infty} (p_k + \eta_k)}.$$

So (B.3) is equivalent to

$$(EY_0 + EY_1 + \sum_{k=1}^{\infty} m_k) (\sum_{k=1}^{\infty} \eta_k) \leq (\sum_{k=1}^{\infty} \Delta_k) (2 + \sum_{k=1}^{\infty} p_k). \tag{B.4}$$

By the stochastic monotonicity of Q_Y , we have

$$E[Y_{k+1} 1_{\{Y_k < r\}} | Y_k] = 1_{\{Y_k < r\}} \int_0^1 y Q_Y(dy | Y_k)$$

$$\leq 1_{\{Y_k < r\}} \int_0^1 y Q_Y(dy | r) =: c_r 1_{\{Y_k < r\}}.$$

Note that

$$c_r = \int_{y \geq r} y Q_Y(dy | r) \geq r. \tag{B.5}$$

Moreover,

$$E[Y_{k+1} 1_{\{r \leq Y_k < r'\}} | Y_k] = 1_{\{r \leq Y_k < r'\}} \int_0^1 y Q_Y(dy | Y_k)$$

$$\geq c_r 1_{\{r \leq Y_k < r'\}}.$$

It follows that

$$m_k = E[Y_{k+1} 1_{\{Y_k < r\}}] \leq c_r p_k, \quad \Delta_k = E[Y_{k+1} 1_{\{r \leq Y_k < r'\}}] \geq c_r \eta_k. \tag{B.6}$$

Since $Y_0 = y_0 < r$,

$$E[Y_1 | Y_0] = \int_0^1 y Q_Y(dy | Y_0) \leq c_r.$$

Hence, $E(Y_0 + Y_1) \leq r + c_r$. By (B.6) and (B.5),

$$\begin{aligned}
 & (EY_0 + EY_1 + \sum_{k=1}^{\infty} m_k)(\sum_{k=1}^{\infty} \eta_k) - (\sum_{k=1}^{\infty} \Delta_k)(2 + \sum_{k=1}^{\infty} p_k) \\
 & \leq (r + c_r + c_r \sum_{k=1}^{\infty} p_k)(\sum_{k=1}^{\infty} \eta_k) - c_r(\sum_{k=1}^{\infty} \eta_k)(2 + \sum_{k=1}^{\infty} p_k) \\
 & = (r - c_r) \sum_{k=1}^{\infty} \eta_k \leq 0,
 \end{aligned}$$

which establishes (B.4). \square

Remark B.1. If for each $y \in [0, 1)$, $Q_Y(dx|y)$ has probability density function $q_Y(x|y) > 0$ for $x \in (y, 1)$, then $c_r > r$ and $\eta_k > 0$ for all $k \geq 1$. In this case, a strict inequality holds for (B.3). \square

Appendix C

We assume (A3). Let $\{x_t^{i,\theta}, t \geq 0\}$ be the Markov chain generated by a θ -threshold policy with $0 < \theta < 1$, where $x_0^{i,\theta}$ is given. By Lemma A.2, $\{x_t^{i,\theta}, t \geq 0\}$ is ergodic. We next define an auxiliary Markov chain $\{Y_t, t \geq 0\}$ with $Y_0 = 0$ and the same transition kernel as $x_t^{i,\theta}$. Denote $S_t = \sum_{i=0}^t Y_i$ for $t \geq 0$. Define $\tau = \inf\{t | Y_t \geq \theta\}$.

Lemma C.1. *We have*

$$\lim_{k \rightarrow \infty} \frac{1}{k} \sum_{t=0}^{k-1} Y_t = \frac{ES_\tau}{1 + E\tau} \quad \text{w.p.1.} \tag{C.1}$$

Proof. By (A3), we can show $E\tau < \infty$. Since $\{Y_t, t \geq 0\}$ has the same transition probability kernel as $\{x_t^{i,\theta}, t \geq 0\}$, it is ergodic, and therefore the left hand side of (C.1) has a constant limit w.p.1. Define $T_0 = 0$ and T_n as the time for $\{Y_t, t \geq 0\}$ to return to state 0 for the n th time. So $T_1 = \tau + 1$. Define $B_n = \sum_{t=T_{n-1}}^{T_n-1} Y_t$ for $n \geq 1$. We observe that $\{Y_t, t \geq 0\}$ is a regenerative process (see e.g. [6, 51] and [7, Theorem 4]) with regeneration times $\{T_n, n \geq 1\}$ and that $\{B_n, n \geq 1\}$ is a sequence of i.i.d. random variables. Note that $B_1 = S_\tau$ is the sum of $\tau + 1$ terms. By the strong law of large numbers for regenerative processes [6, pp. 177], the lemma follows. \square

Suppose $0 < \theta < \theta' < 1$. Then there exist two constants $C_\theta, C_{\theta'}$ such that

$$\lim_{k \rightarrow \infty} \frac{1}{k} \sum_{t=0}^{k-1} x_t^{i,\theta} = C_\theta, \quad \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{t=0}^{k-1} x_t^{i,\theta'} = C_{\theta'}, \quad \text{w.p.1.}$$

Lemma C.2. *We have $C_\theta \leq C_{\theta'}$.*

Proof. Due to the ergodicity of the Markov chain, C_θ (resp., $C_{\theta'}$) does not depend on $x_0^{i,\theta}$ (resp., $x_0^{i,\theta'}$). Therefore, $\lim_{k \rightarrow \infty} \frac{1}{k} \sum_{t=0}^{k-1} Y_t = C_\theta$ w.p.1. The lemma follows from Lemmas C.1 and B.2. \square

Appendix D: An Auxiliary MDP

Assume (A3). This appendix introduces an auxiliary control problem to show the effect of the effort cost on the threshold parameter of the optimal policy. The state and control processes $\{(x_t^i, a_t^i), t \geq 0\}$ are specified by (1)-(2). The cost has the form

$$J_i^r = E \sum_{t=0}^{\infty} \rho^t (R_1(x_t^i) + r 1_{\{a_t^i = a_1\}}), \quad (\text{D.1})$$

where R_1 is continuous and strictly increasing on $[0, 1]$ and $\rho \in (0, 1)$, $r \in (0, \infty)$. Let r take two different values $0 < \gamma_1 < \gamma_2$ and write the corresponding dynamic programming equation

$$v_l(x) = \min \left\{ \rho \int_0^1 v_l(y) Q_0(dy|x) + R_1(x), \quad \rho v_l(0) + R_1(x) + \gamma_l \right\}, \quad l = 1, 2, x \in \mathbf{S}. \quad (\text{D.2})$$

By the method in proving Lemma 1, it can be shown that there exists a unique solution $v_l \in C([0, 1], \mathbb{R})$ and that the optimal policy $a^{i,l}(x)$ is a threshold policy. If $\rho \int_0^1 v_l(y) Q_0(dy|1) < \rho v_l(0) + \gamma_l$, $a^{i,l}(x) \equiv a_0$, and we follow the notation in Section 3 to denote the threshold $\theta_l = 1^+$. Otherwise, $a^{i,l}(x)$ is a θ_l -threshold policy with $\theta_l \in [0, 1]$, i.e., $a^{i,l}(x) = a_1$ if $x \geq \theta_l$, and $a^{i,l}(x) = a_0$ if $x < \theta_l$.

Lemma D.1. *If $\theta_1 \in (0, 1)$, $\theta_2 \neq \theta_1$.*

Proof. We prove by contradiction. Suppose for some $\theta \in (0, 1)$,

$$\theta_1 = \theta_2 = \theta. \quad (\text{D.3})$$

Under (D.3), the resulting optimal policy leads to the representation (see e.g. [23, pp. 22])

$$v_l(x) = E \sum_{t=0}^{\infty} \rho^t \left[R_1(x_t^i) + \gamma_l 1_{\{a_t^i = a_1\}} \right], \quad l = 1, 2,$$

where $\{x_t^i, t \geq 0\}$ is generated by the θ -threshold policy $a_t^i(x_t^i)$ and $x_0^i = x$. Denote $\delta_{21} = \gamma_2 - \gamma_1$.

For fixed $x \geq \theta$ and $x_0^i = x$, denote the resulting optimal state and control processes by $\{(\check{x}_t^i, \check{a}_t^i), t \geq 0\}$. Then $\check{a}_0^i = a_1$ w.p.1., and

$$v_2(x) - v_1(x) = \delta_{21} + \delta_{21} E \sum_{t=1}^{\infty} \rho^t 1_{\{\check{a}_t^i = a_1\}}, \quad x \geq \theta.$$

Next consider $x_0^i = 0$ and denote the optimal state and control processes by $\{(\check{x}_t^i, \check{a}_t^i), t \geq 0\}$. Then

$$v_2(0) - v_1(0) = \delta_{21} E \sum_{t=0}^{\infty} \rho^t 1_{\{\check{a}_t^i = a_1\}} =: \Delta.$$

It is clear that $\hat{x}_1^i = 0$ w.p.1. By the optimality principle, $\{(\hat{x}_t^i, \hat{a}_t^i), t \geq 1\}$ may be interpreted as the optimal state and control processes of the MDP with initial state 0 at $t = 1$. Hence the two processes $\{(\hat{x}_t^i, \hat{a}_t^i), t \geq 1\}$ and $\{(\check{x}_t^i, \check{a}_t^i), t \geq 0\}$, where $\check{x}_0^i = 0$, have the same finite dimensional distributions. In particular, \hat{a}_{t+1}^i and \check{a}_t^i have the same distribution for $t \geq 0$. Therefore,

$$E \sum_{t=1}^{\infty} \rho^{t-1} 1_{\{\hat{a}_t^i = a_1\}} = E \sum_{t=0}^{\infty} \rho^t 1_{\{\check{a}_t^i = a_1\}}.$$

It follows that

$$v_2(x) - v_1(x) = \delta_{21} + \rho \Delta, \quad \forall x \geq \theta. \quad (\text{D.4})$$

Combining (D.2) and (D.3) gives

$$\rho \int_0^1 v_l(y) Q_0(dy|\theta) = \rho v_l(0) + \gamma_l, \quad l = 1, 2,$$

which implies

$$\rho \int_0^1 [v_2(x) - v_1(x)] Q_0(dx|\theta) = \delta_{21} + \rho \Delta. \quad (\text{D.5})$$

By $Q_0([0, \theta]|\theta) = 0$ and (D.4), (D.5) further yields $\rho(\delta_{21} + \rho \Delta) = \delta_{21} + \rho \Delta$, which is impossible since $0 < \rho < 1$ and $\delta_{21} + \rho \Delta > 0$. Therefore, (D.3) does not hold. This completes the proof. \square

For the MDP with cost (D.1), we continue to analyze the dynamic programming equation

$$v_r(x) = \min \left[\rho \int_0^1 v_r(y) Q_0(dy|x) + R_1(x), \quad \rho v_r(0) + R_1(x) + r \right]. \quad (\text{D.6})$$

For each fixed $r \in (0, \infty)$, we obtain the optimal policy as a threshold policy with threshold parameter $\theta(r)$. By evaluating the cost (D.1) associated with the two policies $a_t^i(x_t^i) \equiv a_0$ and $a_t^i(x_t^i) \equiv a_1$, respectively, we have the prior estimate

$$v_r(x) \leq \min \left\{ \frac{R_1(1)}{1-\rho}, R_1(x) + \frac{r + \rho R_1(0)}{1-\rho} \right\}. \quad (\text{D.7})$$

On the other hand, let $\{x_t^i, t \geq 0\}$ with $x_0^i = x$ be generated by any fixed Markov policy. Then

$$E \sum_{t=0}^{\infty} \rho^t (R_1(x_t^i) + r 1_{\{a_t^i = a_1\}}) \geq R_1(x) + \sum_{t=1}^{\infty} \rho^t R_1(0),$$

which implies

$$v_r(x) \geq R_1(x) + \frac{\rho R_1(0)}{1-\rho}. \tag{D.8}$$

If $r > \frac{\rho R_1(1)}{1-\rho}$, it follows from (D.7) that

$$\rho \int_0^1 v_r(y) Q_0(dy|x) < \rho v_r(0) + r, \quad \forall x, \tag{D.9}$$

i.e., $\theta(r) = 1^+$.

Lemma D.2. *There exists $\delta > 0$ such that for all $0 < r < \delta$,*

$$\rho \int_0^1 v_r(y) Q_0(dy|x) > \rho v_r(0) + r, \quad \forall x, \tag{D.10}$$

and so $\theta(r) = 0$.

Proof. By (D.8),

$$\begin{aligned} \rho \int_0^1 v_r(y) Q_0(dy|x) &\geq \rho \int_0^1 R_1(y) Q_0(dy|x) + \frac{\rho^2 R_1(0)}{1-\rho} \\ &\geq \rho \int_0^1 R_1(y) Q_0(dy|0) + \frac{\rho^2 R_1(0)}{1-\rho}, \end{aligned}$$

and (D.7) gives

$$\rho v_r(0) + r \leq \frac{\rho R_1(0)}{1-\rho} + \frac{r}{1-\rho}.$$

Since $R_1(x)$ is strictly increasing,

$$C_{R_1} := \int_0^1 R_1(y) Q_0(dy|0) - R_1(0) > 0.$$

And we have

$$\rho \int_0^1 v_r(y) Q_0(dy|x) - (\rho v_r(0) + r) \geq \rho C_{R_1} - \frac{r}{1-\rho}.$$

It suffices to take $\delta = \rho(1-\rho)C_{R_1}$. \square

Define the nonempty sets

$$\mathcal{R}_{a_0} = \{r > 0 | \text{(D.9) holds}\}, \quad \mathcal{R}_{a_1} = \{r > 0 | \text{(D.10) holds}\}.$$

Remark D.1. We have $(\frac{\rho R_1(1)}{1-\rho}, \infty) \subset \mathcal{R}_{a_0}$ and $(0, \delta) \subset \mathcal{R}_{a_1}$.

Lemma D.3. *Let (r, v_r) be the parameter and the associated solution in (D.6).*

i) If $r > 0$ satisfies

$$\rho \int_0^1 v_r(y) \mathcal{Q}_0(dy|x) \leq \rho v_r(0) + r, \quad \forall x, \quad (\text{D.11})$$

then any $r' > r$ is in \mathcal{R}_{a_0} .

ii) If $r > 0$ satisfies

$$\rho \int_0^1 v_r(y) \mathcal{Q}_0(dy|x) \geq \rho v_r(0) + r, \quad \forall x, \quad (\text{D.12})$$

then any $r' \in (0, r)$ is in \mathcal{R}_{a_1} .

Proof. i) For $r' > r$, $v_{r'}$ is uniquely solved from (D.6) with r' in place of r . We can use (D.11) to verify

$$v_{r'}(x) = \min \left[\rho \int_0^1 v_r(y) \mathcal{Q}_0(dy|x) + R_1(x), \quad \rho v_r(0) + R_1(x) + r' \right].$$

Hence $v_{r'} = v_r$ for all $x \in [0, 1]$. It follows that $\rho \int_0^1 v_{r'}(y) \mathcal{Q}_0(dy|x) < \rho v_{r'}(0) + r'$ for all x . Hence $r' \in \mathcal{R}_{a_0}$.

ii) By (D.6) and (D.12), $v_r(0) = \frac{R_1(0)+r}{1-\rho}$, and subsequently,

$$v_r(x) = \rho v_r(0) + R_1(x) + r = \frac{\rho R_1(0) + r}{1-\rho} + R_1(x).$$

By substituting $v_r(0)$ and $v_r(x)$ into (D.12), we obtain

$$\rho R_1(0) + r \leq \rho \int_0^1 R_1(y) \mathcal{Q}_0(dy|x), \quad \forall x. \quad (\text{D.13})$$

Now for $0 < r' < r$, we construct $v_{r'}(x)$, as a candidate solution to (D.6) with r replaced by r' , to satisfy

$$v_{r'}(0) = \rho v_{r'}(0) + R_1(0) + r', \quad v_{r'}(x) = \rho v_{r'}(0) + R_1(x) + r', \quad (\text{D.14})$$

which gives

$$v_{r'}(x) = \frac{\rho R_1(0) + r'}{1-\rho} + R_1(x). \quad (\text{D.15})$$

We show that $v_{r'}(x)$ in (D.15) satisfies

$$\rho v_{r'}(0) + r' < \rho \int_0^1 v_{r'}(y) \mathcal{Q}_0(dy|x), \quad \forall x, \quad (\text{D.16})$$

which is equivalent to $\rho R_1(0) + r' < \rho \int_0^1 R_1(y) \mathcal{Q}_0(dy|x)$ for all x , which in turn follows from (D.13). By (D.14) and (D.16), $v_{r'}$ indeed satisfies (D.6) with r replaced by r' . So $r' \in \mathcal{R}_{a_1}$. \square

Further define

$$\underline{r} = \sup \mathcal{R}_{a_1}, \quad \bar{r} = \inf \mathcal{R}_{a_0}.$$

Lemma D.4. *i) \underline{r} satisfies $\rho \int_0^1 v_{\underline{r}}(y)Q_0(dy|0) = \rho v_{\underline{r}}(0) + \underline{r}$, and $\theta(\underline{r}) = 0$.*

ii) \bar{r} satisfies $\rho \int_0^1 v_{\bar{r}}(y)Q_0(dy|1) = \rho v_{\bar{r}}(1) = \rho v_{\bar{r}}(0) + \bar{r}$, and $\theta(\bar{r}) = 1$.

iii) We have $0 < \underline{r} < \bar{r} < \infty$.

iv) The threshold $\theta(r)$ as a function of $r \in (0, \infty)$ is continuous and strictly increasing on $[\underline{r}, \bar{r}]$.

Proof. i)-ii) By Lemmas D.2 and D.3, we have $0 < \underline{r} \leq \infty$ and $0 \leq \bar{r} < \infty$. Assume $\underline{r} = \infty$; then $\mathcal{R}_{a_1} = (0, \infty)$ giving $\mathcal{R}_{a_0} = \emptyset$, a contradiction. So $0 < \underline{r} < \infty$. For $\delta > 0$ in Lemma D.2, we have $(0, \delta) \subset \mathcal{R}_{a_1}$. Therefore, $0 < \bar{r} < \infty$. Note that v_r depends on the parameter r continuously, i.e., $\lim_{|r'-r| \rightarrow 0} \sup_x |v_{r'}(x) - v_r(x)| = 0$. Hence

$$\rho \int_0^1 v_{\underline{r}}(y)Q_0(dy|0) \geq \rho v_{\underline{r}}(0) + \underline{r}.$$

Now assume

$$\rho \int_0^1 v_{\underline{r}}(y)Q_0(dy|0) > \rho v_{\underline{r}}(0) + \underline{r}. \tag{D.17}$$

Then there exists a sufficiently small $\varepsilon > 0$ such that (D.17) still holds when $(\underline{r} + \varepsilon, v_{\underline{r}+\varepsilon})$ replaces $(\underline{r}, v_{\underline{r}})$; since $g(x) = \int_0^1 v_{\underline{r}+\varepsilon}(y)Q_0(dy|x)$ is increasing in x , then $\underline{r} + \varepsilon \in \mathcal{R}_{a_1}$, which is impossible. Hence (D.17) does not hold, and this proves i). ii) can be shown in a similar manner.

To show iii), assume

$$0 < \bar{r} < \underline{r} < \infty. \tag{D.18}$$

Then, recalling Remark D.1, there exist $r' \in \mathcal{R}_{a_0}$ and $r'' \in \mathcal{R}_{a_1}$ such that

$$0 < \bar{r} < r' < r'' < \underline{r} < \infty.$$

By Lemma D.3-i), $r'' \in \mathcal{R}_{a_0}$, and then $r'' \in \mathcal{R}_{a_0} \cap \mathcal{R}_{a_1} = \emptyset$, which is impossible. Therefore, (D.18) does not hold and we conclude $0 < \underline{r} \leq \bar{r} < \infty$. We further assume $\underline{r} = \bar{r}$. Then i)-ii) would imply $\int_0^1 v_{\underline{r}}(y)Q_0(dy|0) = v_{\underline{r}}(1)$, which is impossible since $v_{\underline{r}}$ is strictly increasing on $[0, 1]$ and (A3) holds. This proves iii).

iv) By the definition of \underline{r} and \bar{r} , it can be shown using (D.6) that $\theta(r) \in (0, 1)$ for $r \in (\underline{r}, \bar{r})$. By the continuous dependence of the function $v_r(\cdot)$ on r and the method of proving [27, Lemma 10], we can show the continuity of $\theta(r)$ on $(0, 1)$, and further show $\lim_{r \rightarrow \underline{r}^+} \theta(r) = 0$ and $\lim_{r \rightarrow \bar{r}^-} \theta(r) = 1$. So $\theta(r)$ is continuous on $[\underline{r}, \bar{r}]$. If $\theta(r)$ were not strictly increasing on $[\underline{r}, \bar{r}]$, there would exist $\underline{r} < r_1 < r_2 < \bar{r}$ such that

$$\theta(r_1) \geq \theta(r_2). \tag{D.19}$$

If $\theta(r_1) > \theta(r_2)$ in (D.19), by the continuity of $\theta(r)$, $\theta(\underline{r}) = 0$, $\theta(\bar{r}) = 1$, and the intermediate value theorem we may find $r' \in (\underline{r}, r_1)$ such that $\theta(r'_1) = \theta(r_2)$. Next, we replace r_1 by r'_1 . Thus if $\theta(r)$ is not strictly increasing, we may find $r_1 < r_2$ from

(\underline{r}, \bar{r}) such that $\theta(r_1) = \theta(r_2) \in (0, 1)$, which is a contradiction to Lemma D.1. This proves iv). \square

Remark D.2. By Lemmas D.3 and D.4, $\mathcal{R}_{a_1} = (0, \underline{r})$ and $\mathcal{R}_{a_0} = (\bar{r}, \infty)$.

Acknowledgement

We would like to thank Aditya Mahajan for helpful discussions.

References

1. Acemoglu, D., Jensen, M.K.: Aggregate comparative statics. *Games and Economic Behavior* **81**, 27-49 (2013)
2. Acemoglu, D., Jensen, M.K.: Robust comparative statics in large dynamic economies. *Journal of Political Economy* **123**, 587-640 (2015)
3. Adlakha, S., Johari, R., Weintraub, G.Y.: Equilibria of dynamic games with many players: Existence, approximation, and market structure. *J. Econ. Theory* **156**, 269-316 (2015)
4. Altman, E., Stidham, S.: Optimality of monotonic policies for two-action Markovian decision processes, with applications to control of queues with delayed information. *Queueing Systems* **21**, 267-291 (1995)
5. Amir R.: Sensitivity analysis of multisector optimal economic dynamics. *Journal of Mathematical Economics* **25**, 123-141 (1996)
6. Asmussen, S.: *Applied Probability and Queues*, 2nd edn. Springer, New York (2003)
7. Athreya, K.B., Roy, V.: When is a Markov chain regenerative? *Statistics and Probability Letters* **84**, 22-26 (2014)
8. Babichenko, Y.: Best-reply dynamics in large binary-choice anonymous games. *Games and Economic Behavior* **81**, 130-144 (2013)
9. Bardi, M.: Explicit solutions of some linear-quadratic mean field games. *Netw. Heterogeneous Media* **7**, 243-261 (2012)
10. Bauerle, N., Rieder, U.: *Markov Decision Processes with Applications to Finance*. Springer, Berlin (2011)
11. Becker R. A.: Comparative dynamics in aggregate models of optimal capital accumulation. *Quarterly Journal of Economics* **100**, 1235-1256 (1985)
12. Bensoussan, A., Frehse, J., Yam, P.: *Mean Field Games and Mean Field Type Control Theory*. Springer, New York (2013)
13. Biswas, A.: Mean field games with ergodic cost for discrete time Markov processes, preprint, arXiv:1510.08968, 2015.
14. Bonnans, J.F., Shapiro, A.: *Perturbation Analysis of Optimization Problems*. Springer-Verlag, New York (2000)
15. Brock, W.A., Durlauf, S. N.: Discrete choice with social interactions. *Rev. Econ. Studies* **68**, 235-260 (2001)
16. Caines, P.E.: Mean field games. In: Samad, T., Baillieul, J. (eds.) *Encyclopedia of Systems and Control*. Springer-Verlag, Berlin (2014)
17. Caines, P.E., Huang, M., Malhamé, R.P.: Mean Field Games. In: Basar, T., Zaccour, G. (eds.) *Handbook of Dynamic Game Theory*, pp. 345-372, Springer, Berlin (2017)
18. Cardaliaguet, P.: Notes on mean field games, University of Paris, Dauphine (2012)
19. Carmona R., Delarue, F.: *Probabilistic Theory of Mean Field Games with Applications*, vol I and II. Springer, Cham (2018)

20. Dorato, P.: On sensitivity in optimal control systems. *IEEE Transactions on Automatic Control* **8**, 256-257 (1963)
21. Filar, J.A., Vrieze, K.: *Competitive Markov Decision Processes*. Springer, New York (1997)
22. Gomes, D. A., Mohr, J., Souza, R.R.: Discrete time, finite state space mean field games. *J. Math. Pures Appl.* **93** 308-328, (2010)
23. Hernandez-Lerma, O.: *Adaptive Markov Control Processes*. Springer-Verlag, New York (1989)
24. Hicks, J. R.: *Value and Capital*. Clarendon Press, Oxford (1939)
25. Huang, M., Caines, P.E., Malhamé, R.P.: Individual and mass behaviour in large population stochastic wireless power control problems: Centralized and Nash equilibrium solutions. *Proc. 42nd IEEE Conference on Decision and Control*, pp. 98-105, Maui, HI (2003)
26. Huang, M., Caines, P.E., Malhamé, R.P.: Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ϵ -Nash equilibria. *IEEE Trans. Autom. Control* **52**, 1560-1571 (2007)
27. Huang, M., Ma, Y.: Mean field stochastic games: Monotone costs and threshold policies (in Chinese), *Sci. Sin. Math.* (special issue in honour of the 80th birthday of Prof. H-F. Chen) **46**, 1445-1460 (2016)
28. Huang, M., Ma, Y.: Mean field stochastic games with binary action spaces and monotone costs. *arXiv:1701.06661v1*, 2017.
29. Huang, M., Ma, Y.: Mean field stochastic games with binary actions: Stationary threshold policies. *Proc. 56th IEEE Conference on Decision and Control*, Melbourne, Australia, pp. 27-32 (2017)
30. Huang, M., Malhamé, R.P., Caines, P.E.: Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Commun. Inform. Systems* **6**, 221-251 (2006)
31. Huang, M., Zhou, M.: Linear quadratic mean field games: Asymptotic solvability and relation to the fixed point approach. *IEEE Transactions on Automatic Control* (2018, in revision, conditionally accepted)
32. Ito, K., Kunisch, K.: Sensitivity analysis of solutions to optimization problems in Hilbert spaces with applications to optimal control and estimation. *J. Differential Equations* **99**, 1-40 (1992)
33. Jovanovic, B., Rosenthal, R.W.: Anonymous sequential games. *Journal of Mathematical Economics* **17**, 77-87 (1988)
34. Jiang, L., Anantharam, V., Walrand, J.: How bad are selfish investments in network security? *IEEE/ACM Trans. Networking* **19**, 549-560 (2011)
35. Kolokoltsov, V.N.: Nonlinear Markov games on a finite state space (mean-field and binary interactions). *International J. Statistics Probability* **1**, 77-91 (2012)
36. Kress, R.: *Linear Integral Equations*. Springer, Berlin (1989)
37. Lasry, J.-M., Lions, P.-L.: Mean field games. *Japan. J. Math.* **2**, 229-260 (2007)
38. Lelarge, M., Bolot, J.: A local mean field analysis of security investments in networks. *Proc. ACM SIGCOMM NetEcon*, Seattle, WA, pp. 25-30, 2008
39. Li, T., Zhang, J.-F.: Asymptotically optimal decentralized control for large population stochastic multiagent systems. *IEEE Trans. Autom. Control* **53**, 1643-1660 (2008)
40. Manfredia, P., Posta, P.D., d'Onofrio, A., Salinelli, E., Centrone, F., Meo, C., Poletti, P.: Optimal vaccination choice, vaccination games, and rational exemption: An appraisal. *Vaccine* **28**, 98-109 (2010)
41. Meyn, S., Tweedie, R. L.: *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge University Press, Cambridge (2009)
42. Milgrom, P., Shannon, C.: Monotone comparative statics. *Econometrica* **62**, 157-80 (1994)
43. Moon, J., Basar, T.: Linear quadratic risk-sensitive and robust mean field games. *IEEE Trans. Autom. Control* **62**, 1062-1077 (2017)
44. Müller, A., Stoyan, D.: *Comparison Methods for Stochastic Models and Risks*. Wiley, Chichester (2002)
45. Oniki, H.: Comparative dynamics (sensitivity analysis) in optimal control theory. *J. Econ. Theory* **6**, 265-283 (1973)

46. Saldi, N., Basar, T., Raginsky, M.: Markov-Nash equilibria in mean-field games with discounted cost. *SIAM J. Control Optimization* **56**, 4256-4287 (2018)
47. Samuelson, P.A.: *Foundations of Economic Analysis*, enlarged edn., Harvard University Press, Cambridge, MA (1983)
48. Schelling, T.C.: Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. *The Journal of Conflict Resolution* **17**, 381-428 (1973)
49. Selten, R.: An axiomatic theory of a risk dominance measure for bipolar games with linear incentives. *Games and Econ. Behav.* **8**, 213-263 (1995)
50. Shapley, L.S.: Stochastic games. *Proc. Natl. Acad. Sci.* **39**, 1095-1100 (1953)
51. Sigman, K., Wolff, R.W.: A review of regenerative processes. *SIAM Rev.* **35**, 269-288 (1993)
52. Sun, Y.: The exact law of large numbers via Fubini extension and characterization of insurable risks. *J. Econ. Theory* **126**, 31-69 (2006)
53. Topkis, D.M.: *Supermodularity and Complementarity*. Princeton Univ. Press, Princeton (1998)
54. Walker, M., Wooders, J., Amir, R.: Equilibrium play in matches: Binary Markov games. *Games and Economic Behavior* **71**, 487-502 (2011)
55. Weintraub, G.Y., Benkard, C.L., Van Roy, B.: Markov perfect industry dynamics with many firms. *Econometrica* **76**, 1375-1411 (2008)
56. Yong, J.: Linear-quadratic optimal control problems for mean-field stochastic differential equations. *SIAM J. Control Optim.* **51**, 2809-2838 (2013)
57. Zhou, M., Huang, M.: Mean field games with Poisson point processes and impulse control. *Proc. 56th IEEE Conference on Decision and Control*, Melbourne, Australia pp. 3152-3157 (2017)



Equivalence of Fluid Models for $G_t/GI/N + GI$ Queues

Weining Kang and Guodong Pang

Abstract Three different fluid model formulations have been recently developed for $G_t/GI/N + GI$ queues, including a two-parameter fluid model in Whitt (2006) by tracking elapsed service and patience times of each customer, a measure-valued fluid model in Kang and Ramanan (2010) and its extension in Zuñiga (2014) by tracking elapsed service and patience times of each customer, and a measure-valued fluid model in Zhang (2013) by tracking residual service and patience times of each customer. We show that, under general initial conditions, the first two fluid model formulations tracking elapsed times (Whitt's and Kang and Ramanan's fluid models) are equivalent and can be used to describe the same $G_t/GI/N + GI$ queue when the service and patience time distributions have densities, whereas, Zuñiga's fluid model and Zhang's fluid model are equivalent only when the initial conditions for the $G_t/GI/N + GI$ queue satisfy certain assumptions. We identify these conditions under which Zuñiga's fluid model and Zhang's fluid model can be derived from each other for the same system. The equivalence properties discovered provide important implications for the understanding of the recent development for non-Markovian many-server queues.

1 Introduction

Many-server queueing models with abandonment have attracted substantial attention because of their appealing applications to customer contact centers and health-care; see, e.g., [2], [3], [4], [6], and references therein. In the $G_t/GI/N + GI$ model,

Weining Kang

Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, e-mail: wkang@umbc.edu

Guodong Pang

The Harold and Inge Marcus Dept. of Industrial and Manufacturing Eng., College of Engineering, Pennsylvania State University, University Park, PA 16802, e-mail: gup3@psu.edu

there are N parallel servers, and customers arrive with a time-varying arrival rate, require i.i.d. service times, and have i.i.d. patience times; the arrival process, service and patience times are assumed to be mutually independent. The service discipline is first-come-first-served (FCFS) and non-idling, that is, no server will idle whenever there is a customer in queue.

Because of the difficulty in the exact analysis of such stochastic systems, fluid models have been recently developed to approximate the system dynamics and performance measures in a many-server heavy-traffic regime, where the arrival rate and the number of servers get large and service and patience time distributions are fixed. The conventional approach of using total number of customers in the system to describe system dynamics is insufficient to give a complete description and study some performance measures. Thus, measure-valued and two-parameter processes that track elapsed or residual service and patience times of each customer have been recently used to study these stochastic models.

Whitt [21] pioneered the use of two-parameter processes to describe the system dynamics (Definition 1). In particular, $Q(t, y)$ represents the number of customers in queue at time t that have waited for less than or equal to y , and $B(t, y)$ represents the number of customers in service at time t that have received service for less than or equal to y . His idea is to represent these two-parameter processes as integrals of their densities $q(t, y)$ and $b(t, y)$ with respect to y (if they exist), respectively, which satisfy two fundamental evolution equations ((2.14) and (2.15) in [21]), respectively. A queue boundary process plays an important role in determining the real fluid queue size: the two-parameter density function $q(t, y)$ becomes zero for y beyond the queue boundary at each time t . This approach is generalized to study the $G_t/GI/N_t + GI$ model with both time-varying arrival rates and numbers of servers [12] and [13].

Kang and Ramanan [10], following Kaspi and Ramanan [11], used two measure-valued processes to describe the service and queueing dynamics, one tracking the amount of time each customer has been in service, and the other tracking the amount of time each customer has spent in a potential queue, where all customers enter the potential queue upon arrival, and stay there until their patience times run out. The potential queue includes customers waiting in the real queue as well as those that have entered service or even departed but whose patience times have not run out. They also use a frontier waiting-time process to track the waiting time of the customer in front of the queue at each time. This frontier waiting-time process is used to determine the real fluid queue dynamics from the measure-valued process for the potential queue. The description of system dynamics is then completed by the balance equations for the fluid content processes associated with the queue, the service station and the entire system, as well as the non-idling condition; see Definition 2.

We summarize these two approaches of tracking elapsed service and patience times by stating that the two-parameter process approach in Whitt [21] describes the system dynamics by the densities and rates, while the measure-valued process approach in Kang and Ramanan [10] describes the system dynamics by the distributions and counting processes directly. The existence and uniqueness of Whitt's two-parameter fluid model are shown in discrete time under the assumption that the service and patience times have densities in [21]. They also follow, as a special

case, from the existence and uniqueness results established in [12, 13] of the two-parameter fluid model for $G_t/GI/N_t + GI$ queueing model with both time-varying arrival rates and numbers of servers under the assumptions that the system only alternates between overloaded and underloaded regimes (with a finite number of alternations in each finite time interval) and that the service and patience time distributions have piecewise continuous densities. The existence and uniqueness of Kang-Ramanan's fluid model are established in [10] via the fluid limits and more recently in [7] via the characterization of fluid model solution directly under the assumptions that the service time distribution G^S has density and the hazard rate function h^r of patience times is a.e. locally bounded. Zuñiga [23] has recently extended Kang-Ramanan's fluid model for general service time distributions and continuous patience time distributions.

One would expect that the two approaches are equivalent since they are different formulations for the same $G_t/GI/N + GI$ queue. Our first main result is to establish this equivalence in Theorem 1: first, a set of two-parameter fluid equations derived from the measure-valued fluid model satisfies the fluid model equations in [21] (see Proposition 4.1), and second, a set of measure-valued fluid equations derived from the two-parameter fluid model satisfies the fluid model equations in [10] (see Proposition 4.2). The equivalence property we establish provides a proof for the conjecture on the existence and uniqueness of Whitt's two-parameter fluid model under the assumption that the service and patience time distributions have densities (Conjecture 2.2 in [21]). The two-parameter process formulation depends critically on the existence of the densities of the service and patience time distributions, since the densities of the two-parameter processes may not exist for general service and patience time distributions (see Remark 4).

As a different approach, the system dynamics of $G_t/GI/N + GI$ queues can also be described by tracking residual service and patience times. It was conjectured in Section 3.3.2 of Kaspi and Ramanan [11] (in the case of no abandonment) that a measure-valued fluid model that tracks customers' residual service times and patience times can also be formulated in parallel to the fluid model tracking elapsed times. One advantage of considering a fluid model tracking residual times is that it enables us to easily analyze some performance measures, such as the system workload at any given time, which rely directly on the customers' residual service times; see, e.g., [5, 19] for infinite-server models and [11] for $G_t/GI/N$ queues. Such a fluid model tracking residual times, if suitably formulated, should be also equivalent to the above three fluid models tracking elapsed times.

Zhang [22] provided a fluid model tracking residual times for the $G/GI/N + GI$ model with a constant arrival rate (Definition 4). Instead of using the potential queue as described in the fluid models tracking elapsed times, Zhang's model uses a virtual queue to describe the queueing dynamics, where all customers enter the virtual queue upon arrival and stay there until their time to enter service, which may include customers whose patience times have run out already. The existence and uniqueness of this fluid model are shown assuming continuous service time distribution and Lipschitz continuous patience time distributions [22]. We study the relationship of Zhang's fluid model with the above three fluid models, in particular,

focusing on Zuñiga's fluid model, and find that they are not entirely equivalent formulations for the $G/GI/N + GI$ queue under general initial conditions; see Remarks 6-8 in Section 3.3. The disparity lies in the initial conditions assumed for those fluid models, in particular, the assumptions imposed on the initial contents in the virtual queue and in service in Zhang's fluid model. For example, in Kang-Ramanan and Zuñiga's fluid models, it is required that the residual service time of initial content in $v_0(dx)$ should have distribution with density $g^s(x + \cdot)/\bar{G}^s(x)$, whereas, in Zhang's fluid model, there is no requirement on the distribution of the residual service time of initial content in service. We identify the set of necessary and sufficient conditions on the initial contents for the equivalence of Zhang's fluid model and the above three fluid models (Theorems 2 and 3 and Corollary 3.1). It is important to note that in comparison of these different fluid models, they should start with the same input data including the initial conditions.

On the other hand, from Kang-Ramanan and Zuñiga's fluid models, we obtain measure-valued fluid processes tracking residual service and patience times, which, together with the same input data as in those two fluid models, describe the service and real queueing dynamics of the same $G_t/GI/N + GI$ systems. These processes tracking residual times play an important bridging role in the discussion of the non-equivalence of Zhang's fluid model and the fluid models tracking elapsed times.

These equivalence properties established in the paper are significant to understand the fluid dynamics of the $G_t/GI/N + GI$ model from different perspectives. They help to unify the different approaches in the literature, and also highlight their differences and limitations. They provide the flexibility of choosing the most convenient approach among the different formulations, tracking elapsed or residual times, and the possibility of applying results from one formulation to another. Some properties established with one approach can then be directly applied to other models by the equivalence relationship. We illustrate this by two examples. First, an asymptotic periodic property is proved in [15] for the two-parameter fluid model tracking elapsed times for the $G_t/M_t/N_t + GI_t$ queueing model, and thus, should also hold for the associated measure-valued fluid models tracking elapsed and residual times (in the special case of $G_t/M/N + GI$ queues). Second, it is important to show that for a fluid model, the fluid solutions converge uniformly to the steady state over all possible initial states. That has been a difficult task for general non-Markovian many-server models. Thus, the equivalence property in this paper paves the way to show this with possibly any of the fluid models, whichever most convenient (see [16] for some recent attempts in this direction). In addition, the equivalence property results in an algorithm to compute two-parameter processes and relevant quantities under the most general conditions that cannot be computed by previous methods (see [9] and its extension in [17] to fluid models of $G_t/GI/N + GI$ queues under the least-patient first service discipline).

Although these equivalence properties are established for the fluid limits of the associated fluid-scaled stochastic processes in the queueing model, it is conceivable that the proofs for the convergence to these fluid limits may also be unified. The two-parameter approach proves the convergence in the functional space $\mathcal{D}_{\mathcal{G}} = \mathcal{D}([0, \infty), \mathcal{D}([0, \infty), \mathbb{R}))$ endowed with the Skorokhod J_1 topology. The measure-

valued approach proves the convergence in the measure-valued functional space $\mathcal{D}([0, \infty), \mathcal{M}([0, \infty)))$ where $\mathcal{M}([0, \infty))$ is the space of Radon measures on \mathbb{R}_+ endowed with the Borel σ -algebra. Tracking elapsed times enables us to use martingale arguments [10], but tracking residual times uses a different approach to prove the convergence [22]. So it is interesting to ask how these different approaches to establish the convergence are related and what would be the most general assumptions on the system primitives. We believe that these equivalence and coupling properties are useful in the study of other non-Markovian many-server queueing systems and networks.

Organization of the paper. The rest of the paper is organized as follows. We finish this section with some notation. In Section 2, we first review the definitions of the three fluid models tracking elapsed times, and then show their equivalence (Theorem 1), whose proof is given in Section 4. In Section 3, we first state and discuss the fluid measure-valued processes tracking residual times derived from Kang-Ramanan and Zuñiga’s fluid models in Section 3.1. We then review Zhang’s fluid model in Section 3.2 and discuss its connection with the three fluid models tracking elapsed times in Section 3.3.

Notation. We use \mathbb{R} and \mathbb{R}_+ to denote the spaces of real numbers and nonnegative real numbers, respectively. Given any metric space S , $\mathcal{C}_b(S)$ is the space of bounded, continuous real-valued functions on S . Let $\mathcal{C}_c(\mathbb{R}_+)$ be the space of continuous real-valued functions on \mathbb{R}_+ with compact support. Given a Radon measure ξ on $[0, H)$ and an interval $[a, b] \subset [0, H)$, we will use $\xi[a, b]$ to denote $\xi([a, b])$. Let $\mathcal{D}_{[0, \infty)}^{abs}(\mathcal{M}[0, H))$ denote the set of measure-valued processes μ with values in $\mathcal{M}[0, H)$, the space of Radon measures on $[0, H)$, such that for any $t \geq 0$, the measure $\int_0^t \mu_s(\cdot) ds$ is absolutely continuous with respect to the Lebesgue measure on $[0, H)$. Let $\mathcal{D}_{[0, \infty)}(\mathbb{R})$ be the space of real-valued càdlàg functions on $[0, \infty)$. For each real-valued function f defined on $[0, \infty)$, let f^+ and f^- be the positive and the negative parts of f , respectively, that is, $f^+(t) = f(t) \vee 0$ and $f^-(t) = -(f(t) \wedge 0)$ for each $t \geq 0$.

2 Fluid models tracking elapsed times

In the $G_t/GI/N + GI$ fluid models, we let $E(t)$ represent the cumulative amount of fluid content (representing customers) entering the system in the time interval $(0, t]$ for each $t > 0$. Assume that E is a non-decreasing function defined on $[0, \infty)$ with the density function $\lambda(\cdot) \geq 0$, that is,

$$E(t) = \int_0^t \lambda(s) ds, \quad t \geq 0. \tag{1}$$

Let G^s and G^r denote the service and patience time distribution functions, respectively. We assume that $G^s(0+) = G^r(0+) = 0$. Let

$$H^r \doteq \inf\{x \in \mathbb{R}_+ : G^r(x) = 1\}, \quad H^s \doteq \inf\{x \in \mathbb{R}_+ : G^s(x) = 1\}.$$

Then H^r and H^s are right supports of G^r and G^s , respectively.

2.1 Whitt’s two-parameter fluid model

In this section we state a modified version of the two-parameter fluid model in Whitt [21]. We assume that the functions G^s and G^r have density functions g^s and g^r on $[0, \infty)$, respectively. Let the hazard rate functions of G^s and G^r be defined as $h^r \doteq g^r/\bar{G}^r$ on $[0, H^r)$ and $h^s \doteq g^s/\bar{G}^s$ on $[0, H^s)$, respectively, where $\bar{G}^r = 1 - G^r$ and $\bar{G}^s = 1 - G^s$.

Let the two-parameter processes $B(t, y)$ be the amount of fluid content in service at time t that has been in service for less than or equal to y units of time, $\tilde{Q}(t, y)$ be the amount of fluid content in the potential queue at time t that has been in potential queue for less than or equal to y units of time, which may include the fluid content that has entered service or even departed by time t , and $Q(t, y)$ be the portion of $\tilde{Q}(t, y)$ that excludes the fluid content which has entered service by time t . Then it is obvious that $B(t, \infty)$ is the total fluid content in service and $Q(t, \infty)$ is the total fluid content in queue waiting for service.

It is assumed that these three processes are Lebesgue integrable on $[0, \infty)$ with densities $b(t, y)$, $\tilde{q}(t, y)$ and $q(t, y)$ with respect to the second component y , that is,

$$B(t, y) = \int_0^y b(t, x) dx \leq 1, \quad \tilde{Q}(t, y) = \int_0^y \tilde{q}(t, x) dx \geq 0, \tag{2}$$

$$Q(t, y) = \int_0^y q(t, x) dx \geq 0.$$

Let $\tilde{q}(0, x) = q(0, x)$ as a function in x have support in $[0, H^r)$ and $b(0, x)$ as a function in x have support in $[0, H^s)$. Note that in [21], it is not explicitly stated that the service and patience time distributions G^s and G^r can be of finite support.

Definition 1. A pair of functions $(B(t, y), Q(t, y))$ is a two-parameter fluid model tracking elapsed times with the input data $(\lambda(\cdot), \tilde{q}(0, x), b(0, x))$ if it satisfies the following conditions.

(i) The service density function $b(t, x)$ satisfies

$$b(t + u, x + u) = b(t, x) \frac{\bar{G}^s(x + u)}{\bar{G}^s(x)}, \quad x \in [0, H^s), t \geq 0, u > 0. \tag{3}$$

(ii) The potential queue density function $\tilde{q}(t, x)$ satisfies

$$\tilde{q}(t + u, x + u) = \tilde{q}(t, x) \frac{\bar{G}^r(x + u)}{\bar{G}^r(x)}, \quad x \in [0, H^r), t \geq 0, u > 0. \tag{4}$$

(iii) There exists a queue boundary function $w(t)$ such that $\tilde{Q}(t, w(t)) = Q(t, \infty)$ and then the queue density function $q(t, x)$ satisfies

$$q(t, x) = \begin{cases} \tilde{q}(t, x), & x \leq w(t), \\ 0, & x > w(t). \end{cases} \tag{5}$$

(iv) The density functions $b(t, x)$, $\tilde{q}(t, x)$ and $q(t, x)$ satisfy the following boundary properties:

$$b(t, 0) = \begin{cases} \lambda(t), & \text{if } B(t, \infty) < 1, \\ \sigma(t) \wedge \lambda(t), & \text{if } B(t, \infty) = 1, \text{ and } Q(t, \infty) = 0, \\ \sigma(t), & \text{if } B(t, \infty) = 1, \text{ and } Q(t, \infty) > 0, \end{cases} \tag{6}$$

$$\tilde{q}(t, 0) = \lambda(t), \tag{7}$$

and

$$q(t, 0) = \begin{cases} \lambda(t), & \text{if } Q(t, \infty) > 0 \text{ (} w(t) > 0 \text{)}, \\ \lambda(t) - (\sigma(t) \wedge \lambda(t)), & \text{if } B(t, \infty) = 1, \text{ and } Q(t, \infty) = 0, \\ 0, & \text{if } B(t, \infty) < 1, \end{cases} \tag{8}$$

where

$$\sigma(t) = \int_{[0, H^s)} b(t, x) h^s(x) dx, \quad t \geq 0. \tag{9}$$

(v) The densities $\lambda(t)$, $q(t, x)$, $b(t, x)$ and $\alpha(t)$ satisfy the balance equation:

$$\int_0^t \lambda(s) ds + \int_0^\infty q(0, x) dx = \int_0^\infty q(t, x) dx + \int_0^t b(s, 0) ds + \int_0^t \alpha(s) ds, \tag{10}$$

where

$$\alpha(t) = \int_{[0, H^r)} q(t, x) h^r(x) dx, \quad t \geq 0. \tag{11}$$

(vi) The densities $b(t, x)$, $q(t, x)$ satisfy the non-idling condition:

$$\left(\int_0^\infty (b(t, x) + q(t, x)) dx - 1 \right)^+ = \int_0^\infty q(t, x) dx, \tag{12}$$

$$\left(\int_0^\infty (b(t, x) + q(t, x)) dx \right) \wedge 1 = \int_0^\infty b(t, x) dx, \tag{13}$$

$$\left(\int_0^\infty b(t, x) dx - 1 \right) \int_0^\infty q(t, x) dx = 0. \tag{14}$$

In [21], equations (3) and (4) are called the first and second fundamental evolution equations, respectively. Note that the first fundamental evolution equation (3) essentially says that the fluid content in service that has not completed service remains in service. Similarly, the second fundamental evolution equation (4) essentially says that the fluid content in the potential queue that has not reached its

patience time remains in the potential queue. For each time t , the queue boundary quantity $w(t)$ divides the fluid content in the potential queue into two portions. The fluid content on the left side of $w(t)$ is still in queue waiting for service and the fluid content on the right side of $w(t)$ has entered service or even departed. The quantities $b(t, 0)$, $\tilde{q}(t, 0)$, $q(t, 0)$ in condition (iv) above are exactly the rates at time t at which the fluid content enters service, the potential queue and the queue, respectively. The quantities $\sigma(t)$ in (9) and $\alpha(t)$ in (11) are precisely the total service rate and the total abandonment rate at each time t , respectively. At last, the balance equation (10) is implicit in the definition of the fluid model in [21] and stated in equation (6) in [12]. We remark that the non-idling condition (vi) is implicit in [21], but is explicitly stated in a subsequent paper by Liu and Whitt [12]. We also remark that if $\lambda(t) > 0$, then the no-idling conditions in (12)–(14) are redundant, since they can be derived by the conditions in (6)–(8). Indeed, notice from (8) that $q(t, 0) = 0$ when $B(t) < 1$ and $q(t, 0) = \lambda(t) > 0$ when $Q(t) > 0$. Thus, $B(t) < 1$ and $Q(t) > 0$ cannot happen at the same time since $\lambda(t) > 0$. This will also imply that $B(t) = X(t) \wedge 1$ and $Q(t) = (X(t) - 1)^+$ for all $t \geq 0$.

Remark 1. (Existence and uniqueness of Whitt's fluid model.) Whitt [21] has shown the existence and uniqueness of the two-parameter fluid model for $G_t/GI/N + GI$ queues in discrete time by proving a functional weak law of large numbers (FWLLN), and conjectured them in continuous time (cf. Conjecture 2.2 of [21]). The existence and uniqueness of the two-parameter fluid model for $G_t/GI/N_t + GI$ queues with time-dependent staffing are shown in Liu and Whitt [12, 13], by an explicit characterization of the solution to the fluid model in [12] and by proving an FWLLN in [13], under the additional assumptions that the system only alternates between overloaded and underloaded regimes (with a finite number of alternations in each finite time interval) and that the service and patience time distributions have piecewise continuous densities. Thus, by specializing their argument to $G_t/GI/N + GI$ queues, the conjecture is proved but with the previously mentioned additional assumptions.

In this paper, we prove the conjecture under the assumption that the service and patience time distributions have densities, without assuming, a priori, that the system only alternates between overloaded and underloaded regimes, by applying the equivalence between the two fluid models in Definitions 1 and 2 established in Theorem 1 below and the existence and uniqueness of Kang-Ramanan's fluid model established in [7, 10]. We remark that the existence of the densities of the service and patience time distributions is critical for the formulation of Whitt's two-parameter fluid model, because the densities of $B(t, y)$ and $Q(t, y)$ with respect to y may not exist when the service and/or patience time distributions are general (see Remark 4).

2.2 Kang-Ramanan’s measure-valued fluid model

In this section, we state the measure-valued fluid model in Kang and Ramanan [10]. They use two measure-valued processes to describe the service and queuing dynamics. Let ν_t be a nonnegative finite measure on $[0, \infty)$ with support in $[0, H^s)$ such that $\nu_t(dx)$, $x \in [0, H^s)$, represents the amount of fluid content of customers in service whose time spent in service by time t lies in the range $[x, x + dx)$. Let η_t be another nonnegative finite measure on $[0, \infty)$ with support in $[0, H^r)$ such that $\eta_t(dx)$, $x \in [0, H^r)$, represents the amount of fluid content in the potential queue whose time spent there by time t lies in the range $[x, x + dx)$, where the potential queue is an artificial queue that includes the fluid content of customers in queue waiting for service and also the fluid content of customers that has entered service or even departed, but whose patience time has not been reached.

We assume that the functions G^s and G^r have density functions g^s and g^r on $[0, \infty)$, respectively. Let \mathcal{S}_0 denote the set of triples (η, ν, x) such that $1 - \nu[0, H^s) = [1 - x]^+$ and $\nu[0, H^s) + \eta[0, H^r) = x$, where η is a non-negative finite measure on $[0, \infty)$ with support in $[0, H^r)$, ν is a non-negative finite measure on $[0, \infty)$ with support in $[0, H^s)$, and $x \in \mathbb{R}_+$. The set \mathcal{S}_0 represents all possible measures of (η, ν) and values of x that the initial state of the measure-valued fluid model (η, ν, X) can take, satisfying the non-idling condition.

Definition 2. A triple of functions (η, ν, X) is a measure-valued fluid model tracking elapsed times with the input data $(\lambda(\cdot), \eta_0, \nu_0, X(0))$ such that $(\eta_0, \nu_0, X(0)) \in \mathcal{S}_0$ if it satisfies the following equations. For every $\psi \in \mathcal{C}_b(\mathbb{R}_+)$ and $t \geq 0$,

$$\int_0^\infty \psi(x)\eta_t(dx) = \int_{[0, H^r)} \psi(x+t) \frac{\bar{G}^r(x+t)}{\bar{G}^r(x)} \eta_0(dx) + \int_0^t \psi(t-s) \bar{G}^r(t-s) \lambda(s) ds, \tag{15}$$

$$\int_0^\infty \psi(x)\nu_t(dx) = \int_{[0, H^s)} \psi(x+t) \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) + \int_{[0, t]} \psi(t-s) \bar{G}^s(t-s) dK(s), \tag{16}$$

where

$$K(t) = B(t) + D(t) - B(0) = \nu_t[0, H^s) + D(t) - \nu_0[0, H^s), \tag{17}$$

$$D(t) = \int_0^t \left(\int_{[0, H^s)} h^s(x) \nu_s(dx) \right) ds, \tag{18}$$

$$E(t) + Q(0) = Q(t) + K(t) + R(t), \tag{19}$$

$$R(t) = \int_0^t \left(\int_{[0, \chi(s)]} h^r(x) \eta_s(dx) \right) ds, \tag{20}$$

$$\chi(s) = \inf\{x \in [0, H^r) : \eta_s[0, x] \geq Q(s)\}, \tag{21}$$

$$Q(t) = (X(t) - 1)^+, \tag{22}$$

$$B(t) = v_t[0, \infty) = X(t) \wedge 1 = 1 - (1 - X(t))^+, \tag{23}$$

and

$$Q(t)(1 - B(t)) = 0. \tag{24}$$

In this fluid model, $B(t)$ represents the total fluid content of customers in service, $Q(t)$ represents the total fluid content of customers in queue waiting for service, and $X(t)$ represents the total fluid content of customers in the system at each time t . Then, by (22) and (23),

$$X(t) = B(t) + Q(t). \tag{25}$$

The additional quantities $K(t)$, $R(t)$, $D(t)$, $\chi(t)$ can naturally be interpreted, respectively, as the cumulative amount of fluid content that has entered service by time t , the cumulative amount of fluid content that has abandoned from the queue by time t , the amount of fluid content that has departed the system after service completion by time t , and the waiting time of the fluid content at the head of the queue at time t , that is, the fluid content in queue with the longest waiting time.

For completeness, we now provide an intuitive explanation for these fluid equations. The equation (15) governs the evolution of the measure-valued process η_t . Note that when $x \leq t$, the amount of fluid content $\eta_t(dx)$ is the fraction of the amount of fluid content $\lambda(t - x)$ arriving to the system at time $t - x$ and whose time in the system since its arrival is more than x by time t . It is easy to see that this fraction equals to $\bar{G}^r(x)$. When $x > t$, the amount of fluid content $\eta_t(dx)$ is the fraction of the amount of fluid content $\eta_0(d(x - t))$ initially in queue and whose waiting time is more than x by time t given that it is more than $x - t$ at time 0. This fraction equals to $\bar{G}^r(x) / \bar{G}^r(x - t)$. This shows that (15) holds. A similar observation yields (16). The equations (17)–(19) are simply mass conservation equations for the queue and the server station, respectively. Since $v_s(dx)$, $x \in [0, s]$, represents the amount of fluid content in service whose time in service lies in the range $[x, x + dx)$ at time s , and $h^s(x)$ represents the fraction of the amount of fluid content with time in service x (that is, with service time no less than x) that would depart from the system while having time in service in $[x, x + dx)$. Hence, it is natural to expect $\int_{[0, H^s)} h^s(x) v_s(dx)$ to represent the departure rate of fluid content from the fluid system at time s and thus, expect (18) holds. A similar explanation can be applied to (20) except that, to consider the real reneging rate, we can only consider $x < \chi(s)$ since all the fluid content with the time in the system more than $\chi(s)$ has entered service by time s . The equation (24) represents the usual non-idling condition.

By adding (19) and (17) together and using (25), we see that

$$E(t) + X(0) = X(t) + R(t) + D(t). \tag{26}$$

By the representations of E , R and D in (1), (20) and (18), we have from (26) that X is absolutely continuous. In turn, using the fact that $||[n - a]^+ - [n - b]^+| \leq |a - b|$, it is easy to see from (23) and (17) that B and then K are absolutely continuous. So there exists a Lebesgue integrable function κ such that

$$K(t) = \int_0^t \kappa(s) ds, \quad t \geq 0. \tag{27}$$

By (17) and (18), the process K has the following representation:

$$K(t) = B(t) - B(0) + \int_0^t \left(\int_{[0, H^s)} h^s(x) v_s(dx) \right) ds. \tag{28}$$

Then it follows from the same argument as in deriving (3.12) of [11] that the process κ satisfies for a.e. $t \in \mathbb{R}_+$,

$$\kappa(t) = \begin{cases} \lambda(t) & \text{if } X(t) < 1, \\ \lambda(t) \wedge \int_{[0, H^s)} h^s(x) v_t(dx) & \text{if } X(t) = 1, \\ \int_{[0, H^s)} h^s(x) v_t(dx) & \text{if } X(t) > 1. \end{cases} \tag{29}$$

Remark 2. (Existence and uniqueness of Kang-Ramanan’s fluid model.) Under the assumptions that the hazard rate functions h^r and h^s are either bounded or lower semi-continuous, Kang and Ramanan [10] established the existence of the measure-valued fluid model in Definition 2 by proving an FWLLN and also showed its uniqueness via the fluid model characterization. The existence and uniqueness of Kang-Ramanan’s fluid model directly from the characterization of its solution is established in Kang [7], under the weaker assumptions that the service time distribution G^s has density and the hazard rate function h^r is a.e. locally bounded.

Now we state our first result on the equivalence between the two fluid models described in Definitions 1 and 2. Its proof is deferred to Section 4. As a consequence, it also gives a proof for Conjecture 2.2 of [21] under the assumption that the service and patience time distributions have densities and h^r is a.e. locally bounded.

Theorem 1. *Existence and uniqueness of Whitt’s fluid model in Definition 1 is equivalent to existence and uniqueness of Kang-Ramanan’s fluid model in Definition 2 for the $G_t/GI/N + GI$ queue with the time-dependent arrival rate $\lambda(\cdot)$ and the initial data $(\eta_0, v_0, X(0)) \in \mathcal{S}_0$, where $\eta_0(dx) = \tilde{q}(0, x)dx = q(0, x)dx$ and $v_0(dx) = b(0, x)dx$.*

2.3 Zuñiga’s fluid model

Recently, Zuñiga [23] extended Kang-Ramanan’s fluid model without assuming that the patience time distribution G^r and service time distribution G^s have densities. In this section, we state this extended Kang-Ramanan’s fluid model and establish some useful properties on certain quantities in the model, which are needed in the subsequent analysis.

Define a measure M^r on $[0, H^r]$ by

$$dM^r(x) \doteq \mathbf{1}_{\{x < H^r\}} \bar{G}^r(x-)^{-1} dG^r(x) + \mathbf{1}_{\{G^r(H^r-) < 1\}} \delta_{H^r}(dx),$$

and a measure M^s on $[0, H^s]$ by

$$dM^s(x) \doteq \mathbf{1}_{\{x < H^s\}} \bar{G}^s(x-)^{-1} dG^s(x) + \mathbf{1}_{\{G^s(H^s-) < 1\}} \delta_{H^s}(dx).$$

Note that in Zuñiga [23], it is assumed that G^r is continuous on $[0, \infty)$ in Assumption 2.1 therein. Thus, in Zuñiga’s fluid model stated in Definition 3.4 of [23], the measure M^r (in Definition 3.4 of [23], the author uses H^r instead), the extra term $\mathbf{1}_{\{G^r(H^r-) < 1\}} \delta_{H^r}(dx)$ is not needed. Since here we do not make the continuity assumption on G^r in the following definition, we need to add this term just like the similar term in M^s .

Definition 3. A triple of processes $(\eta, v, X) \in \mathcal{D}_{[0, \infty)}^{abs}(\mathcal{M}[0, H^r]) \times \mathcal{D}_{[0, \infty)}^{abs}(\mathcal{M}[0, H^s]) \times \mathcal{D}_{[0, \infty)}(\mathbb{R})$ is a solution to an extended Kang-Ramanan’s measure-valued fluid model with the input data $(\lambda(\cdot), \eta_0, v_0, X(0))$ such that $(\eta_0, v_0, X(0)) \in \mathcal{S}_0$ if q_t and p_t , the densities of $\int_0^t v_s(\cdot) ds$ and $\int_0^t \eta_s(\cdot) ds$, respectively, satisfy the following conditions. There exist $K(\cdot)$, a process of bounded variation started at 0, $\chi(\cdot)$, $B(\cdot)$, $Q(\cdot)$, $D(\cdot)$, $R(\cdot)$ such that for every $\psi \in \mathcal{C}_b(\mathbb{R}_+)$ and $t \geq 0$, (15)–(17), (19), (21)–(24) hold and

$$D(t) = \int_{[0, H^s]} q_t(x) dM^s(x), \tag{30}$$

$$R(t) = \int_{[0, H^r]} \int_{[0, t]} \mathbf{1}_{\{x \leq \chi(s)\}} d_s p_s(x) dM^r(x), \tag{31}$$

where the integral with respect to $p_s(x)$ is defined as a Lebesgue-Stieltjes integral in s .

Remark 3. Zuñiga’s fluid model stated in Definition 3 is equivalent to Definition 3.4 of [23] due to Lemma 4.1 and Remark 4.2 of [23] and the given input data $(\lambda(\cdot), \eta_0, v_0, X(0))$. The main difference of Zuñiga’s fluid model from Kang-Ramanan’s fluid model in Definition 2 is that the processes D and R satisfy (30) and (31) instead of (18) and (20) due to the lack of existence of densities of G^s and G^r , respectively. By Lemma 4.1 of [23], the densities q_t and p_t can be written as

$$q_t(x) = \bar{G}^s(x-)K((t-x)^+) + \int_{[(x-t)^+, x]} \frac{\bar{G}^s(x-)}{\bar{G}^s(y)} v_0(dy), \tag{32}$$

and

$$p_t(x) = \bar{G}^r(x-)E((t-x)^+) + \int_{[(x-t)^+, x]} \frac{\bar{G}^r(x-)}{\bar{G}^r(y)} \eta_0(dy). \tag{33}$$

When G^s and G^r are assumed to have densities, g^r and g^s , respectively, Zuñiga’s fluid model is reduced to Kang-Ramanan’s fluid model. Zuñiga’s fluid model admits a unique solution (established in Theorem 3.5 via an FWLLN and Theorem 4.4 via the characterization of the fluid model in [23]) under the assumptions that G^r is continuous, η_0 is diffuse, and v_0 is diffuse if G^s is not continuous (Assumption 3.1 of [23]).

We end this section by showing the following critical lemma for Zuñiga’s fluid model in Definition 3, which will be used in Section 3 in discussing the relationship

of the fluid models tracking elapsed times stated in this section and a fluid model tracking residual times stated in Section 3.2.

Lemma 1. *In Definition 3, the processes D and R have the following representations: for each $t \geq 0$,*

$$D(t) = \int_{[0, H^s)} \frac{G^s(y+t) - G^s(y)}{\bar{G}^s(y)} v_0(dy) + \int_0^t G^s(t-s) dK(s), \quad (34)$$

$$\begin{aligned} R(t) = & \int_{[0, H^r)} \left(\int_{(y, y+t]} \mathbf{1}_{\{y \leq \chi(x-y) - (x-y)\}} dG^r(x) \right) \bar{G}^r(y)^{-1} \eta_0(dy) \\ & + \int_0^t \int_{[0, H^r)} \mathbf{1}_{\{x \leq s \wedge \chi(s)\}} \lambda(s-x) dG^r(x) ds. \end{aligned} \quad (35)$$

Moreover, the process $K(t)$ is non-decreasing and the process $\chi(t)$ satisfies the following property:

$$\chi(t) - \chi(s) \leq t - s \text{ whenever } 0 \leq s < t < \infty. \quad (36)$$

Remark 4. It is evident that the representation of the process D in (34) implies that $D(t)$ is not absolute continuous when the service time distribution does not have density. Thus, we cannot write the total service rate (departure rate) as in (9). Although the two-parameter processes $B(t, y)$, $\bar{Q}(t, y)$ and $Q(t, y)$ can be obtained as in (2) from the Zuñiga's fluid model (v_t, η_t, X) in Definition 3, their densities with respect to y may not exist and the associated two-parameter fluid model using densities $b(t, x)$ and $q(t, x)$ cannot be formulated with the densities as in Definition 1.

Proof of Lemma 1. By (30) and (32), applying interchange of the order of integration and integration by parts, we easily obtain (34). To show $R(t)$ in (35), from (31) and (33), we obtain that for each $t \geq 0$,

$$\begin{aligned} R(t) &= \int_{[0, H^r)} \int_0^t \mathbf{1}_{\{x \leq \chi(s) \wedge s\}} \bar{G}^r(x-) \lambda(s-x) ds dM^r(x) \\ &+ \int_{[0, H^r)} \int_{[0, t]} \mathbf{1}_{\{x \leq \chi(s)\}} d_s \left(\int_{[(x-s)^+, x]} \frac{\bar{G}^r(x-)}{\bar{G}^r(y)} \eta_0(dy) \right) dM^r(x) \\ &= \int_0^t \int_{[0, H^r)} \mathbf{1}_{\{x \leq \chi(s) \wedge s\}} \lambda(s-x) \bar{G}^r(x-) dM^r(x) ds \\ &+ \int_{[0, H^r)} \int_{[0, x \wedge t]} \mathbf{1}_{\{x \leq \chi(s)\}} d_s \left(\int_{[x-s, H^r)} \mathbf{1}_{\{y < x\}} \frac{\bar{G}^r(x-)}{\bar{G}^r(y)} \eta_0(dy) \right) dM^r(x) \\ &= \int_0^t \int_{[0, H^r)} \mathbf{1}_{\{x \leq \chi(s) \wedge s\}} \lambda(s-x) \bar{G}^r(x-) dM^r(x) ds \\ &+ \int_{[0, H^r)} \int_{[[x-t]^+, x]} \mathbf{1}_{\{x \leq \chi(x-s)\}} \mathbf{1}_{\{s < x\}} \frac{\bar{G}^r(x-)}{\bar{G}^r(s)} \eta_0(ds) dM^r(x), \end{aligned} \quad (37)$$

$$\begin{aligned}
 &= \int_0^t \int_{[0, H^r]} \mathbf{1}_{\{x \leq s \wedge \chi(s)\}} \lambda(s-x) \bar{G}^r(x-) dM^r(x) ds \\
 &\quad + \int_{[0, H^r]} \left(\int_{(y, y+t]} \mathbf{1}_{\{y \leq \chi(x-y) - (x-y)\}} \bar{G}^r(x-) dM^r(x) \right) \bar{G}^r(y)^{-1} \eta_0(dy) \\
 &= \int_0^t \int_{[0, H^r]} \mathbf{1}_{\{x \leq s \wedge \chi(s)\}} \lambda(s-x) dG^r(x) ds \\
 &\quad + \int_{[0, H^r]} \left(\int_{(y, y+t]} \mathbf{1}_{\{y \leq \chi(x-y) - (x-y)\}} dG^r(x) \right) \bar{G}^r(y)^{-1} \eta_0(dy),
 \end{aligned}$$

where the second term in the second equality follows from Theorem 3.6.1 of [1] with $X = [[x-t]^+, x]$ $Y = [0, x \wedge t]$, $f(s) = x - s$ and μ such that $\mu[a, b] = \int_{[a, b]} \frac{\mathbf{1}_{\{y < x\}}}{\bar{G}^r(y)} \eta_0(dy)$ and the last equality follows from the interchange of the order of integrations.

We next prove the non-decreasing property of $K(t)$. It follows from this representation of $R(t)$ in (35) that Lemma 4.4 of [10] holds, that is, for any $0 \leq a \leq b < \infty$, if $Q(t) = 0$ (equivalently, $\chi(t) = 0$) for all $t \in [a, b]$, then $R(b) - R(a) = 0$. Then the proof for the non-decreasing property of $K(t)$ will follow the same argument in Lemma 4.5 in [10] using (37).

We now prove the property of $\chi(t)$ in (36). By a similar argument as in Lemma 3.4 of [10] on time shifts, to prove the lemma, without loss of generality, we may assume that $s = 0$ in (36). Suppose that the property of $\chi(t)$ in (36) does not hold, that is, there is a time $t_2 > 0$ such that $\chi(t_2) > \chi(0) + t_2$. Let

$$t_1 \doteq \sup\{u \leq t_2 : \chi(u) \leq \chi(0) + u\}.$$

Then $\chi(t_1-) \leq \chi(0) + t_1$ and for each $u \in [t_1, t_2]$,

$$\chi(u) \geq \chi(0) + u \geq \chi(t_1-) + (u - t_1) \text{ and } \chi(t_2) > \chi(t_1-) + (t_2 - t_1). \tag{38}$$

By (37), it is clear that $R(t) - R(t-) \geq 0$ for each $t > 0$. By applying the above display and time shift at t_1 , we have

$$\begin{aligned}
 &R(t_2) - R(t_1) \\
 &= \int_{[0, H^r]} \left(\int_{(y, y+t_2-t_1]} \mathbf{1}_{\{y \leq \chi(t_1+x-y) - (x-y)\}} dG^r(x) \right) \bar{G}^r(y)^{-1} \eta_{t_1}(dy) \\
 &\quad + \int_0^{t_2-t_1} \left(\int_{[0, H^r]} \mathbf{1}_{\{u \leq s \wedge \chi(t_1+s)\}} \lambda(t_1+s-u) dG^r(u) \right) ds.
 \end{aligned}$$

It follows from (38) that $s \wedge \chi(t_1+s) = s$ and $\chi(t_1+s) - s \geq \chi(t_1-)$ for each $s \in (0, t_2 - t_1]$. Hence the above display implies that

$$\begin{aligned}
 &R(t_2) - R(t_1-) \\
 &\geq \int_{[0, H^r]} \left(\int_{(y, y+t_2-t_1]} \mathbf{1}_{\{y \leq \chi(t_1-)\}} dG^r(x) \right) \bar{G}^r(y)^{-1} \eta_{t_1}(dy)
 \end{aligned}$$

$$\begin{aligned}
 & + \int_0^{t_2-t_1} \left(\int_{[0, H^r]} \mathbf{1}_{\{u \leq s\}} \lambda(t_1 + s - u) dG^r(u) \right) ds \\
 = & \int_{[0, H^r]} \mathbf{1}_{\{y \leq \chi(t_1-)\}} \frac{G^r(y + t_2 - t_1) - G^r(y)}{\bar{G}^r(y)} \eta_{t_1}(dy) \\
 & + \int_0^{t_2-t_1} G^r(t_2 - t_1 - u) \lambda(t_1 + u) du,
 \end{aligned}$$

where and the second term on the right hand side of the last display follows from Proposition 0.4.5 of [20]. Since (19) holds for Zuñiga’s fluid model (η, ν, X) and K is non-decreasing, then the above three displays imply that

$$\begin{aligned}
 Q(t_2) & = Q(t_1-) + (E(t_2) - E(t_1-)) - (R(t_2) - R(t_1-)) - (K(t_2) - K(t_1-)) \\
 & \leq \eta_{t_1}[0, \chi(t_1-)] + \int_{t_1}^{t_2} \lambda(u) du - \int_0^{t_2-t_1} G^r(t_2 - t_1 - u) \lambda(t_1 + u) du \\
 & \quad - \int_{[0, H^r]} \mathbf{1}_{[0, \chi(t_1-)]}(x) \frac{G^r(x + (t_2 - t_1)) - G^r(x)}{\bar{G}^r(x)} \eta_{t_1}(dx) \\
 & = \int_0^{t_2-t_1} \bar{G}^r(t_2 - t_1 - u) \lambda(t_1 + u) du \\
 & \quad + \int_{[0, H^r]} \mathbf{1}_{[0, \chi(t_1-)+(t_2-t_1)]}(x + (t_2 - t_1)) \frac{\bar{G}^r(x + (t_2 - t_1))}{\bar{G}^r(x)} \eta_{t_1}(dx) \\
 & = \eta_{t_2}[0, \chi(t_1-) + (t_2 - t_1)],
 \end{aligned}$$

where the last inequality follows from (15). From this and the definition of χ , we have $\chi(t_2) \leq \chi(t_1-) + (t_2 - t_1)$, which contradicts (38). Thus, the lemma is proved. ■

3 Measure-valued fluid models tracking residual times

We first state the two measure-valued processes tracking residual times that arise from Zuñiga’s fluid model for the same $G_t/GI/N + GI$ queueing system in Section 3.1. Here we do not define a new fluid model tracking residual times, but only introduce the two processes themselves. We then state Zhang’s fluid model in Section 3.2, and discuss its connection with the three fluid models tracking elapsed times in Section 3.3. The two measure-valued processes tracking residual times introduced in Section 3.1 play an important bridging role in making the connection.

3.1 Measure-valued processes tracking residual times from Zuñiga’s fluid model

Zuñiga’s fluid model naturally gives rise to the following two measure-valued processes v_t^ℓ and η_t^ℓ . For each $t \geq 0$, clearly the mapping

$$\begin{aligned} \psi \mapsto & \int_{[0, H^s)} \left(\int_{(y+t, \infty)} \frac{\psi(x-y-t)}{\bar{G}^s(y)} dG^s(x) \right) v_0(dy) \\ & + \int_{[0, t]} \left(\int_{(t-s, \infty)} \psi(x-t+s) dG^s(x) \right) dK(s) \end{aligned}$$

is a positive linear functional on $\mathcal{C}_c(\mathbb{R}_+)$ since K is non-decreasing by Lemma 1. Then by Riesz-Markov-Kakutani representation theorem, there is a unique regular Borel measure v_t^ℓ with support $[0, H^s)$ such that for every $\psi \in \mathcal{C}_b(\mathbb{R}_+)$,

$$\begin{aligned} \int_0^\infty \psi(x) v_t^\ell(dx) = & \int_{[0, H^s)} \left(\int_{(y+t, \infty)} \frac{\psi(x-y-t)}{\bar{G}^s(y)} dG^s(x) \right) v_0(dy) \\ & + \int_0^t \left(\int_{(t-s, \infty)} \psi(x-t+s) dG^s(x) \right) dK(s). \end{aligned} \tag{1}$$

Similarly, for each $t \geq 0$, there is a unique regular Borel measure η_t^ℓ with support $[0, H^r)$ such that for every $\psi \in \mathcal{C}_b(\mathbb{R}_+)$,

$$\begin{aligned} \int_0^\infty \psi(x) \eta_t^\ell(dx) = & \mathbf{1}_{\{\zeta(t) \leq 0\}} \int_{[0, -\zeta(t)]} \left(\int_{(y+t, \infty)} \frac{\psi(x-y-t)}{\bar{G}^r(y)} dG^r(x) \right) \eta_0(dy) \\ & + \int_{\zeta^+(t)}^t \left(\int_{(t-s, \infty)} \psi(x-t+s) dG^r(x) \right) \lambda(s) ds, \end{aligned} \tag{2}$$

where

$$\zeta(t) = t - \chi(t). \tag{3}$$

Since $\chi(t)$ represents the elapsed patience time of the fluid content of customers that has been in queue the longest at time t , then the quantity $\zeta(t)$ can be interpreted as the arrival time of the fluid content of customers that has been in queue the longest at time t . It is clear that $\zeta(t) \leq t$ for each $t \geq 0$. At time 0, $\zeta(0) = -\chi(0)$ represents the arrival time of the oldest fluid content in queue initially, and thus, it follow from (21) that

$$\zeta(0) = -\inf\{x \in [0, H^r) : \eta_0[0, x] \geq X(0) - v_0[0, H^s)\}. \tag{4}$$

We first argue that v^ℓ and η^ℓ are two measure-valued processes tracking residual times of fluid content of customers in service and in queue, respectively.

For each $z \geq 0$, by plugging $\psi(x) = \mathbf{1}_{(z, \infty)}(x)$ into (1) and (2), we have

$$v_t^\ell(z, \infty) = \int_{[0, H^s)} \frac{\bar{G}^s(y+t+z)}{\bar{G}^s(y)} v_0(dy) \tag{5}$$

$$\begin{aligned}
 & + \int_{[0,t]} \bar{G}^s(t-s+z)dK(s), \\
 \eta_t^\ell(z, \infty) = & \mathbf{1}_{\{\zeta(t) \leq 0\}} \int_{[0, -\zeta(t)]} \frac{\bar{G}^r(y+t+z)}{\bar{G}^r(y)} \eta_0(dy) \\
 & + \int_{\zeta^+(t)}^t \bar{G}^r(t-s+z)\lambda(s)ds.
 \end{aligned} \tag{6}$$

By (15) and (16), we obtain

$$\begin{aligned}
 \eta_{t+z}[z, \chi(t) + z] = & \int_{[0, H^r)} \mathbf{1}_{[z, \chi(t)+z]}(y+t+z) \frac{\bar{G}^r(y+t+z)}{\bar{G}^r(y)} v_0(dy) \\
 & + \int_0^t \mathbf{1}_{[z, \chi(t)+z]}(t+z-s) \bar{G}^r(t+z-s)\lambda(s)ds \\
 = & \mathbf{1}_{\{\zeta(t) \leq 0\}} \int_{[0, -\zeta(t)]} \frac{\bar{G}^r(y+t+z)}{\bar{G}^r(y)} \eta_0(dy) \\
 & + \int_{\zeta^+(t)}^t \bar{G}^r(t-s+z)\lambda(s)ds,
 \end{aligned}$$

and

$$v_{t+z}[z, \infty) = \int_{[0, H^s)} \frac{\bar{G}^s(y+t+z)}{\bar{G}^s(y)} v_0(dy) + \int_{[0,t]} \bar{G}^s(t+z-s)dK(s).$$

Hence, we obtained the following *coupling* property between (v, η) and (v^ℓ, η^ℓ) :

$$v_t^\ell(z, \infty) = v_{t+z}[z, \infty) \text{ and } \eta_t^\ell(z, \infty) = \eta_{t+z}[z, \chi(t) + z], \text{ quad } z \geq 0. \tag{7}$$

Intuitively, $v_{t+z}[z, \infty)$ represents the amount of fluid content in service at time $t + z$ with elapsed service time at least z , which is precisely the amount of fluid content in service at time t that will still be in service at time $t + z$ and then is equal to the amount of fluid content in service at time t that has residual service time greater than z . (Note that the fluid content in service at time t that has residual service time exactly equal to z will depart from service and hence will not be in service at time $t + z$.) Thus, by the first equality in (7), $v_t^\ell(z, \infty)$ represents the amount of fluid content in service at time t that has residual service time greater than z , that is, v_t^ℓ keeps track of the residual time of fluid content in service at time t . Similarly, $\eta_{t+z}[z, \chi(t) + z]$ represents the amount of fluid content in the potential queue at time $t + z$ with elapsed patience time between z and $\chi(t) + z$, which is precisely the amount of fluid content in queue at time t that will not abandon by time $t + z$. This amount of fluid content is equal to the amount of fluid content that has residual patience time more than z units of time at time t and then is represented by $\eta_t^\ell(z, \infty)$ by the second equality in (7). Then η_t^ℓ keeps track of the residual patience times of customers in queue at time t .

When $t = 0$, (5), (6) and (7) become: for each $z \geq 0$,

$$v_0^\ell(z, \infty) = \int_{[0, H^s)} \frac{\bar{G}^s(y+z)}{\bar{G}^s(y)} v_0(dy) = v_z[z, \infty), \tag{8}$$

$$\eta_0^\ell(z, \infty) = \mathbf{1}_{\{\zeta(0) \leq 0\}} \int_{[0, -\zeta(0)]} \frac{\bar{G}^r(y+z)}{\bar{G}^r(y)} \eta_0(dy) = \eta_z[z, \chi(0) + z]. \tag{9}$$

Remark 5. When G^r and G^s have densities g^r and g^s , respectively, (1) and (2) are equivalent to the following representations:

$$\begin{aligned} \int_0^\infty \psi(x) v_t^\ell(dx) &= \int_{[0, H^s)} \left(\int_0^\infty \frac{g^s(y+t+x)}{\bar{G}^s(y)} \psi(x) dx \right) v_0(dy) \\ &\quad + \int_{[0, t]} \left(\int_0^\infty g^s(t-s+x) \psi(x) dx \right) dK(s), \end{aligned} \tag{10}$$

$$\begin{aligned} \int_0^\infty \psi(x) \eta_t^\ell(dx) &= \mathbf{1}_{\{\zeta(t) \leq 0\}} \int_{[0, -\zeta(t)]} \left(\int_0^\infty \frac{g^r(y+t+x)}{\bar{G}^r(y)} \psi(x) dx \right) \eta_0(dy) \\ &\quad + \int_{\zeta^+(t)}^t \left(\int_0^\infty g^r(t-s+x) \psi(x) dx \right) \lambda(s) ds. \end{aligned} \tag{11}$$

In this case, for each $t \geq 0$, the two measures η_t^ℓ and v_t^ℓ have densities $b_\ell(t, x)$ and $q_\ell(t, x)$, respectively, which can be expressed as

$$b_\ell(t, y) = \int_{[0, H^s)} \frac{g^s(x+t+y)}{\bar{G}^s(x)} v_0(dx) + \int_{[0, t]} g^s(y+t-u) dK(u), \tag{12}$$

and

$$q_\ell(t, y) = \mathbf{1}_{\{\zeta(t) \leq 0\}} \int_{[0, -\zeta(t)]} \frac{g^r(x+t+y)}{\bar{G}^r(x)} \eta_0(dx) + \int_{\zeta^+(t)}^t g^r(y+t-u) \lambda(u) du. \tag{13}$$

3.2 Zhang’s fluid model

Zhang [22] uses a so-called *virtual queue* to describe the queueing dynamics, instead of the potential queue used in the three fluid models in Section 2. In the definitions of both potential and virtual queues, all customers enter them upon arrival. *The difference between them lies in how customers depart.* Customers can leave the potential queue only when their patience expires, that is, at the instant when their remaining patience times are zeros. Whereas, customers can only leave the virtual queue in their turns of service. Customers in the virtual queue may have already run out of patience (i.e., the remaining patience time is negative) at their turns of service. Whenever a server becomes free, the server will check the oldest customer in the virtual queue. If the customer being checked has not abandoned yet (its remaining patience time is still positive), then the server will start serving this customer

and this customer is removed from the virtual queue, and otherwise, this customer is simply removed from the virtual queue and the server will turn to check the next oldest customer. We now state Zhang’s fluid model.

Definition 4. (Zhang’s fluid model in [22].) Assume that the fluid arrival rate $\lambda(t) = \lambda$ for each $t \geq 0$, where $\lambda > 0$ is a constant. A pair of measure-valued processes $(\mathcal{R}, \mathcal{L})$ is a solution to the fluid model if the following conditions are satisfied:

(i) $(\mathcal{R}, \mathcal{L})$ satisfies the following two equations:

$$\mathcal{R}_t(C_x) = \lambda \int_{t-Q_v(t)/\lambda}^t \bar{G}^r(t+x-s)ds, \quad x \in \mathbb{R}, \tag{14}$$

and

$$\mathcal{L}_t(C_x) = \mathcal{L}_0(C_x+t) + \int_0^t \bar{G}^r(Q_v(s)/\lambda)\bar{G}^s(t+x-s)dL_v(s), \quad x \in \mathbb{R}_+, \tag{15}$$

where $C_x \doteq (x, \infty)$ for $x \in \mathbb{R}$, $Q_v(t) = \mathcal{R}_t(\mathbb{R})$ is of bounded variation and $L_v(t) = \lambda t - Q_v(t)$;

(ii) the non-idling conditions in (23) and (24) hold for $B(t) = \mathcal{L}_t(\mathbb{R}_+)$, $Q(t) = \mathcal{R}_t(\mathbb{R}_+)$ and $X(t) = B(t) + Q(t)$;

(iii) the initial condition $(\mathcal{R}_0, \mathcal{L}_0)$ satisfies

$$\mathcal{R}_0(C_x) = \lambda \int_0^{Q_v(0)/\lambda} \bar{G}^r(x+s)ds, \quad x \in \mathbb{R}, \text{ and } \mathcal{L}_0(\{0\}) = 0, \tag{16}$$

and the non-idling condition at time 0 in (23) and (24).

In Zhang’s fluid model, $\mathcal{R}_t(C_x)$ can be interpreted as the fluid content of customers in the virtual queue with residual patience times strictly bigger than x and $\mathcal{L}_t(C_x)$ can be interpreted as the fluid content of customers in service with residual service times strictly bigger than x at each time t . Then $Q(t)$, $B(t)$, $Q_v(t)$ and $L_v(t)$ represent, respectively, the total fluid content of the real queue at time t , the total fluid content of customers in service at time t , the total fluid content in the virtual queue at time t , and the cumulative customers removed from the virtual queue by time t . The existence and uniqueness of Zhang’s fluid model are proved in Theorem 3.1 of [22] by an explicit characterization of its solution, under the assumptions that the service time distribution G^s is continuous and the patience time distribution G^r is Lipschitz continuous.

3.3 Connection between Zhang’s fluid model and the three fluid models in Section 2

Among the three fluid models in Section 2, we have showed in Theorem 1 that Whitt’s fluid model is equivalent to Kang-Ramanan fluid model and in Remark 3

that Zuñiga’s fluid model extends Kang-Ramanan’s fluid model by relaxing the assumption on the existence of densities of G^r and G^s . Since Zhang’s fluid model keeps track of customers’ residual times and does not need G^r and G^s to have densities ([22] does assume that G^s is continuous and G^r is Lipschitz continuous to establish existence and uniqueness), while Zuñiga’s fluid model keeps track of customers’ elapsed times and also does not need G^r and G^s to have densities, it is natural to question if Zhang’s fluid model and Zuñiga’s fluid model are in fact equivalent in describing system dynamics of the same $G_t/GI/N + GI$ queues. If so, this will enable researchers to borrow results from either one of the two to study the system performance of $G_t/GI/N + GI$ queues.

In this section we provide a detailed discussion on Zhang’s fluid model in connection with Zuñiga’s fluid model (and hence Kang-Ramanan’s fluid model and Whitt’s fluid model). The three fluid models in Section 2 allow time-varying arrival rate $\lambda(\cdot)$, whereas, Zhang’s fluid model requires a constant arrival rate λ . Thus the discussion in this section will focus on the three formulations with a constant arrival rate. It is important that these formulations must have the same system input data including the initial conditions when making the comparisons. We first show by a series of remarks that *Zhang’s fluid model is not entirely equivalent to the three fluid models tracking elapsed times for the same $G/GI/N + GI$ queueing system under general initial conditions, that is, Zhang’s fluid model and the three fluid models tracking elapsed times may not be formulated simultaneously for the same $G/GI/N + GI$ queueing system under certain general initial conditions.*

Remark 6. (On the arrival rate.) The imposed condition on \mathcal{R}_0 in Zhang’s fluid model requires that the initial fluid content of customers in the virtual queue depends on the arrival rate λ after time 0, whereas in real life applications, the customers’ arrival patterns before time 0 and after time 0 are likely different. Thus, Zhang’s fluid model may not be appropriate for those applications. In contrast, the three fluid models tracking elapsed times do not have this restriction.

Remark 7. (The initial condition on \mathcal{R}_0 .) Zhang’s fluid model requires that the system initial condition \mathcal{R}_0 satisfies (16), that is,

$$\mathcal{R}_0(C_x) = \lambda \int_0^{Q_v(0)/\lambda} \bar{G}^r(x+s)ds, \quad x \in \mathbb{R}. \tag{17}$$

Let \mathcal{R}_0^+ be the restriction of \mathcal{R}_0 on $[0, \infty)$. Then \mathcal{R}_0^+ keeps track of the residual patience times of the fluid content of customers initially in queue. So if Zhang’s fluid model were equivalent to Zuñiga’s fluid model for the same $G/GI/N + GI$ queueing system assuming a constant arrival rate, we must have $\mathcal{R}_0^+ = \eta_0^\ell$ in (9), that is,

$$\mathcal{R}_0^+(C_x) = \mathbf{1}_{\{\zeta(0) \leq 0\}} \int_{[0, -\zeta(0)]} \frac{\bar{G}^r(y+x)}{\bar{G}^r(y)} \eta_0(dy), \quad \forall x \geq 0,$$

where η_0 is the initial condition for the η in Definition 3, and then

$$\lambda \int_0^{Q_v(0)/\lambda} \bar{G}^r(x+s)ds = \mathbf{1}_{\{\zeta(0)\leq 0\}} \int_{[0,-\zeta(0)]} \frac{\bar{G}^r(y+x)}{\bar{G}^r(y)} \eta_0(dy), \quad \forall x \geq 0. \quad (18)$$

We first note that there may not be a unique η_0 satisfying (18) for the given \mathcal{R}_0 . For example, when G^r has density $g^r(x) = e^{-x}$, $x \in \mathbb{R}_+$,

$$\mathbf{1}_{\{\zeta(0)\leq 0\}} \int_{[0,-\zeta(0)]} \frac{\bar{G}^r(y+x)}{\bar{G}^r(y)} \eta_0(dy) = \mathbf{1}_{\{\zeta(0)\leq 0\}} e^{-x} \eta_0[0, -\zeta(0)],$$

and

$$\lambda \int_0^{Q_v(0)/\lambda} \bar{G}^r(x+s)ds = \lambda e^{-x} (1 - e^{-Q_v(0)/\lambda}).$$

Thus, any η_0 satisfying $\mathbf{1}_{\{\zeta(0)\leq 0\}} \eta_0[0, -\zeta(0)] = \lambda (1 - e^{-Q_v(0)/\lambda})$ will satisfy (18).

Moreover, it is clear that the above display (18) does not hold for an arbitrary initial condition η_0 . For example, if $\eta_0(dx) = \lambda^\dagger \bar{G}^r(x)dx$ for some positive $\lambda^\dagger \neq \lambda$, then

$$\begin{aligned} & \mathbf{1}_{\{\zeta(0)\leq 0\}} \int_{[0,-\zeta(0)]} \frac{\bar{G}^r(y+x)}{\bar{G}^r(y)} \eta_0(dy) \\ &= \lambda^\dagger \mathbf{1}_{\{\zeta(0)\leq 0\}} \int_{[0,-\zeta(0)]} \bar{G}^r(y+x)dy, \end{aligned}$$

which is not equal to $\mathcal{R}_0^+(C_x)$ in (17) even if $-\zeta(0) = Q_v(0)/\lambda$. Thus, for a fluid $G/GI/N + GI$ queueing system with a constant arrival rate λ after time 0, the initial conditions $\eta_0(dx) = \lambda^\dagger \bar{G}^r(x)dx$ for $\lambda^\dagger \neq \lambda$ and $(v_0, X(0))$ such that $(\eta_0, v_0, X(0)) \in \mathcal{S}_0$, Zuñiga’s fluid model can be well formulated, but there is no corresponding Zhang’s fluid model $(\mathcal{R}, \mathcal{Z})$ that describes the same system.

Remark 8. (The initial condition on \mathcal{L}_0 .) Zhang’s fluid model only requires that $\mathcal{L}_0(\{0\}) = 0$. This condition is rather general. We show by an example that for a $G/GI/N + GI$ queueing system, although Zhang’s fluid model can be formulated with that initial condition \mathcal{L}_0 , there may not exist an (unique) initial measure v_0 to formulate a corresponding Zuñiga’s fluid model for the same system.

Consider the service time distribution G^s being exponential with unit rate, that is, $g^s(x) = e^{-x}$, $x \in \mathbb{R}_+$. Let \mathcal{L}_0 be the measure that tracks the residual service times of fluid content of customers initially in service and satisfies $\mathcal{L}_0(\{0\}) = 0$, and assume that Zhang’s fluid model can be formulated with \mathcal{L}_0 . Suppose that Zuñiga’s fluid model can also be formulated for some measure v_0 , which tracks the elapsed service times of fluid content of customers initially in service. By (8), if Zhang’s fluid model and Zuñiga’s fluid model were equivalent, \mathcal{L}_0 and v_0 must satisfy the following equation:

$$\mathcal{L}_0(C_x) = \int_{[0,H^s)} \frac{\bar{G}^s(y+x)}{\bar{G}^s(y)} v_0(dy) = v_0[0, H^s) e^{-x}, \quad x \geq 0.$$

If the given \mathcal{L}_0 satisfies $\mathcal{L}_0(C_x) = ce^{-x}$ for some constant $c > 0$, then any such measure ν_0 in Zuñiga’s fluid model satisfying $c = \nu_0[0, H^s)$ will satisfy the above display. However, on the other hand, if the given \mathcal{L}_0 , satisfying $\mathcal{L}_0(\{0\}) = 0$, does not have an exponential density, then this contradicts the above equation resulting from the equivalence, and implies that no corresponding measure ν_0 can be found for Zuñiga’s fluid model to be well formulated for the given queueing system.

From the discussion in Remarks 6, 7 and 8, it is clear that the class of fluid many-server queueing systems where Zhang’s fluid model can be formulated is *not* the same as the class of fluid many-server queueing systems where Zuñiga’s fluid model (and hence Kang-Ramanan’s fluid model and Whitt’s fluid model) can be formulated.

We next look more closely into the conditions on fluid $G/GI/N + GI$ queueing systems where Zhang’s fluid model and the three fluid models tracking elapsed times can all be used to describe the system dynamics for the same system. To simplify the exposition, we focus on Zhang’s fluid model and Zuñiga’s fluid model. Our findings are stated in the following two theorems.

Theorem 2. *Given a Zhang’s fluid model $(\mathcal{R}, \mathcal{L})$ for a $G/GI/N + GI$ queueing system with arrival rate λ , there exists a Zuñiga’s fluid model (η, ν, X) for the same queueing system with the input data $(\lambda, \eta_0, \nu_0, X(0))$ such that $(\eta_0, \nu_0, X(0)) \in \mathcal{S}_0$ with*

$$\eta_0(dx) \doteq \lambda \mathbf{1}_{[0, Q_v(0)/\lambda]}(x) \bar{G}^r(x) dx, \tag{19}$$

if and only if, for the given \mathcal{L}_0, ν_0 satisfying

$$\mathcal{L}_0(C_x) = \int_{[0, H^s)} \frac{\bar{G}^s(y+x)}{\bar{G}^s(y)} \nu_0(dy), \quad x \geq 0. \tag{20}$$

Proof. The “only if” part follows directly from the discussion in Remark 8. We now focus on “if” part.

Let η_0 be as given in (19). For each $t \geq 0$, the following mapping

$$\psi \mapsto \int_{[0, H^r)} \psi(x+t) \frac{\bar{G}^r(x+t)}{\bar{G}^r(x)} \eta_0(dx) + \lambda \int_0^t \psi(t-s) \bar{G}^r(t-s) ds$$

is a positive linear functional on $\mathcal{C}_c(\mathbb{R}_+)$. Then by Riesz-Markov-Kakutani representation theorem, there is a unique regular Borel measure η_t on \mathbb{R}_+ such that (15) holds. It is clear that η_t has support $[0, H^r)$.

For each $t \geq 0$, define

$$K(t) \doteq \int_0^t \bar{G}^r(Q_v(s)/\lambda) dL_v(s) \text{ and } R(t) \doteq \lambda \int_0^t \bar{G}^r(Q_v(s)/\lambda) ds.$$

Then, for each $t \geq 0$, with the above K and the given ν_0 satisfying (20), the mapping

$$\psi \mapsto \int_{[0, H^s)} \psi(x+t) \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) + \int_{[0, t]} \psi(t-s) \bar{G}^s(t-s) dK(s)$$

is a positive linear functional on $\mathcal{C}_c(\mathbb{R}_+)$. By Riesz-Markov-Kakutani representation theorem, there is a unique regular Borel measure ν_t that satisfies (16). Let B, Q, X be the associated processes in Zhang’s fluid model and for each $t \geq 0$, define

$$D(t) \doteq \int_{[0, H^r]} \frac{G^s(y+t) - G^s(y)}{\bar{G}^s(y)} \nu_0(dy) + \int_0^t G^s(t-s) dK(s).$$

We show that (η, ν, X) satisfies Definition 3.

From (14), it is clear that

$$Q(t) = \mathcal{R}_t(\mathbb{R}_+) = \lambda \int_{t-Q_v(t)/\lambda}^t \bar{G}^r(t-s) ds = \lambda G_d^r(Q_v(t)/\lambda), \tag{21}$$

where $G_d^r(x) = \int_0^x \bar{G}^r(s) ds$. It is established in the proof of Theorem 3.1 of [22] that $Q(t)/\lambda < G_d^r(\infty) = G_d^r(H^r)$. Then it follows that $Q_v(t)/\lambda < H^r$. Since Q_v is of bounded variation by (14), it follows that Q is also of bounded variation and by the chain rule formula (Proposition 4.6 in Chapter 0 of [20])

$$Q(t) = Q(0) + \int_0^t \bar{G}^r(Q_v(s)/\lambda) dQ_v(s).$$

Thus, by the definition of K and the above display for $Q(t)$,

$$\begin{aligned} K(t) &= Q(0) - \left(Q(t) - \lambda \int_0^t \bar{G}^r(Q_v(s)/\lambda) ds \right) \\ &= Q(0) - \left(Q(t) - \lambda \int_0^t \bar{G}^r((G_d^r)^{-1}(Q(s)/\lambda)) ds \right). \end{aligned}$$

Then it follows from Lemma A.3 of [22] that K is non-decreasing. Simple calculation also shows that

$$\begin{aligned} &Q(t) + K(t) + R(t) \\ &= Q(0) + \int_0^t \bar{G}^r(Q_v(s)/\lambda) dQ_v(s) + \int_0^t \bar{G}^r(Q_v(s)/\lambda) d(\lambda s - Q_v(s)) \\ &\quad + \lambda \int_0^t G^r(Q_v(s)/\lambda) ds \\ &= Q(0) + \lambda t, \end{aligned}$$

which establishes (19). For each $t \geq 0$, define $\chi(t)$ by the right hand side of (21). It follows from the construction of η_t and the given η_0 in (19) that

$$\begin{aligned} Q(t) = \eta_t[0, \chi(t)] &= \lambda \int_0^{[\chi(t)-t]^+ \wedge Q_v(0)/\lambda} \bar{G}^r(x+t) dx \\ &\quad + \lambda \int_{[t-\chi(t)]^+}^t \bar{G}^r(t-s) ds. \end{aligned} \tag{22}$$

When $\chi(t) > t$, the above display is reduced to

$$Q(t)/\lambda = \int_0^{\chi(t) \wedge (t + Q_v(0)/\lambda)} \bar{G}^r(s) ds.$$

Comparing this with (21), we have $Q_v(t)/\lambda = \chi(t) \wedge (t + Q_v(0)/\lambda)$. When $\chi(t) \leq t$, the display in (22) is reduced to $Q(t)/\lambda = \int_0^{\chi(t)} \bar{G}^r(s) ds$ and hence $Q_v(t)/\lambda = \chi(t)$. Combining the two cases, we have for each $t \geq 0$,

$$Q_v(t)/\lambda = \chi(t) \wedge (t + Q_v(0)/\lambda). \tag{23}$$

For each $t \geq 0$, it follows from (19) and the definition of M^r that

$$\begin{aligned} & \int_{[0, H^r]} \int_{[0, t]} \mathbf{1}_{\{x \leq \chi(s)\}} ds \left(\int_{[(x-s)^+, x]} \frac{\bar{G}^r(x-)}{\bar{G}^r(y)} \eta_0(dy) \right) dM^r(x) \\ & + \lambda \int_{[0, H^r]} \int_{[0, t]} \mathbf{1}_{\{x \leq \chi(s) \wedge s\}} \bar{G}^r(x-) ds dM^r(x) \\ = & \lambda \int_{[0, H^r]} \int_{[0, x \wedge t]} \mathbf{1}_{\{x \leq \chi(s)\}} ds \left(\int_{[x-s, x]} \frac{\bar{G}^r(x-)}{\bar{G}^r(y)} \mathbf{1}_{[0, Q_v(0)/\lambda]}(y) \bar{G}^r(y) dy \right) dM^r(x) \\ & + \lambda \int_{[0, t]} \int_{[0, H^r]} \mathbf{1}_{\{x \leq \chi(s) \wedge s\}} dG^r(x) ds \\ = & \lambda \int_{[0, H^r]} \int_0^{t \wedge x} \mathbf{1}_{\{x \leq \chi(s)\}} ds \left(\int_{[x-s, x]} \mathbf{1}_{[0, Q_v(0)/\lambda]}(y) dy \right) dG^r(x) \\ & + \lambda \int_0^t G^r(\chi(s) \wedge s) ds \\ = & \lambda \int_{[0, H^r]} \int_0^t \mathbf{1}_{\{s \leq x \leq \chi(s)\}} \mathbf{1}_{[0, Q_v(0)/\lambda]}(x-s) ds dG^r(x) \\ & + \lambda \int_0^t G^r(\chi(s) \wedge s) ds \\ = & \lambda \int_0^t G^r(\chi(s) \wedge (s + Q_v(0)/\lambda)) ds = \lambda \int_0^t G^r(Q_v(s)/\lambda) ds, \end{aligned}$$

where the second to the last equality follows from the fact that

$$\mathbf{1}_{\{s \leq x \leq \chi(s)\}} \mathbf{1}_{[0, Q_v(0)/\lambda]}(x-s) = \mathbf{1}_{\{s \leq x \leq \chi(s) \wedge (s + Q_v(0)/\lambda)\}} = \mathbf{1}_{\{s \wedge \chi(s) \leq x \leq \chi(s) \wedge (s + Q_v(0)/\lambda)\}},$$

and the last equality follows from (23). This, together with the definition of $R(t)$ and (33), implies that (31) holds.

By using (4.5) of [23], we obtain

$$\begin{aligned} & \int_{[0, H^s]} \left(\bar{G}^s(x-) K([t-x]^+) + \int_{[x-t]^+}^x \frac{\bar{G}^s(x-)}{\bar{G}^s(y)} v_0(dy) \right) dM^s(x) \\ = & \int_{[0, H^s]} \frac{G^s(y+t) - G^s(y)}{\bar{G}^s(y)} v_0(dy) + \int_0^t G^s(t-s) dK(s) \end{aligned}$$

which is equal to the process $D(t)$ by definition, and implies that (30) holds.

For each $t \geq 0$, (17) holds by applying interchange of the order of integration to (16) and using the definitions of D and B . The properties (22)–(24) follow from property (ii) of Zhang’s fluid model. Thus, this completes the proof that (η, v, X) is a Zuñiga’s fluid model satisfying Definition 3. Clearly from the construction, both the given Zhang’s fluid model and the constructed Zuñiga’s fluid model describe the same $G/GI/N + GI$ queueing system. ■

Theorem 3. *Given a Zuñiga’s fluid model (η, v, X) for a $G/GI/N + GI$ queueing system with the input data $(\lambda, \eta_0, v_0, X(0))$ such that $(\eta_0, v_0, X(0)) \in \mathcal{S}_0$, there exists a Zhang’s fluid model $(\mathcal{R}, \mathcal{Z}^r)$ for the same queueing system with arrival rate λ if and only if η_0 satisfies the following condition: for each $t \geq 0$, there exists a solution z_t , independent of $x \geq 0$, to the equation in z :*

$$\lambda \int_{t \wedge \chi(t)}^z \bar{G}^r(x+s) ds = \mathbf{1}_{\{\chi(t) \geq t\}} \int_{[0, \chi(t)-t]} \frac{\bar{G}^r(y+t+x)}{\bar{G}^r(y)} \eta_0(dy), \tag{24}$$

such that

$$\begin{aligned} \lambda \int_0^t G^r(z_s) ds &= \lambda \int_0^t G^r(\chi(s) \wedge s) ds \\ &+ \int_{[0, H^r)} \left(\int_{(y, y+t]} \mathbf{1}_{\{x \leq \chi(x-y)\}} dG^r(x) \right) \bar{G}^r(y)^{-1} \eta_0(dy). \end{aligned} \tag{25}$$

In this case, \mathcal{Z}_0 can be chosen as defined by (20) for the given v_0 , and \mathcal{R}_0 can be chosen as defined by (16) for $Q_v(0) = z_0 \lambda$, where z_0 is the solution, independent of $x \geq 0$, that satisfies (24) for $t = 0$.

Remark 9. When G^r has a density g^r , the conditions (24) and (25) can be replaced as follows: for each $t \geq 0$, there exists a solution z_t , independent of $x \geq 0$, to the equation in z :

$$\lambda G^r(x+z) = \lambda G^r(x+t \wedge \chi(t)) + \mathbf{1}_{\{\chi(t) \geq t\}} \int_{[0, \chi(t)-t]} \frac{g^r(y+t+x)}{\bar{G}^r(y)} \eta_0(dy). \tag{26}$$

In fact, (24) follows from (26) directly by integrating both sides of (24) in x . It follows from (26) with $x = 0$ that

$$\begin{aligned} &\lambda \int_0^t G^r(\chi(s) \wedge s) ds \\ &+ \int_{[0, H^r)} \left(\int_{(y, y+t]} \mathbf{1}_{\{x \leq \chi(x-y)\}} dG^r(x) \right) \bar{G}^r(y)^{-1} \eta_0(dy) \\ &= \lambda \int_0^t G^r(\chi(s) \wedge s) ds \\ &+ \int_{[0, H^r)} \left(\int_0^t \mathbf{1}_{\{y \leq \chi(x)-x\}} g^r(x+y) dx \right) \bar{G}^r(y)^{-1} \eta_0(dy) \end{aligned}$$

$$\begin{aligned}
 &= \lambda \int_0^t G^r(\chi(s) \wedge s) ds + \int_0^t \lambda (\bar{G}^r(s \wedge \chi(s)) - \bar{G}^r(z_s)) ds \\
 &= \lambda \int_0^t G^r(z_s) ds.
 \end{aligned}$$

Thus, (25) holds.

Proof of Theorem 3. We first show the “only if” part. Recall that in Zhang’s fluid model, \mathcal{R}_t^+ , the restriction of \mathcal{R}_t on $[0, \infty)$, tracks the residual patience times of the fluid content of customers in queue at time t . If there exists a Zhang’s fluid model $(\mathcal{R}, \mathcal{Z})$ to describe the same $G/GI/N + GI$ queueing system together with Zuñiga’s fluid model (η, v, X) , the measure \mathcal{R}_t must satisfies (see (6))

$$\mathcal{R}_t(C_x) = \mathbf{1}_{\{\chi(t) \geq t\}} \int_{[0, \chi(t)-t]} \frac{\bar{G}^r(y+t+x)}{\bar{G}^r(y)} \eta_0(dy) + \lambda \int_0^{t \wedge \chi(t)} \bar{G}^r(s+x) ds,$$

for each $t \geq 0$ and $x \geq 0$, and hence η_0 must satisfy that for each $t \geq 0$ and $x \geq 0$,

$$\lambda \int_{t \wedge \chi(t)}^{Q_v(t)/\lambda} \bar{G}^r(x+s) ds = \mathbf{1}_{\{\chi(t) \geq t\}} \int_{[0, \chi(t)-t]} \frac{\bar{G}^r(y+t+x)}{\bar{G}^r(y)} \eta_0(dy). \tag{27}$$

When $t = 0$, (27) is reduced to

$$\mathcal{R}_0(C_x) = \mathbf{1}_{\{\chi(0) \geq 0\}} \int_{[0, \chi(0)]} \frac{\bar{G}^r(y+x)}{\bar{G}^r(y)} \eta_0(dy), \quad x \geq 0, \tag{28}$$

which is discussed in Remark 7. Moreover, in Zhang’s fluid model, since customers in queue will renege when their residual patience times reach zero, then by differentiating (14) in x and letting $x = 0$, we have the abandonment rate at time t is given by

$$\lambda (\bar{G}^r(x) - \bar{G}^r(x + Q_v(t)/\lambda)) \Big|_{x=0} = G^r(Q_v(t)/\lambda).$$

Then $R(t)$, the cumulative abandonment by time t , is given by $\int_0^t G^r(Q_v(s)/\lambda) ds$. On the other hand, by Zuñiga’s fluid model, $R(t)$ is given by (35). Then η_0 must also satisfy that for each $t > 0$,

$$\begin{aligned}
 \lambda \int_0^t G^r(Q_v(s)/\lambda) ds &= \lambda \int_0^t G^r(\chi(s) \wedge s) ds \\
 &+ \int_{[0, H^r)} \left(\int_{(y, y+t]} \mathbf{1}_{\{x \leq \chi(x-y)\}} dG^r(x) \right) \bar{G}^r(y)^{-1} \eta_0(dy).
 \end{aligned} \tag{29}$$

Note that for each $t \geq 0$, $Q_v(t)$ satisfies (24) and (25), independent of $x \geq 0$. Hence the “only if” part is established.

For the “if” part, let \mathcal{Z}_0 and \mathcal{R}_0 be defined as in the statement of the theorem. It is clear that the defined \mathcal{R}_0 and η_0 satisfy (28) and $(\mathcal{R}_0, \mathcal{Z}_0)$ satisfies property (iii) of Zhang’s fluid model. Let $\chi(t), B(t), Q(t), K(t), D(t), R(t)$ be the associated auxiliary processes from Zuñiga’s fluid model (η, v, X) . For each $t > 0$, define

$$Q_v(t) \doteq \lambda z_t \quad \text{and} \quad L_v(t) \doteq \lambda t - Q_v(t),$$

where z_t is the solution, independent of $x \geq 0$, that satisfies (24) and (25). Define \mathcal{R}_t and \mathcal{Z}_t by the right hand sides of (14) and (15), respectively. We show that the pair of processes $(\mathcal{R}, \mathcal{Z})$ satisfies Zhang's fluid model. In fact, it suffices to verify conditions (i) and (ii) of Zhang's fluid model.

From the definition of $Q_v(t)$, we have for each $x \geq 0$,

$$\lambda \int_{t \wedge \chi(t)}^{Q_v(t)/\lambda} \bar{G}^r(x+s) ds = \mathbf{1}_{\{\chi(t) \geq t\}} \int_{[0, \chi(t)-t]} \frac{\bar{G}^r(y+t+x)}{\bar{G}^r(y)} \eta_0(dy).$$

Combining this, the construction of \mathcal{R} and (15), we have

$$\begin{aligned} \mathcal{R}_t(\mathbb{R}_+) &= \mathbf{1}_{\{\chi(t) \geq t\}} \int_{[0, \chi(t)-t]} \frac{\bar{G}^r(y+t)}{\bar{G}^r(y)} \eta_0(dy) + \lambda \int_0^{t \wedge \chi(t)} \bar{G}^r(s) ds \\ &= \eta_t[0, \chi(t)] = Q(t). \end{aligned} \quad (30)$$

Since $Q(t)$ is of bounded variation by (19), the previous display implies that Q_v and hence L_v are also of bounded variation. Thus, condition (i) of Zhang's fluid model holds.

Next we show that condition (ii) of Zhang's fluid model holds for $B^*(t) = \mathcal{Z}_t(\mathbb{R}_+)$, $Q^*(t) = \mathcal{R}_t(\mathbb{R}_+)$ and $X^*(t) = B^*(t) + Q^*(t)$. Note that for each $t \geq 0$, we have showed that $Q^*(t) = Q(t)$. By using the definition of B^* , the construction of \mathcal{Z} and the property of \mathcal{Z}_0 , we have

$$\begin{aligned} B^*(t) &= \mathcal{Z}_t(\mathbb{R}_+) \\ &= \mathcal{Z}_0(C_t) + \int_0^t \bar{G}^r(Q_v(s)/\lambda) \bar{G}^s(t-s) dL_v(s) \\ &= \int_{[0, H^s]} \frac{\bar{G}^s(y+t)}{\bar{G}^s(y)} \nu_0(dy) + \int_0^t \bar{G}^r(Q_v(s)/\lambda) \bar{G}^s(t-s) dL_v(s). \end{aligned} \quad (31)$$

By (35), (25) and $\lambda(t) = \lambda$ for each $t \geq 0$, we have

$$\begin{aligned} R(t) &= \int_{[0, H^r]} \left(\int_{(y, y+t]} \mathbf{1}_{\{y \leq \chi(x-y) - (x-y)\}} dG^r(x) \right) \bar{G}^r(y)^{-1} \eta_0(dy) \\ &\quad + \lambda \int_0^t \int_{[0, H^r]} \mathbf{1}_{\{x \leq s \wedge \chi(s)\}} dG^r(x) ds \\ &= \int_{[0, H^r]} \left(\int_{(y, y+t]} \mathbf{1}_{\{x \leq \chi(x-y)\}} dG^r(x) \right) \bar{G}^r(y)^{-1} \eta_0(dy) \\ &\quad + \lambda \int_0^t G^r(\chi(s) \wedge s) ds \\ &= \lambda \int_0^t G^r(Q_v(s)/\lambda) ds. \end{aligned}$$

In addition, since $G_d^r(Q_v(t)/\lambda) = Q(t)/\lambda$ by (30), by the chain rule formula,

$$Q(t) = Q(0) + \int_0^t \bar{G}^r(Q_v(s)/\lambda) dQ_v(s).$$

These, together with (19), imply that

$$K(t) = \lambda t + Q(0) - Q(t) - R(t) = \int_0^t \bar{G}^r(Q_v(s)/\lambda) dL_v(s).$$

Hence, by (31), $B^*(t) = B(t)$ and then $X^*(t) = X(t)$. Since B, Q, X satisfy (22)–(24), then B^*, Q^*, X^* satisfy condition (ii) of Zhang’s fluid model. This completes the proof that $(\mathcal{R}, \mathcal{X})$ is a Zhang’s fluid model. Clearly from the construction, both the given Zuñiga’s fluid model and the constructed Zhang’s fluid model describe the same $G/GI/N + GI$ queueing system. ■

Corollary 3.1 *Given a Zuñiga’s fluid model (η, v, X) for a $G/GI/N + GI$ queueing system with the input data $(\lambda, \eta_0, v_0, X(0))$ such that $(\eta_0, v_0, X(0)) \in \mathcal{S}_0$ and $\eta_0(dx) = \lambda \mathbf{1}_{[0,a]}(x) \bar{G}^r(x) dx$ for some $a \geq 0$, then one can construct a Zhang’s fluid model $(\mathcal{R}, \mathcal{X})$ for the same queueing system with arrival rate λ , \mathcal{X}_0 defined by (20) for the given v_0 , and \mathcal{R}_0 defined by (16) for $Q_v(0) = a\lambda$.*

Proof. It suffices to check that the given η_0 satisfies (24) and (25). Note that for the given η_0 , the equation in (24) becomes

$$\int_{t \wedge \chi(t)}^z \bar{G}^r(x+s) ds = \mathbf{1}_{\{\chi(t) \geq t\}} \int_t^{t+(\chi(t)-t) \wedge a} \bar{G}^r(s+x) ds.$$

For $t \geq 0$ such that $\chi(t) \geq t$, we can choose $z_t = t + (\chi(t) - t) \wedge a$ and for $t \geq 0$ such that $\chi(t) < t$, we can choose $z_t = \chi(t)$. Clearly, in either case, z_t does not depend on $x \geq 0$. Now we show that z_t satisfies (25). Note that for the given η_0 , by (37),

$$\begin{aligned} & \int_{[0, H^r)} \left(\int_{(y, y+t]} \mathbf{1}_{\{x \leq \chi(x-y)\}} dG^r(x) \right) \bar{G}^r(y)^{-1} \eta_0(dy) \\ &= \int_{[0, H^r]} \int_{[0, t]} \mathbf{1}_{\{x \leq \chi(s)\}} ds \left(\int_{[(x-s)^+, x]} \frac{\bar{G}^r(x-)}{\bar{G}^r(y)} \eta_0(dy) \right) dM^r(x) \\ &= \lambda \int_{[0, H^r]} \int_{[0, t]} \mathbf{1}_{\{x \leq \chi(s)\}} \left(\int_{(x-s)^+}^x \mathbf{1}_{[0, a]}(y) dy \right) dG^r(x) \\ &= \lambda \int_{[0, H^r]} \int_{[0, t]} \mathbf{1}_{\{x \leq \chi(s)\}} \mathbf{1}_{\{s < x \leq s+a\}} ds dG^r(x) \\ &= \lambda \int_0^t (G^r(\chi(s) \wedge (s+a)) - G^r(s \wedge \chi(s))) ds. \end{aligned}$$

It follows that

$$\begin{aligned} & \lambda \int_0^t G^r(\chi(s) \wedge s) ds \\ & + \int_{[0, H^r)} \left(\int_{(y, y+t]} \mathbf{1}_{\{x \leq \chi(x-y)\}} dG^r(x) \right) \bar{G}^r(y)^{-1} \eta_0(dy) \end{aligned}$$

$$\begin{aligned}
 &= \lambda \int_0^t G^r(\chi(s) \wedge s) ds + \lambda \int_0^t (G^r(\chi(s) \wedge (s+a)) - G^r(s \wedge \chi(s))) ds \\
 &= \lambda \int_0^t \mathbf{1}_{\{\chi(s) \geq s\}} (G^r(s) + G^r(\chi(s) \wedge (s+a)) - G^r(s)) ds \\
 &\quad + \lambda \int_0^t \mathbf{1}_{\{\chi(s) < s\}} G^r(\chi(s)) ds \\
 &= \lambda \int_0^t G^r(z_s) ds.
 \end{aligned}$$

Thus, (25) holds for the choice of z_t and hence the corollary follows directly from Lemma 3. ■

4 Proof of Theorem 2.1

In this section, we prove Theorem 1, the equivalence between the two fluid models tracking elapsed times described in Sections 2.2 and 2.1. We first derive a set of two-parameter fluid equations from a measure-valued fluid model (η, ν, X) in Definition 2 and show that it is a two-parameter fluid model; see Proposition 4.1. We then derive a set of measure-valued fluid equations from a two-parameter fluid model $(B(t, y), Q(t, y))$ in Definition 1 and show that it is a measure-valued fluid model; see Proposition 4.2. Thus we conclude that the existence and uniqueness of the two fluid models are equivalent.

Recall that $\chi(t)$ in (21) represents the waiting time of the fluid content at the head of the queue. Namely, the fluid content in the potential queue must be in queue waiting for service if the waiting time is less than $\chi(t)$, but must have abandoned otherwise. By the FCFS service discipline, the definition of the potential queue and the role of $w(t)$, we see that

$$\chi(t) = w(t). \tag{1}$$

We also observe that evidently, for each $y \geq 0$,

$$B(t, y) = \nu_t[0, y], \quad \tilde{Q}(t, y) = \eta_t[0, y], \quad \text{and} \quad Q(t, y) = \eta_t[0, y \wedge \chi(t)]. \tag{2}$$

We first start with the measure-valued fluid model (η, ν, X) in Definition 2, and show that the two-parameter processes $(B(t, y), Q(t, y))$ in (2) satisfy Definition 1. For this we need to assume that η_0 and ν_0 have densities $\tilde{q}(0, x)$ and $b(0, x)$, respectively, since they are required in the definition of the two-parameter fluid model.

Proposition 4.1 *Let (η, ν, X) be a measure-valued fluid model tracking elapsed times with the input data $(\lambda(\cdot), \eta_0, \nu_0, X(0))$ such that $(\eta_0, \nu_0, X(0)) \in \mathcal{S}_0$. Suppose that η_0 and ν_0 have densities $\tilde{q}(0, x)$ and $b(0, x)$, respectively. Then, $(B(t, y), Q(t, y))$ given by (2) is a two-parameter fluid model tracking elapsed times with the input data $(\lambda(\cdot), \tilde{q}(0, x), b(0, x))$ and $q(0, x) = \tilde{q}(0, x)$.*

Proof. Let (η, ν, X) be a measure-valued fluid model tracking elapsed times with the input data $(\lambda(\cdot), \eta_0, \nu_0, X(0))$ such that $(\eta_0, \nu_0, X(0)) \in \mathcal{S}_0$ and η_0 and ν_0 have densities $\tilde{q}(0, x)$ and $b(0, x)$, respectively. Since $b(0, x)$ and $\tilde{q}(0, x)$ denote the densities of ν_0 and η_0 , respectively, it follows that $b(0, x) = 0$ for each $x \geq H^s$ and $\tilde{q}(0, x) = 0$ for each $x \geq H^r$. For each $t \geq 0$ and $y \geq 0$, by letting $\psi_y(x) = \mathbf{1}(0 \leq x \leq y)$ in (15) and (16), respectively (Corollary 4.2 in [10] shows that (15) and (16) hold for any bounded Borel measurable function ψ), $B(t, y)$ and $\tilde{Q}(t, y)$ satisfy the following equations, respectively:

$$\begin{aligned} B(t, y) &= \int_0^{(y-t)^+ \wedge H^s} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} b(0, x) dx + \int_{(t-y)^+}^t \bar{G}^s(t-s) \kappa(s) ds \\ &= \int_0^{(y-t)^+ \wedge H^s} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} b(0, x) dx + \int_0^{y \wedge t} \bar{G}^s(s) \kappa(t-s) ds, \end{aligned} \tag{3}$$

$$\begin{aligned} \tilde{Q}(t, y) &= \int_0^{(y-t)^+ \wedge H^r} \frac{\bar{G}^r(x+t)}{\bar{G}^r(x)} \tilde{q}(0, x) dx + \int_{(t-y)^+}^t \bar{G}^r(t-s) \lambda(s) ds \\ &= \int_0^{(y-t)^+ \wedge H^r} \frac{\bar{G}^r(x+t)}{\bar{G}^r(x)} \tilde{q}(0, x) dx + \int_0^{y \wedge t} \bar{G}^r(s) \lambda(t-s) ds. \end{aligned} \tag{4}$$

Then from (3) and (4), $B(t, y)$ and $\tilde{Q}(t, y)$ have densities $b(t, y)$ and $\tilde{q}(t, y)$, respectively, with the representation:

$$b(t, y) = \begin{cases} \bar{G}^s(y) \kappa(t-y) & \text{if } y < t \wedge H^s, \\ \frac{\bar{G}^s(y)}{\bar{G}^s(y-t)} b(0, y-t) & \text{if } t < y < t + H^s, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

and

$$\tilde{q}(t, y) = \begin{cases} \bar{G}^r(y) \lambda(t-y) & \text{if } y < t \wedge H^r, \\ \frac{\bar{G}^r(y)}{\bar{G}^r(y-t)} \tilde{q}(0, y-t) & \text{if } t < y < t + H^r, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

From this, it is easy to check that the two fundamental evolution equations in (3) and (4) are satisfied. It is clear from the last equation in (2) that $Q(t, y)$ satisfies the following equation:

$$Q(t, y) = \int_0^{(y \wedge \chi(t) - t)^+ \wedge H^r} \frac{\bar{G}^r(x+t)}{\bar{G}^r(x)} \tilde{q}(0, x) dx + \int_0^{y \wedge \chi(t) \wedge t} \bar{G}^r(s) \lambda(t-s) ds, \tag{7}$$

Then, by comparing with (4), we have that

$$Q(t, y) = \begin{cases} \tilde{Q}(t, y) & \text{if } y < \chi(t), \\ Q(t) & \text{if } y \geq \chi(t). \end{cases} \tag{8}$$

Now, define $q(t, y)$ by

$$q(t, y) = \begin{cases} \tilde{q}(t, y) & \text{if } y < \chi(t), \\ 0 & \text{if } y > \chi(t), \\ \lambda(t) - \lambda(t) \wedge \int_{[0, H^s)} h^s(x) b(t, x) dx & \text{if } y = \chi(t), B(t) = 1, \\ 0 & \text{if } y = \chi(t), B(t) < 1. \end{cases} \tag{9}$$

Note that

$$\int_0^y q(t, x) dx = \int_0^y \tilde{q}(t, x) dx = \tilde{Q}(t, y) = Q(t, y), \text{ if } y < \chi(t),$$

and

$$\int_0^y q(t, x) dx = \int_0^{\chi(t)} \tilde{q}(t, x) dx = \tilde{Q}(t, \chi(t)) = Q(t) = Q(t, y), \text{ if } y \geq \chi(t).$$

Thus, $q(t, y)$ is a density function of $Q(t, y)$. Since $(\eta_0, \nu_0, X(0)) \in \mathcal{S}_0$, it is clear that $q(0, x) = \tilde{q}(0, x)$.

From (20) and (21), we obtain the following expression of $R(t)$ using the process $\tilde{Q}(t, y)$ in (4),

$$\begin{aligned} R(t) &= \int_0^t \left(\int_0^{\chi(s) \wedge H^r} h^r(x) \eta_s(dx) \right) ds \\ &= \int_0^t \left(\int_0^{(\chi(s)-s)^+ \wedge H^r} \frac{g^r(x+s)}{\bar{G}^r(x)} \tilde{q}(0, x) dx + \int_0^{s \wedge \chi(s)} g^r(x) \lambda(s-x) dx \right) ds \\ &= \int_0^t \left(\int_0^{\chi(s) \wedge H^r} h^r(x) \tilde{Q}(t, dx) \right) ds, \end{aligned} \tag{10}$$

and from (18), we obtain the following expression of $D(t)$ using the process $B(t, y)$ in (3),

$$\begin{aligned} D(t) &= \int_0^t \left(\int_{[0, H^s)} h^s(x) \nu_s(dx) \right) ds \\ &= \int_0^t \left(\int_{[0, H^s)} \frac{g^s(x+s)}{\bar{G}^s(x)} b(0, x) dx + \int_0^s g^s(x) \kappa(s-x) dx \right) ds \\ &= \int_0^t \left(\int_{[0, H^s)} h^s(x) B(t, dx) \right) ds. \end{aligned} \tag{11}$$

From (10), (11) and (8), we can see that $D(t)$ and $R(t)$ have densities $\sigma(t)$ and $\alpha(t)$, respectively, and they satisfy (9), that is,

$$\sigma(t) = \int_{[0, H^s)} b(t, x) h^s(x) dx, \quad \alpha(t) = \int_{[0, H^r)} q(t, x) h^r(x) dx, \quad t \geq 0. \tag{12}$$

To complete the proof, it is enough to show that $b(t, 0)$, $\tilde{q}(t, 0)$ and $q(t, 0)$ from (5), (6) and (9) satisfy (6), (7) and (8), respectively. Note that $b(t, 0) = \kappa(t)$ by (5). Combining this with (29) and (12), $b(t, 0)$ satisfies (6). By (6), $\tilde{q}(t, 0) = \lambda(t)$ and then satisfies (7). On the other hand, from (9) and (12),

$$q(t, 0) = \begin{cases} \tilde{q}(t, 0) = \lambda(t) & \text{if } 0 < \chi(t), \\ \lambda(t) - \lambda(t) \wedge \sigma(t) & \text{if } 0 = \chi(t), B(t) = 1, \\ 0 & \text{if } 0 = \chi(t), B(t) < 1. \end{cases} \tag{13}$$

This implies that $q(t, 0)$ satisfies (8). Finally, the rate balance equation (10) follows from the balance equation (19), by noting that $Q(t) = \int_0^\infty q(t, x) dx$, $K(t) = \int_0^t b(s, 0) ds$ and $R(t) = \int_0^t \alpha(s) ds$. ■

We next show that a set of measure-valued equations (ν, η, X) derived from a two-parameter fluid model in Definition 1 satisfies Definition 2.

Proposition 4.2 *Let $(B(t, y), Q(t, y))$ be a two-parameter fluid model tracking elapsed times with the input data $(\lambda(\cdot), \tilde{q}(0, x), b(0, x))$ and $q(0, x) = \tilde{q}(0, x)$. For each $t \geq 0$, let $\eta_t[0, y] \doteq \tilde{Q}(t, y)$ and $\nu_t[0, y] \doteq B(t, y)$ for each $y \geq 0$ and define $X(t) \doteq B(t, \infty) + Q(t, \infty)$. Then, (η, ν, X) is a measure-valued fluid model tracking elapsed times with the input data $(\lambda(\cdot), \eta_0, \nu_0, X(0))$ such that $(\eta_0, \nu_0, X(0)) \in \mathcal{S}_0$.*

Proof. Fix $(B(t, y), Q(t, y))$ and the triple of functions (η, ν, X) defined from it. It is clear from the two fundamental evolution equations (3) and (4) that for each $t \geq 0$, $\tilde{q}(t, x)$ as a function in x has support in $[0, H^r)$ and $b(t, x)$ as a function in x has support in $[0, H^s)$. It then follows that η_t has support in $[0, H^r)$ and ν_t has support in $[0, H^s)$ for each $t \geq 0$. Also it is clear that $(\eta_0, \nu_0, X(0)) \in \mathcal{S}_0$.

We first show that ν satisfies (16). For every $\psi \in \mathcal{C}_b(\mathbb{R}_+)$ and $t \geq 0$,

$$\int_0^\infty \psi(x) \nu_t(dx) = \int_0^\infty \psi(x) b(t, x) dx = \int_0^t \psi(x) b(t, x) dx + \int_t^\infty \psi(x) b(t, x) dx. \tag{14}$$

For the first term on the right-hand side of (14), we can use the first fundamental evolution equation (3) to yield that

$$\int_0^t \psi(x) b(t, x) dx = \int_0^t \psi(x) b(t-x, 0) \frac{\bar{G}^s(x)}{\bar{G}^s(0)} dx = \int_0^t \psi(x) \bar{G}^s(x) b(t-x, 0) dx. \tag{15}$$

For the second term on the right-hand side of (14), another application of the first fundamental evolution equation (3) yields that

$$\begin{aligned} \int_t^\infty \psi(x) b(t, x) dx &= \int_{t \wedge H^s}^{H^s} \psi(x) b(t, x) dx \\ &= \int_{t \wedge H^s}^{H^s} \psi(x) b(0, x-t) \frac{\bar{G}^s(x)}{\bar{G}^s(x-t)} dx \\ &= \int_0^{t \vee H^s - t} \psi(x+t) \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} b(0, x) dx \end{aligned} \tag{16}$$

$$= \int_0^{H^s} \psi(x+t) \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} b(0,x) dx.$$

The last equality in (16) follows from the fact that $\bar{G}^s(x+t) = 0$ if $x \in (t \vee H^s - t, H^s)$. For each $t \geq 0$, let

$$K(t) \doteq \int_0^t b(s,0) ds. \tag{17}$$

Combining the above four displays, we obtain that v satisfies (16). An analogous argument using the second fundamental evolution equation (4) shows that

$$\begin{aligned} \int_0^\infty \psi(x) \eta_t(dx) &= \int_0^\infty \psi(x) \tilde{q}(t,x) dx \\ &= \int_0^t \psi(x) \tilde{q}(t,x) dx + \int_t^\infty \psi(x) \tilde{q}(t,x) dx \\ &= \int_0^t \psi(x) \bar{G}^r(x) \tilde{q}(t-x,0) dx + \int_0^{H^r} \psi(x+t) \frac{\bar{G}^r(x+t)}{\bar{G}^r(x)} \tilde{q}(0,x) dx. \end{aligned} \tag{18}$$

By (7), $\tilde{q}(t-x,0) = \lambda(t-x)$. Thus, η satisfies (15).

Next, for each $t \geq 0$, define $B(t) \doteq B(t, \infty)$, $Q(t) \doteq Q(t, \infty)$, $D(t) \doteq \int_0^t \sigma(x) dx$, $R(t) \doteq \int_0^t \alpha(x) dx$. From (9), D satisfies (18). From (9) again, (5) and (1), R satisfies (20). Since v satisfies (16), by choosing $\psi = \mathbf{1}$ in (16), we have

$$B(t) = \int_0^{H^s} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} v_0(dx) + \int_0^t \bar{G}^s(t-s) dK(s)$$

and by choosing $\psi = h^s$ in (16), we have

$$\begin{aligned} D(t) &= \int_0^t \int_0^{H^s} b(s,x) h^s(x) dx ds \\ &= \int_0^t \left(\int_0^{H^s} \frac{g^s(x+s)}{\bar{G}^s(x)} v_0(dx) + \int_0^s g^s(s-x) dK(x) \right) ds \\ &= \int_0^{H^s} \frac{\bar{G}^s(x) - \bar{G}^s(x+t)}{\bar{G}^s(x)} v_0(dx) + \int_0^t \int_0^s g^s(s-x) dK(x) ds \\ &= B(0) - \int_0^{H^s} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} v_0(dx) + \int_0^t G^s(t-s) dK(s) \\ &= B(0) - B(t) + K(t). \end{aligned}$$

This shows that (17) is satisfied. The relationship (10) directly implies that (19).

Now, (21) follows directly from (1). Finally the non-idling conditions in (22)–(24) directly follows from those in (12)–(14). This completes the proof of the proposition. ■

Acknowledgement

We thank Ward Whitt and Kavita Ramanan for many helpful discussions and suggestions on the paper.

References

- [1] Bogachev, V. I. (2007) *Measure Theory*, Vol I, Springer, Berlin.
- [2] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005) Statistical analysis of a telephone call center: a queueing-science perspective. *J. Amer. Stat. Ass.* Vol. 100, No. 469, 36–50.
- [3] Gans, N, G. Koole and A. Mandelbaum. (2003) Telephone call centers: tutorial, review and research prospects. *Manufacturing & Service Operations Management*. **5**, 79–141.
- [4] Garnett, O., A. Mandelbaum and M.I. Reiman. (2002) Designing a call center with impatient customers. *Manufacturing & Service Operations Management*. **4**, 208–227.
- [5] Glynn, P.W. and W. Whitt. (1991) A new view of the heavy-traffic limit theorem for infinite-server queues. *Adv. Appl. Prob.* **23**, 188–209.
- [6] Green, L.V. (2006) Queueing analysis in healthcare. *Patient Flow: Reducing Delay in Healthcare Delivery*, edited by R. Hall. Springer.
- [7] Kang, W. (2014) Existence and Uniqueness of a Fluid Model for Many-Server Queues with Abandonment. *Operations Research Letters*. **42** (6-7), 478–483.
- [8] Kang, W. and G. Pang. (2014) Fluid limit of a many-server queueing network with abandonment. *Working paper*.
- [9] Kang, W. and G. Pang. (2015) An algorithm to compute two-parameter fluid models for $G_t/GI/N + GI$ queues. *Working paper*.
- [10] Kang, W. and K. Ramanan. (2010) Fluid limits of many-server queues with reneging. *Annals of Applied Probability*. **20**, 2204–2260.
- [11] Kaspi, H. and K. Ramanan. (2011) Law of large numbers limits for many-server queues. *Annals of Applied Probability*. **21**, 33–114.
- [12] Liu, Y. and W. Whitt. (2012) The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems*. Vol. 71, No. 4, 405–444.
- [13] Liu, Y. and W. Whitt. (2012) A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Operations Research Letters*. Vol.40, 307–312.
- [14] Liu, Y. and W. Whitt. (2011) A network of time-varying many-server fluid queues with customer abandonment. *Operations Research*. Vol. 59, 835–846.
- [15] Liu, Y. and W. Whitt. (2011) Large-time asymptotics for the $G_t/M_t/s_t + GI$ many-server fluid queue with abandonment. *Queueing Systems*. **67**, 145–182.
- [16] Long, Z. and J. Zhang. (2014) Convergence to equilibrium states for fluid models of many-server queues with abandonment. *Operations Research Letters*. **42** (6-7), 388–393.

- [17] Mandelbaum, A. and P. Momcilovic. (2017) Personalized queues: the customer view via a fluid model of serving least-patient first. *Queueing Systems*. Vol. 87, No. 1–2, 23–53.
- [18] Pang, G., R. Talreja and W. Whitt. (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*., Vol. 4, 193–267.
- [19] Pang, G. and W. Whitt. (2010) Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems*. 65, 325–364.
- [20] Revuz, D. and M. Yor. (1991) *Continuous Martingales and Brownian Motion*, Springer-Verlag, Berlin, Heidelberg, New York.
- [21] Whitt, W. (2006) Fluid models for multiserver queues with abandonment. *Operations Research*. Vol. 54, No. 1, 37–54.
- [22] Zhang, J. (2013) Fluid models of many-server queues with abandonment. *Queueing Systems*. Vol. 73, No. 2, 147–193.
- [23] Zuñaiga, A. W. (2014) Fluid limits of many-server queues with abandonments, general service and continuous patience time distributions. *Stochastic Processes and their Applications*. Vol. 124, No. 3, 1436–1468.



Stochastic HJB Equations and Regular Singular Points*

Arthur J Krener

Abstract We consider finite and infinite horizon, stochastic, smooth optimal control problems in continuous time where the coefficients of the white Gaussian noise terms in the dynamics vanish at the origin. For infinite horizon problems we show how the Taylor polynomials of the optimal cost and the optimal feedback can be computed degree by degree. The degree two part of the optimal cost and the degree one part of the optimal feedback are found by solving new stochastic algebraic Riccati equations (SARE). If SARE is solvable the higher degree terms can be found by solving linear algebraic equations. This is a generalization of the work of Al'brekht who showed how the Taylor polynomials of the optimal cost and the optimal feedback for deterministic problems can be computed degree by degree. For finite horizon problems we show how the Taylor polynomials of the optimal cost and the optimal feedback can also be computed degree by degree. The degree two part of the optimal cost and the degree one part of the optimal feedback are found by solving stochastic differential Riccati equations (SDRE). The higher degree terms can be found by solving linear differential equations.

1 Introduction

Euler solved some second order, linear, variable coefficient ODEs by power series around a regular singular point [2]. The equations that he considered are of the form

$$P(x) \frac{d^2y}{dx^2} + Q(x) \frac{dy}{dx} + R(x)y = 0 \quad (1)$$

A. J. Krener

Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA 93943, e-mail: ajkrenner@nps.edu

* Research supported in part by AFOSR under FA9550-17-1-0219.

and this ODE has a regular singular point at $x = 0$ if $P(x) = O(x)^2$ and $Q(x) = O(x)$. If we divide by $P(x)$ then we obtain

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = 0$$

where $p(x) = \frac{Q(x)}{P(x)}$ and $q(x) = \frac{R(x)}{P(x)}$.

Because $P(x) = O(x)^2$ the coefficients $p(x)$ and $q(x)$ can be singular at $x = 0$. Suppose that we have the series expansion

$$xp(x) = \sum_{n=0}^{\infty} p_n x^n, \quad x^2q(x) = \sum_{n=0}^{\infty} q_n x^n$$

Euler assumed that for some ρ the solution had a power series expansion of the form

$$\begin{aligned} y(x) &= \sum_{n=0}^{\infty} a_n x^{\rho+n} \\ y'(x) &= \sum_{n=0}^{\infty} (\rho+n)a_n x^{\rho+n-1} \\ y''(x) &= \sum_{n=0}^{\infty} (\rho+n)(\rho+n-1)a_n x^{\rho+n-2} \end{aligned}$$

He plugged these series into the ODE (1) and collected the coefficient of x^ρ to get the so-called indicial equation

$$F(\rho) = p_0\rho(\rho-1) + q_0\rho + r_0 = 0$$

This quadratic has two possibly complex roots ρ_1, ρ_2 and corresponding to each root there is a series solution of ODE. For each root setting the coefficient of $x^{\rho+n}$ equal to zero yields the recursion relation.

$$F(\rho+n)a_n + \sum_{k=0}^{n-1} a_k ((\rho+k)p_{n-k} + q_{n-k}) = 0$$

so if $F(\rho+n) \neq 0$ we can solve for a_n as a function of a_k , $0 \leq k < n$. Assuming $F(\rho_i + n)$ is never zero this yields a series solution to the ODE for each ρ_i that depends on its first coefficient a_0 . The sum of these two solutions each depending on a free constant yields the general solution to the ODE.

More recently Al'brekht [1] considered an infinite horizon, deterministic optimal control problem

$$\min_{u(\cdot)} \int_0^{\infty} l(x, u) dt$$

subject to

$$\begin{aligned} \dot{x} &= f(x, u) \\ x(0) &= x^0 \end{aligned}$$

It is well-known if the optimal cost $\pi(x^0)$ and optimal feedback $u(t) = \kappa(x(t))$ exist and are smooth then they satisfy the Hamilton-Jacobi-Bellman Equations (HJB)

$$\begin{aligned} 0 &= \min_u \left\{ \frac{\partial \pi}{\partial x}(x) f(x, u) + l(x, u) \right\} \\ \kappa(x) &= \operatorname{argmin}_u \left\{ \frac{\partial \pi}{\partial x}(x) f(x, u) + l(x, u) \right\} \end{aligned}$$

If the quantity to be minimized is smooth with respect to u then the HJB equations imply that

$$\begin{aligned} 0 &= \frac{\partial \pi}{\partial x}(x) f(x, \kappa(x)) + l(x, \kappa(x)) \\ 0 &= \frac{\partial \pi}{\partial x}(x) \frac{\partial f}{\partial u}(x, \kappa(x)) + \frac{\partial l}{\partial u}(x, \kappa(x)) \end{aligned}$$

We call these the simplified HJB equations. Of course the simplified HJB equations do not imply the HJB equations if the quantity to be minimized is not strictly convex in u .

Al’brekht assumed that $l(x, u)$ and $f(x, u)$ are smooth and have Taylor polynomial expansions

$$\begin{aligned} l(x, u) &= \frac{1}{2} (x' Q x + 2x' S u + u' R u) + l^{[3]}(x, u) + l^{[4]}(x, u) \\ &\quad + \dots + l^{[d+1]}(x, u) + O(x, u)^{d+2} \end{aligned}$$

$$f(x, u) = Fx + Gu + f^{[2]}(x, u) + f^{[3]}(x, u) + \dots + f^{[d]}(x, u) + O(x, u)^{d+1}$$

and the unknowns $\pi(x)$ and $\kappa(x)$ have similar Taylor polynomial expansions

$$\pi(x, u) = \frac{1}{2} x' P x + \pi^{[3]}(x) + \pi^{[3]}(x) + \dots + \pi^{[d+1]}(x) + O(x)^{d+2}$$

$$\kappa(x) = Kx + \kappa^{[2]}(x) + \kappa^{[3]}(x) + \dots + \kappa^{[d]}(x) + O(x)^{d+1}$$

He plugged these into the simplified HJB equations and solved degree by degree. At the leading degrees, two in the first simplified HJB equation and one in the second simplified HJB equation, he obtained the familiar LQR equations

$$0 = F'P + PF + Q - (PG + S)R^{-1}(G'P + S')$$

$$0 = K'R + (PG + S)$$

Then he derived linear recursion relations for the higher degree terms, $\pi^{[3]}(x), \kappa^{[2]}(x), \pi^{[4]}(x), \kappa^{[3]}(x), \dots$

The reason why Al'brekht's method succeeds is that the first simplified HJB equations has a regular singular point at $x = 0$. One indication of this is the coefficient $f(x, \kappa(x))$ of $\frac{\partial \pi}{\partial x}(x)$ vanishes at $x = 0$. But this is not the only place where $\frac{\partial \pi}{\partial x}(x)$ appears. The second simplified HJB equation allows us to express $\kappa(x)$ in terms of $\frac{\partial \pi}{\partial x}(x)$. When we plug this into the first simplified HJB equation we get a nonlinear first order PDE where $\frac{\partial \pi}{\partial x}(x)$ appears quadratically in the Lagrangian. To have a regular singular point at $x = 0$ we must have $\frac{\partial \pi}{\partial x} = O(x)$ and $\pi(x)$ must be $O(x)^2$.

This paper extends Al'brekht's Method to stochastic, infinite horizon optimal control problems in continuous time. We also extend Al'brekht's Method to stochastic, finite horizon optimal control problems in continuous time. To do so we must make the assumption that the coefficients of noises are of order $O(x, u)$. Stochastic HJB equations are nonlinear, second order PDEs. This assumption ensures that coefficients of the second partial derivatives of $\pi(x)$ are of order $O(x, u)^2$ and so such stochastic HJB PDEs have regular singular points at the origin.

The rest of this paper is organized as follows. In the next section we study the simplest example of the problems that we are considering, linear quadratic regulators with bilinear noise. These lead to new type of algebraic Riccati equation which may or may not have a solution. In Section 3 we give an example of this. Section 4 presents the extension of Al'brekht's method to stochastic, infinite horizon optimal control problems that are not linear-quadratic but where the coefficients of the noise are $O(x, u)$. Section 5 contains an example of such a problem where the Taylor polynomials of the optimal cost and the optimal feedback are computed to degrees six and five respectively by the extension of Al'brekht's method. In Section 6 we turn to stochastic optimal control problems over finite horizons where the coefficients of the noise are again $O(x, u)$. At the lowest degrees these problems lead to a type of stochastic differential Riccati that is well-know. What is new is that the higher degree terms of the optimal cost and the optimal feedback can be computed degree by degree by solving linear ODEs. We conclude in Section 7 and we close with acknowledgements.

2 Linear Quadratic Regulator with Bilinear Noise

The simplest version of the problems of interest is an infinite horizon, stochastic Linear Quadratic Regulator with Bilinear Noise (LQGB),

$$\min_{u(\cdot)} \frac{1}{2} \mathbb{E} \int_0^{\infty} (x' Q x + 2x' S u + u' R u) dt$$

subject to

$$dx = (Fx + Gu) dt + \sum_{k=1}^r (C_k x + D_k u) dw_k$$

$$x(0) = x^0$$

In a previous version of this paper [4] we studied the case with $D_k = 0$.

The state x is n dimensional, the control u is m dimensional and $w(t) = (w_1(t), \dots, w_r(t))'$ is standard r dimensional Brownian motion. The matrices are sized accordingly, in particular, C_k is an $n \times n$ matrix and D_k is an $n \times m$ matrix for each $k = 1, \dots, r$.

To the best of our knowledge such problems have not been considered before. The finite horizon version of this problem can be found in Chapter 6 of the excellent treatise by Yong and Zhou [6]. We will also treat finite horizon problems in Section 6 but not in the same generality as Yong and Zhou. Throughout this note we will require that the coefficient of the noise is $O(x, u)$. Yong and Zhou allow the coefficient to be $O(1)$ in their linear-quadratic problems. The reason why we require $O(x, u)$ is that then the associated stochastic Hamilton-Jacobi-Bellman equations for nonlinear extensions of LQGB have regular singular points at the origin. Hence they are amenable to solution by power series techniques. If the noise is $O(1)$ these power series techniques have closure problems, the equations for lower degree terms depend on higher degree terms. If the coefficients of the noise is $O(x, u)$ then the equations can be solved degree by degree.

A first order partial differential equation with independent variable x has a regular singular point at $x = 0$ if the coefficients the first order partial derivatives are $O(x)$. A second order partial differential equation has a regular singular point at $x = 0$ if the coefficients the first order partial derivatives are $O(x)$ and the coefficients the second order partial derivatives are $O(x)^2$. For more on regular singular points we refer the reader to [2].

If we can find a smooth scalar valued function $\pi(x)$ and a smooth m vector valued $\kappa(x)$ satisfying the stochastic Hamilton-Jacobi-Bellman equations (SHJB)

$$0 = \min_u \left\{ \frac{\partial \pi}{\partial x}(x)(Fx + Gu) + \frac{1}{2} (x' Qx + 2x' Su + u' Ru) \right. \\ \left. + \frac{1}{2} \sum_{k=1}^r (x' C'_k + u' D'_k) \frac{\partial^2 \pi}{\partial x^2}(x)(C_k x + D_k u) \right\} \tag{2}$$

$$\kappa(x) = \operatorname{argmin}_u \left\{ \frac{\partial \pi}{\partial x}(x)(Fx + Gu) + \frac{1}{2} (x' Qx + 2x' Su + u' Ru) \right. \\ \left. + \frac{1}{2} \sum_{k=1}^r (x' C'_k + u' D'_k) \frac{\partial^2 \pi}{\partial x^2}(x)(C_k x + D_k u) \right\} \tag{3}$$

then by a standard verification argument [3] one can show that $\pi(x^0)$ is the optimal cost of starting at x^0 and $u(0) = \kappa(x^0)$ is the optimal control at x^0 .

We make the standard assumptions of deterministic LQR,

1. The matrix

$$\begin{bmatrix} Q & S \\ S' & R \end{bmatrix}$$

is nonnegative definite.

2. The matrix R is positive definite.
3. The pair F, G is stabilizable.
4. The pair $Q^{1/2}, F$ is detectable.

Because of the linear dynamics and quadratic cost, we expect that $\pi(x)$ is a quadratic function of x and $\kappa(x)$ is a linear function of x ,

$$\begin{aligned} \pi(x) &= \frac{1}{2}x'Px \\ \kappa(x) &= Kx \end{aligned}$$

Then the stochastic Hamilton-Jacobi-Bellman equations (2, 3) simplify to

$$0 = F'P + PF + Q + \sum_k C_k'PC_k - K' \left(R + \sum_k D_k'PD_k \right) K \quad (4)$$

$$K = - \left(R + \sum_k D_k'PD_k \right)^{-1} \left(G'P + S + \sum_k D_k'PC_k \right) \quad (5)$$

We call these equations (4, 5) the Stochastic Algebraic Riccati Equations (SARE). They reduce to the deterministic Algebraic Riccati Equations (ARE) if $C_k = 0$ and $D_k = 0$ for all k .

Here is an iterative method for solving SARE. Let $P_{(0)}$ be the solution of the deterministic ARE

$$0 = P_{(0)}F + F'P_{(0)} + Q - (P_{(0)}G + S)R^{-1}(G'P_{(0)} + S')$$

and $K_{(0)}$ be given by

$$K_{(0)} = -R^{-1}(G'P_{(0)} + S')$$

Given $P_{(\tau-1)}$ define

$$Q_{(\tau)} = Q + \sum_{k=1}^r C_k'P_{(\tau-1)}C_k$$

$$R_{(\tau)} = R + \sum_{k=1}^r D_k'P_{(\tau-1)}D_k$$

$$S_{(\tau)} = S + \sum_{k=1}^r C_k'P_{(\tau-1)}D_k$$

Let $P_{(\tau)}$ be the solution of

$$0 = P_{(\tau)}F + F'P_{(\tau)} + Q_{(\tau)} - (P_{(\tau)}G + S_{(\tau)})R_{(\tau)}^{-1}(G'P_{(\tau)} + S'_{(\tau)})$$

and

$$K_{(\tau)} = -R_{(\tau)}^{-1} \left(G'P_{(\tau)} + S'_{(\tau)} \right)$$

If the iteration on $P_{(\tau)}$ nearly converges, that is, for some τ , $P_{(\tau)} \approx P_{(\tau-1)}$ then $P_{(\tau)}$ and $K_{(\tau)}$ are approximate solutions to SARE

The solution P of the deterministic ARE is the kernel of the optimal cost of a deterministic LQR and since

$$\begin{bmatrix} Q & S \\ S' & R \end{bmatrix} \leq \begin{bmatrix} Q_{(\tau-1)} & S_{(\tau-1)} \\ S'_{(\tau-1)} & R_{(\tau-1)} \end{bmatrix} \leq \begin{bmatrix} Q_{(\tau)} & S_{(\tau)} \\ S'_{(\tau)} & R_{(\tau)} \end{bmatrix}$$

it follows that $P_{(0)} \leq P_{(\tau-1)} \leq P_{(\tau)}$, the iteration is monotonically increasing. We have found computationally, using MATLAB's `are.m`, that if matrices C_k and D_k are not too big then the iteration converges. But if the C_k and D_k are about the same size as F and G or larger then the iteration can diverge. Further study of this issue is needed. The iteration does converge in the following simple example.

Another issue which deserves further study is whether the first and second standard assumptions of deterministic LQR can be weakened. It is known [6] that finite horizon stochastic LQR problems with indefinite or even negative definite R can have finite solutions when the control enters the coefficient of the noise. One might think that if R is negative definite then by using larger and larger control actions one could drive the quantity to be minimized to negative infinity. But the volatility in x cause by large control actions can cause the quadratic terms in x to get very large thereby canceling the negative effect of the quadratic terms in u . The reason why can be seen in the above iteration. For some $\tau^* > 0$ it may happen that

$$\begin{bmatrix} Q_{(\tau^*)} & S_{(\tau^*)} \\ S'_{(\tau^*)} & R_{(\tau^*)} \end{bmatrix} \geq 0 \\ R_{(\tau^*)} > 0$$

then this will happen for all $\tau > \tau^*$ even though this might not be true when $\tau = 0$. For such problems in the above iteration one should use a solver like MATLAB's `are.m` that can handle them.

3 LQGB Example

Here is a simple example with $n = 2, m = 1, r = 2$.

$$\min_u \frac{1}{2} \int_0^\infty \|x\|^2 + u^2 dt$$

subject to

$$\begin{aligned} dx_1 &= x_2 dt + 0.1x_1 dw_1 \\ dx_2 &= u dt + 0.1(x_2 + u) dw_2 \end{aligned}$$

In other words

$$\begin{aligned} Q &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & S &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}, & R &= 1 \\ F &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, & G &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ C_1 &= \begin{bmatrix} 0.1 & 0 \\ 0 & 0 \end{bmatrix}, & C_2 &= \begin{bmatrix} 0 & 0 \\ 0 & 0.1 \end{bmatrix} \\ D_1 &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, & D_2 &= \begin{bmatrix} 0 \\ 0.1 \end{bmatrix} \end{aligned}$$

The solution of the noiseless ARE is

$$\begin{aligned} P &= \begin{bmatrix} 1.7321 & 1.000 \\ 1.000 & 1.7321 \end{bmatrix} \\ K &= - \begin{bmatrix} 1.0000 & 1.7321 \end{bmatrix} \end{aligned}$$

The eigenvalues of the noiseless closed loop matrix $F + GK$ are $-0.8660 \pm 0.5000i$.

Using `are.m` the above iteration converges to the solution of the noisy SARE in eight iterations, the solution is

$$\begin{aligned} P &= \begin{bmatrix} 1.7625 & 1.0176 \\ 1.0176 & 1.7524 \end{bmatrix} \\ K &= - \begin{bmatrix} 1.0176 & 1.7524 \end{bmatrix} \end{aligned}$$

The eigenvalues of the noisy closed loop matrix $F + GK$ are $-0.8762 \pm 0.4999i$.

As expected the noisy system is more difficult to control than the noiseless system. It should be noted that the above iteration diverged to infinity when the noise coefficients were increased from 0.1 to 1.

4 Nonlinear Infinite Horizon HJB

Suppose the problem is not linear-quadratic, the dynamics is given by an Ito equation

$$dx = f(x, u) dt + \sum_{k=1}^r \gamma_k(x, u) dw_k$$

and the criterion to be minimized is

$$\min_{u(\cdot)} E \int_0^\infty l(x, u) dt$$

We assume that $f(x, u), \gamma_k(x, u), l(x, u)$ are smooth functions that have Taylor polynomial expansions around $x = 0, u = 0$,

$$\begin{aligned} f(x, u) &= Fx + Gu + f^{[2]}(x, u) + \dots + f^{[d]}(x, u) + O(x, u)^{d+1} \\ \gamma_k(x, u) &= C_k x + D_k u + \gamma_k^{[2]}(x, u) + \dots + \gamma_k^{[d]}(x, u) + O(x)^{d+1} \\ l(x, u) &= \frac{1}{2} (x' Q x + 2x' S u + u' R u) + l^{[3]}(x, u) + \dots + l^{[d+1]}(x, u) + O(x, u)^{d+2} \end{aligned}$$

where $^{[d]}$ indicates the homogeneous polynomial terms of degree d .

The stochastic Hamilton-Jacobi-Bellman equations are

$$\begin{aligned} 0 = \min_u \left\{ \frac{\partial \pi}{\partial x}(x) f(x, u) + l(x, u) \right. \\ \left. + \frac{1}{2} \sum_{k=1}^r \gamma'_k(x, u) \frac{\partial^2 \pi}{\partial x^2}(x) \gamma_k(x, u) \right\} \end{aligned} \tag{6}$$

$$\begin{aligned} \kappa(x) = \operatorname{argmin}_u \left\{ \frac{\partial \pi}{\partial x}(x) f(x, u) + l(x, u) \right. \\ \left. + \frac{1}{2} \sum_{k=1}^r \gamma'_k(x, u) \frac{\partial^2 \pi}{\partial x^2}(x) \gamma_k(x, u) \right\} \end{aligned} \tag{7}$$

If the control enters the dynamics affinely,

$$\begin{aligned} f(x, u) &= f^0(x) + f^u(x)u \\ \gamma_k(x, u) &= \gamma_k^0(x) + \gamma_k^u(x)u \end{aligned}$$

and $l(x, u)$ is always strictly convex in u for every x then the quantity to be minimized in (6) is strictly convex in u .

Whether (6) is strictly convex or not because it is smooth, the HJB equations (6, 7) imply

$$\begin{aligned} 0 = \frac{\partial \pi}{\partial x}(x) f(x, \kappa(x)) + l(x, \kappa(x)) \\ + \frac{1}{2} \sum_{k=1}^r \gamma'_k(x, \kappa(x)) \frac{\partial^2 \pi}{\partial x^2}(x) \gamma_k(x, \kappa(x)) \end{aligned} \tag{8}$$

$$0 = \frac{\partial \pi}{\partial x}(x) \frac{\partial f}{\partial u}(x, \kappa(x)) + \frac{\partial l}{\partial u}(x, \kappa(x)) \tag{9}$$

$$+ \sum_{k=1}^r \gamma'_k(x, \kappa(x)) \frac{\partial^2 \pi}{\partial x^2}(x) \frac{\partial \gamma_k}{\partial u}(x, \kappa(x))$$

These are the simplified stochastic HJB equations. Of course when (6) is not strictly convex in u these equations do not imply the stochastic HJB equations because of the possibility of multiple local minima.

Because $f(x, u) = O(x, u)$ and $\gamma_k(x, u) = O(x, u)$, (8) has a regular singular point at $x = 0, u = 0$ and so is amenable to power series solution techniques. If $\gamma_k(x, u) = O(1)$ then there is persistent noise that must be overcome by persistent control action. Presumably then the infinite horizon optimal cost is infinite.

Following Al'brekht [1] we assume that the optimal cost and the optimal feedback have Taylor polynomial expansions

$$\begin{aligned} \pi(x) &= \frac{1}{2} x' P x + \pi^{[3]}(x) + \dots + \pi^{[d+1]}(x) + O(x)^{d+2} \\ \kappa(x) &= Kx + \kappa^{[2]}(x) + \dots + \kappa^{[d]}(x) + O(x)^{d+1} \end{aligned}$$

We plug all these expansions into the simplified SHJB equations (8, 9). At lowest degrees, degree two in (8) and degree one in (9) we get the familiar SARE (4, 5).

If (4, 5 are solvable then we may proceed to the next degrees, degree three in (8) and degree two in (9).

$$\begin{aligned} 0 &= \frac{\partial \pi^{[3]}}{\partial x}(x)(F + GK)x + x' P f^{[2]}(x, Kx) + l^{[3]}(x, Kx) \\ &+ \frac{1}{2} \sum_k x'(C'_k + K' D'_k) \frac{\partial^2 \pi^{[3]}}{\partial x^2}(x)(C_k + D_k K)x \\ &+ \sum_k x'(C'_k + K' D'_k) P \gamma_k^{[2]}(x, Kx) \end{aligned} \tag{10}$$

$$\begin{aligned} 0 &= \frac{\partial \pi^{[3]}}{\partial x}(x)G + x' P \frac{\partial f^{[2]}}{\partial u}(x, Kx) + \frac{\partial l^{[3]}}{\partial u}(x, Kx) \\ &+ \sum_k x'(C_k + D_k K)' \left(P \frac{\partial \gamma_k^{[2]}}{\partial u}(x, Kx) + \frac{\partial^2 \pi^{[3]}}{\partial x^2}(x) D_k \right) \\ &+ \sum_k \gamma_k^{[2]}(x, Kx) P D_k + (\kappa^{[2]}(x))' \left(R + \sum_k D'_k P D_k \right) \end{aligned} \tag{11}$$

Notice the first equation (10) is a square linear equation for the unknown $\pi^{[3]}(x)$, the other unknown $\kappa^{[2]}(x)$ does not appear in it. If we can solve the first equation (10) for $\pi^{[3]}(x)$ and if $R + \sum_k D_k P D_k$ is invertible. then we can solve the second equation (11) for $\kappa^{[2]}(x)$.

In the deterministic case the eigenvalues of the linear operator

$$\pi^{[3]}(x) \mapsto \frac{\partial \pi^{[3]}}{\partial x}(x)(F + GK)x \tag{12}$$

are the sums of three eigenvalues of $F + GK$. Under the standard LQR assumptions all the eigenvalues of $F + GK$ are in the open left half plane so any sum of three eigenvalues of $F + GK$ is different from zero and the operator (12) is invertible.

In the stochastic case the relevant linear operator is a sum of two operators

$$\begin{aligned} \pi^{[3]}(x) \mapsto & \frac{\partial \pi^{[3]}}{\partial x}(x)(F + GK)x \\ & + \frac{1}{2} \sum_k x'(C'_k + K'D'_k) \frac{\partial^2 \pi^{[3]}}{\partial x^2}(x)(C_k + D_k K)x \end{aligned} \tag{13}$$

Consider a simple version of the second operator, for some C ,

$$\pi^{[3]}(x) \mapsto \frac{1}{2} x' C' \frac{\partial^2 \pi^{[3]}}{\partial x^2}(x) C x \tag{14}$$

Suppose C has a complete set of left eigenpairs, $\lambda_i \in \mathcal{C}$, $w^i \in \mathcal{C}^{1 \times n}$ for $i = 1, \dots, n$,

$$w^i C = \lambda_i w^i$$

Then the eigenvalues of (14) are of the form $\lambda_{i_1} \lambda_{i_2} + \lambda_{i_2} \lambda_{i_3} + \lambda_{i_3} \lambda_{i_1}$ and the corresponding eigenvectors are $(w^{i_1} x)(w^{i_2} x)(w^{i_3} x)$ for $1 \leq i_1 \leq i_2 \leq i_3$. But this analysis does not completely clarify whether the operator (13) is invertible. Here is one case where we know it is invertible.

Consider the space of cubic polynomials $\pi(x)$. We can norm this space using the standard L_2 norm on the vector of coefficients of $\pi(x)$ which we denote by $\|\pi(x)\|$. Then there is an induced norm on operators like (12), (13) and

$$\pi^{[3]}(x) \mapsto \frac{1}{2} \sum_k x'(C'_k + K'D'_k) \frac{\partial^2 \pi^{[3]}}{\partial x^2}(x)(C_k + D_k K)x$$

Since the operator (12) is invertible its inverse has an operator norm $\rho < \infty$. If all the eigenvalues of $F + GK$ have real parts less than $-\tau$ then $\frac{1}{\rho} \geq 3\tau$. Let σ be the supremum operator norms of $C_k + D_k K$ for $k = 1, \dots, r$. Then from the discussion above we know that the operator norm of (15) is bounded above by $\frac{3r\sigma^2}{2}$

Lemma 1. *If $\tau > \frac{r\sigma^2}{2}$ then the operator (13) is invertible.*

Proof. Suppose (13) is not invertible then there exist a cubic polynomial $\pi(x) \neq 0$ such that

$$\frac{\partial \pi^{[3]}}{\partial x}(x)(F + GK)x = -\frac{1}{2} \sum_k x'(C'_k + K'D'_k) \frac{\partial^2 \pi^{[3]}}{\partial x^2}(x)(C_k + D_k K)x$$

so

$$\left\| \frac{\partial \pi^{[3]}}{\partial x}(x)(F + GK)x \right\| = \left\| \frac{1}{2} \sum_k x'(C'_k + K'D'_k) \frac{\partial^2 \pi^{[3]}}{\partial x^2}(x)(C_k + D_k K)x \right\|$$

But we know that

$$\left\| \frac{\partial \pi^{[3]}}{\partial x}(x)(F + GK)x \right\| \geq \frac{1}{\rho} \|\pi(x)\| \geq 3\tau \|\pi(x)\| > \frac{3r\sigma^2}{2} \|\pi(x)\|$$

while

$$\left\| \frac{1}{2} \sum_k x'(C'_k + K'D_k) \frac{\partial^2 \pi^{[3]}}{\partial x^2}(x)(C_k + D_k K)x \right\| \leq \frac{3r\sigma^2}{2} \|\pi(x)\|$$

The takeaway message from this lemma is that if the nonzero entries of C_k, D_k are small relative to the nonzero entries of F, G then we can expect that (13) will be invertible.

There are at least two ways to try solve (10), the iterative approach or the direct approach. The iterative approach takes advantage of the MATLAB software `hjb.m` that we have written to solve the deterministic version of these equations [5]. This suggests an iteration scheme similar to the above for solving SARE. Let $\pi_{(0)}^{[3]}$ be the solution of the deterministic version of (10) where $C_k = 0, D_k = 0$. Given $\pi_{(\tau-1)}^{[3]}$ define

$$l_{(\tau)}^{[3]}(x, u) = l^{[3]}(x, u) + \frac{1}{2} \sum_k x'(C'_k + K'D_k) \frac{\partial^2 \pi_{\tau-1}^{[3]}}{\partial x^2}(x)(C_k + D_k K)x \\ + \sum_k x'(C'_k + K'D_k) P \gamma_k^{[2]}(x, u)$$

and let $\pi_{(\tau)}^{[3]}$ be the solution of

$$0 = \frac{\partial \pi_{(\tau)}^{[3]}}{\partial x}(x)(F + GK)x + x' P f^{[2]}(x, Kx) + l_{(\tau)}^{[3]}(x, Kx)$$

If this iteration converges then we have the solution to (10). More recently we have written software to find the Talor polynomials of $\pi(x)$ and $\kappa(x)$ directly, see `shjb.m` in our Nonlinear Systems Toolbox [5].

If (10) is solvable then solving (11) for $\kappa^{[2]}(x)$ is straightforward assuming that $R + \sum_k (C_k + D_k K)' P (C_k + D_k K)$ is invertible. If these equations are solvable then we can move on to the equations for $\pi^{[4]}(x)$ and $\kappa^{[3]}(x)$ and higher degrees.

It should be noted that if the Lagrangian is an even function and the dynamics is an odd function then the optimal cost $\pi(x)$ is an even function and the optimal feedback $\kappa(x)$ is an odd function.

5 Nonlinear Example

Here is a simple example with $n = 2, m = 1, r = 3$. Consider a pendulum of length 1 m and mass 1 kg orbiting approximately 400 kilometers above Earth on the International Space Station (ISS). The "gravity constant" at this height is approximately $g = 8.7 \text{ m/sec}^2$. The pendulum can be controlled by a torque u that can be applied at the pivot and there is damping at the pivot with linear damping constant $c = 0.1 \text{ kg/sec}$ and cubic damping constant $c_3 = 0.05 \text{ kg sec/m}^2$. Let x_1 denote the angle of pendulum measured counter clockwise from the outward pointing ray from the center of the Earth and let x_2 denote the angular velocity. The deterministic equations of motion are

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= lg \sin x_1 - c_1 x_2 - c_3 x_2^3 + u \end{aligned}$$

But the shape of the earth is not a perfect sphere and its density is not uniform so there are fluctuations in the "gravity constant". We set these fluctuations at around one percent although they are probably smaller. There might also be fluctuations in the damping constants of around one percent. Further assume that the commanded torque is not always realized and the relative error in the actual torque fluctuates around one percent. We model these stochastically by three white noises

$$\begin{aligned} dx_1 &= x_2 dt \\ dx_2 &= (lg \sin x_1 - c_1 x_2 - c_3 x_2^3 + u) dt \\ &\quad + 0.01lg \sin x_1 dw_1 - 0.01(c_1 x_2 + c_3 x_2^3) dw_2 + 0.01u dw_3 \end{aligned}$$

This is an example about how stochastic models with noise coefficients of order $O(x, u)$ can arise. If the noise is modeling an uncertain environment then its coefficients are likely to be $O(1)$. But if it is the model that is uncertain then noise coefficients are likely to be $O(x, u)$.

The goal is to find a feedback $u = \kappa(x)$ that stabilizes the pendulum to straight up in spite of the noises so we take the criterion to be

$$\min_u \frac{1}{2} \int_0^\infty \|x\|^2 + u^2 dt$$

Then

$$\begin{aligned} F &= \begin{bmatrix} 0 & 1 \\ 8.7 & 0.1 \end{bmatrix}, & G &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \\ Q &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & R &= 1, & S &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ C_1 &= \begin{bmatrix} 0 & 0 \\ 0.087 & 0 \end{bmatrix}, & C_2 &= \begin{bmatrix} 0 & 0 \\ 0 & -0.001 \end{bmatrix}, & C_3 &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

$$D_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad D_3 = \begin{bmatrix} 0 \\ 0.01 \end{bmatrix}$$

Because the Lagrangian is an even function and the dynamics is an odd function of x, u , we know that $\pi(x)$ is an even function of x and $\kappa(x)$ is an odd function of x .

We have computed the optimal cost $\pi(x)$ to degree 6 and the optimal feedback $\kappa(x)$ to degree 5,

$$\begin{aligned} \pi(x) &= 26.7042x_1^2 + 17.4701x_1x_2 + 2.9488x_2^2 \\ &\quad - 4.6153x_1^4 - 2.9012x_1^3x_2 - 0.5535x_1^2x_2^2 - 0.0802x_1x_2^3 - 0.0157x_2^4 \\ &\quad 0.3361x_1^6 + 0.1468x_1^5x_2 - 0.0015x_1^4x_2^2 - 0.0077x_1^3x_2^3 \\ &\quad - 0.0022x_1^2x_2^4 - 0.0003x_1x_2^5 + 0.000025058x_2^6 \\ \kappa(x) &= -17.4598x_1 - 5.8941x_2 \\ &\quad + 2.8995x_1^3 + 1.1064x_1^2x_2 + 0.2404x_1x_2^2 + 0.0628x_2^3 \\ &\quad - 0.1467x_1^5 + 0.0031x_1^4x_2 + 0.0232x_1^3x_2^2 \\ &\quad + 0.0089x_1^2x_2^3 + 0.0014x_1x_2^4 - 0.0002x_2^5 \end{aligned}$$

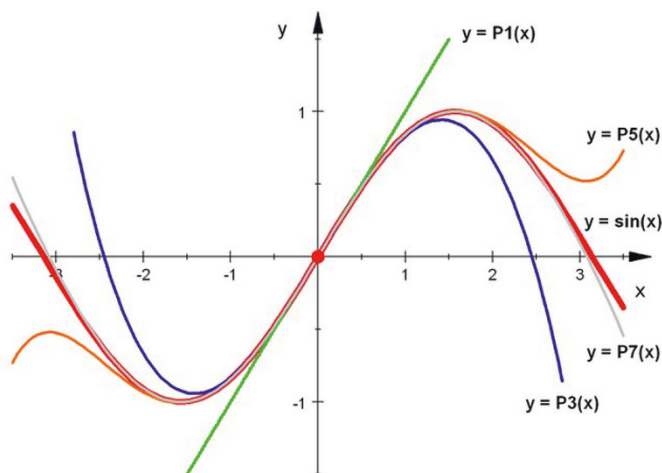


Fig. 1 Taylor approximations of $\sin(x)$

In making this computation we are approximating $\sin x_1$ by its Taylor polynomials

$$\sin x_1 = x_1 - \frac{x_1^3}{6} + \frac{x_1^5}{120} + \dots$$

The alternating signs of the odd terms in these polynomials are reflected in the nearly alternating signs in the Taylor polynomials of the optimal cost $\pi(x)$ and op-

timal feedback $\kappa(x)$. If we take a first degree approximation to $\sin x_1$ we are overestimating the gravitational force pulling the pendulum from its upright position pointing so $\pi^{[2]}(x)$ overestimates the optimal cost and the feedback $u = \kappa^{[1]}(x)$ is stronger than it needs to be. The latter could be a problem if there is a bound on the magnitude of u that we ignored in the analysis. If we take a third degree approximation to $\sin x_1$ then $\pi^{[2]}(x) + \pi^{[4]}(x)$ underestimates the optimal cost and the feedback $u = \kappa^{[1]}(x) + \kappa^{[3]}(x)$ is weaker than it needs to be. If we take a fifth degree approximation to $\sin x_1$ then $\pi^{[2]}(x) + \pi^{[4]}(x) + \pi^{[6]}(x)$ overestimates the optimal cost but by a smaller margin than $\pi^{[2]}(x)$. The feedback $u = \kappa^{[1]}(x) + \kappa^{[3]}(x) + \kappa^{[5]}(x)$ is stronger than it needs to be but by a smaller margin than $u = \kappa^{[1]}(x)$. This smaller margin may be important if there is a saturation limit on the control. Moreover since the quintic feedback is less aggressive than the linear feedback this may be advantageous in contolling a more complicated like a double pendulum.

6 Finite Horizon Stochastic Nonlinear Optimal Control Problem

Consider the finite horizon stochastic nonlinear optimal control problem,

$$\min_{u(\cdot)} E \left\{ \int_0^T l(t, x, u) dt + \pi_T(x(T)) \right\}$$

subject to

$$dx = f(t, x, u)dt + \sum_{k=1}^r \gamma_k(t, x, u)dw_k$$

$$x(0) = x^0$$

Again we assume that f, l, γ_k, π_T are sufficiently smooth.

If they exist and are smooth the optimal cost $\pi(t, x)$ of starting at x at time t and the optimal feedback $u(t) = \kappa(t, x(t))$ satisfy the time dependent Hamilton-Jacobi-Bellman equations (HJB)

$$0 = \min_u \left\{ \frac{\partial \pi}{\partial t}(t, x) + \frac{\partial \pi}{\partial x}(t, x)f(t, x, u) + l(t, x, u) \right. \\ \left. + \frac{1}{2} \sum_{l=1}^k \gamma'_k(t, x, u) \frac{\partial^2 \pi}{\partial x^2}(t, x) \gamma_k(t, x, u) \right\}$$

$$0 = \operatorname{argmin}_u \left\{ \frac{\partial \pi}{\partial x}(t, x)f(t, x, u) + l(t, x, u) \right. \\ \left. + \frac{1}{2} \sum_{k=1}^r \gamma'_k(t, x, u) \frac{\partial^2 \pi}{\partial x^2}(t, x) \gamma_k(t, x, u) \right\}$$

If the quantity to be minimized is strictly convex in u then HJB equations simplify to

$$0 = \frac{\partial \pi}{\partial t}(t, x) + \frac{\partial \pi}{\partial x}(t, x) f(t, x, \kappa(x)) + l(t, x, \kappa(x)) + \frac{1}{2} \sum_{k=1}^r \gamma'_k(t, x, \kappa(x)) \frac{\partial^2 \pi}{\partial x^2}(t, x) \gamma_k(t, x, \kappa(x)) \tag{1}$$

$$0 = \frac{\partial \pi}{\partial x}(x) \frac{\partial f}{\partial u}(t, x, \kappa(x)) + \frac{\partial l}{\partial u}(t, x, \kappa(x)) + \sum_{k=1}^r \gamma'_k(t, x, \kappa(x)) \frac{\partial^2 \pi}{\partial x^2}(x) \frac{\partial \gamma_k}{\partial u}(t, x, \kappa(x)) \tag{2}$$

Even if the quantity to be minimized is not convex in u then HJB equations imply these simplified equations but not necessarily vice versa.

These simplified equations are integrated backward in time from the final condition

$$\pi(T, x) = \pi_T(x) \tag{3}$$

Again we assume that we have the following Taylor expansions

$$\begin{aligned} f(t, x, u) &= F(t)x + G(t)u + f^{[2]}(t, x, u) + f^{[3]}(t, x, u) + \dots \\ l(t, x, u) &= \frac{1}{2} (x'Q(t)x + 2x'S(t)u + u'R(t)u) + l^{[3]}(t, x, u) + l^{[4]}(t, x, u) + \dots \\ \gamma_k(t, x, u) &= C_k(t)x + D_k(t)u + \gamma_k^{[2]}(t, x, u) + \gamma_k^{[3]}(t, x, u) + \dots \\ \pi_T(x) &= \frac{1}{2} x'P_T x + \pi_T^{[3]}(x) + \pi_T^{[4]}(x) + \dots \\ \pi(t, x) &= \frac{1}{2} x'P(t)x + \pi^{[3]}(t, x) + \pi^{[4]}(t, x) + \dots \\ \kappa(t, x) &= K(t)x + \kappa^{[2]}(t, x) + \kappa^{[3]}(t, x) + \dots \end{aligned}$$

where $^{[r]}$ indicates terms of homogeneous degree r in x, u with coefficients that are continuous functions of t .

The key assumption is that $\gamma_k(t, 0, 0) = 0$ for then (1) has a regular singular point at $x = 0$ and so is amenable to power series methods.

We plug these expansions into the simplified time dependent HJB equations and collect terms of lowest degree, that is, degree two in (1), degree one in (2) and degree two in (3).

$$\begin{aligned} 0 &= \dot{P}(t) + P(t)F(t) + F'(t)P(t) + Q(t) - K'(t)R(t)K(t) \\ &\quad + \sum_k (C'_k(t) + K'(t)D'_k(t)) P(t) (C_k(t) + D_k(t)K(t)) \end{aligned}$$

$$K(t) = - \left(R(t) + \sum_{k=1}^r D'_k(t)P(t)D_k(t) \right)^{-1} (G'(t)P(t) + S(t))$$

$$P(T) = P_T$$

We call these equations the stochastic differential Riccati equations (SDRE). Similar equations in more generality can be found in [6] but since we are interested in nonlinear problems we require that $\gamma_k(t, x, u) = O(x, u)$ so that the stochastic HJB equations have a regular singular point at the origin.

If SDRE are solvable we may proceed to the next degrees, degree three in (1), and degree two in (3).

$$0 = \frac{\partial \pi^{[3]}}{\partial t}(t, x) + \frac{\partial \pi^{[3]}}{\partial x}(t, x)(F(t) + G(t)K(t))x$$

$$+ x'P(t)f^{[2]}(t, x, K(t)x) + l^{[3]}(t, x, Kx)$$

$$+ \frac{1}{2} \sum_k x' C'_k(t) \frac{\partial^2 \pi^{[3]}}{\partial x^2}(t, x) (C_k + D_k(t)K(t)) (t)x$$

$$+ \sum_k x' (C'_k(t) + K'(t)D'_k(t)) P(t) \gamma_k^{[2]}(t, x)$$

$$0 = \frac{\partial \pi^{[3]}}{\partial x}(t, x)G(t) + x'P(t) \frac{\partial f^{[2]}}{\partial u}(t, x, K(t)x) + \frac{\partial l^{[3]}}{\partial u}(t, x, K(t)x)$$

$$+ \sum_k x' (C_k(t) + D_k(t)K(t))' \left(P(t) \frac{\partial \gamma_k^{[2]}}{\partial u}(x, K(t)x) + \frac{\partial^2 \pi^{[3]}}{\partial x^2}(x) D_k(t) \right)$$

$$+ \sum_k \gamma_k^{[2]}(x, K(t)x) P(t) D_k(t) + (\kappa^{[2]}(t, x))' \left(R(t) + \sum_k D'_k(t) P D_k(t) \right)$$

Notice again the unknown $\kappa^{[2]}(t, x)$ does not appear in the first equation which is linear ode for $\pi^{[3]}(t, x)$ running backward in time from the terminal condition,

$$\pi^{[3]}(t, x) = \pi_T^{[3]}(x)$$

After we have solved it then the second equation for $\kappa^{[2]}(t, x)$ is easily solved because of the standard assumption that $R(t)$ is invertible and hence $R(t) + G'(t)P(t)G(t) + \sum_k D'_k(t)P(t)D_k(t)$ is invertible.

The higher degree terms can be found in a similar fashion.

7 Conclusion

We have considered nonlinear continuous time stochastic optimal control problems over infinite and finite horizons under the assumption the coefficients of the noise terms are $O(x, u)$. This assumption implies that corresponding Hamilton-Jacobi-Bellman equations have regular singular points at the origin and so we can compute the Taylor polynomials of the optimal cost and optimal feedback degree by degree. At the lowest degrees, two in optimal cost and one in the optimal feedback, we obtained Riccati equations. The infinite horizon Riccati equation is a new algebraic equation while the finite horizon Riccati equation is a familiar differential equation. The infinite horizon higher degree terms are found by solving linear algebraic equations while the finite horizon higher degree terms are found by solving linear differential equations. We have written general purpose MATLAB code to solve the algebraic equations.

8 Acknowledgements

The author would like to thank Mark Davis, Wendell Fleming, Alan Laub, George Yin and especially Peter Caines for their helpful comments and AFOSR for their support.

References

1. E. G. Al'brekht, *On the optimal stabilization of nonlinear systems*, PMM-J. Appl. Math. Mech., 25:1254-1266, 1961.
2. W. E. Boyce and R. C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, Tenth Edition, Wiley, New Jersey, 2009.
3. W. Fleming and R. Rishel, *Deterministic and Stochastic Optimal Control*, Springer, New York, 1975.
4. A. J. Krener, Stochastic HJB Equations and Regular Singular Points, arXiv:1806.04120v1 [math.OC] 11 Jun 2018.
5. A. J. Krener, Nonlinear Systems Toolbox 2018. Available upon request to the author.
6. J. Yong and X. J. Zhou, *Stochastic Controls, Hamiltonian Systems and HJB Equations*, Springer, New York, 1999.



Information Diffusion in Social Networks: Friendship Paradox Based Models and Statistical Inference

Vikram Krishnamurthy and Buddhika Nettasinghe

Abstract Dynamic models and statistical inference for the diffusion of information in social networks is an area which has witnessed remarkable progress in the last decade due to the proliferation of social networks. Modeling and inference of diffusion of information has applications in targeted advertising and marketing, forecasting elections, predicting investor sentiment and identifying epidemic outbreaks. This chapter discusses three important aspects related to information diffusion in social networks: (i) How does observation bias due to the *friendship paradox* (on average your friends have more friends than you do) and *monophilic contagion* (influence of friends of friends) affect the information diffusion dynamics? (ii) How can social networks adapt their structural connectivity depending on the state of information diffusion? (iii) How one can estimate the state of the network induced by information diffusion? The motivation for these three topics stems from recent results in network science and social sensing.

1 Introduction

Information diffusion refers to how the opinions (states) of individual nodes in a social network (graph) evolve with time. The two phenomena that give rise to information diffusion in social networks are *contagion* and *homophily*. Contagion-based diffusions are driven by influence of neighbors whereas homophily-based diffusions

Vikram Krishnamurthy
Cornell Tech and School of Electrical & Computer Engineering, Cornell University
e-mail: vikramk@cornell.edu

Buddhika Nettasinghe
Cornell Tech and School of Electrical & Computer Engineering, Cornell University
e-mail: dwn26@cornell.edu

This research was supported by the Army Research Office under grant W911NF-17-1-0335 and National Science Foundation under grant 1714180.

are driven by properties of nodes (which are correlated among neighbors) [3, 63, 84]. Dynamic models and statistical inference for such information diffusion processes in social networks (such as news, innovations, cultural fads, etc) has witnessed remarkable progress in the last decade due to the proliferation of social media networks such as Facebook, Twitter, YouTube, Snapchat and also online reputation systems such as Yelp and Tripadvisor. Models and inference methods for information diffusion in social networks are useful in a wide range of applications including selecting influential individuals for targeted advertising and marketing [41, 68, 83], localization of natural disasters [82], forecasting elections [69] and predicting sentiment of investors in financial markets [75, 8]. For example, [4] shows that models based on the rate of Tweets for a particular product can outperform market-based prediction methods.

This chapter deals with the contagion-based information diffusion in large scale social networks. In such contagion-based information diffusion (henceforth referred to as information diffusion) processes, states (which represent opinions, voting intentions, purchase of a product, etc.) of individuals in the network evolve over time as a probabilistic function of the states of their neighbors. Popular models for studying information diffusion processes over networks include Susceptible-Infected (SI), Susceptible-Infected-Susceptible (SIS), Susceptible-Infected-Recovered (SIR) and Susceptible-Exposed-Infected-Recovered (SEIR) [32, 14]. Apart from these models, several recent works also investigated information diffusions using real-world social network datasets: [80] studied the spread of hashtags on Twitter, [6] conducted large scale field experiments to identify the causal effects of peer influence in information diffusion, [56] studied how the network structure affects dynamics of information flow using Digg and Twitter datasets to track how interest in new stories spread over them.

Main Topics and Organization

In this chapter, we consider a discrete time version of the SIS model on an undirected network which involves two steps (detailed in Sec. 2) at each time instant. In the first step, a randomly sampled individual (agent) m from the population observes $d(m)$ (degree of m) number of randomly selected agents (neighbors of m). In the second step, based on the $d(m)$ observations, the state of agent m evolves probabilistically to one of the two possible states: infected or susceptible.

In the context of this discrete time SIS model, next, we briefly discuss the main topics studied in this chapter, motivation for studying them and how they are organized throughout this chapter. Further, how the main topics discussed in different sections are interconnected with each other and unified under the main theme of dynamic modeling and statistical inference of information diffusion processes is illustrated in Fig. 1.

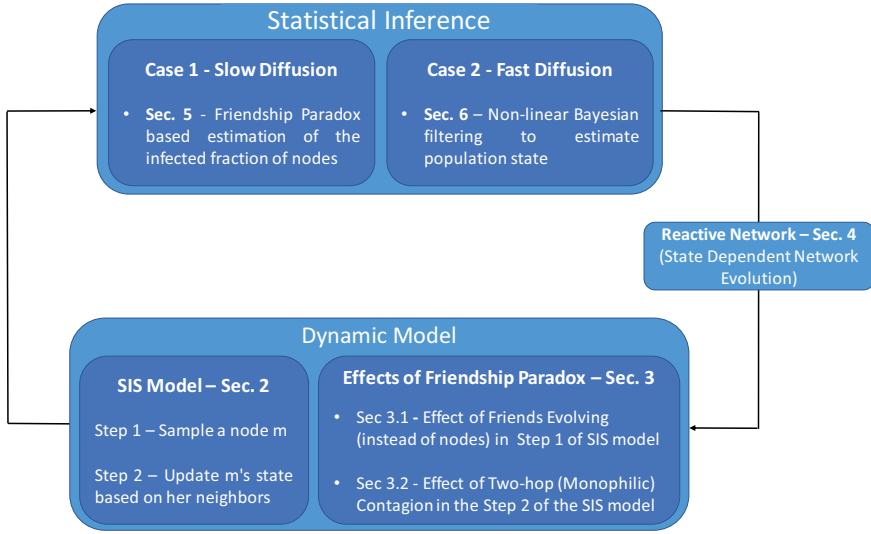


Fig. 1: Block diagram illustrating the main topics covered and their organization in this chapter. The main topics are unified under the central theme of dynamic modeling and statistical inference of information diffusion processes over social networks.

1. Friendship Paradox based Variants of the SIS model

The first topic (Sec. 3) studies the effects of friendship paradox on the SIS model. The *friendship paradox* refers to a graph theoretic consequence that was introduced in 1991 by Scott. L. Feld in [17]. Feld’s original statement of the friendship paradox is “on average, the number of friends of a random friend is always greater than or equal to the number of friends of a random individual”. Here, a random friend refers to a random end node Y of a randomly chosen edge (a pair of friends) in the network. This friendship paradox is formally stated as follows:

Theorem 1 (Friendship Paradox [17]). *Let $G = (V, E)$ be an undirected graph, X be a node chosen uniformly from V and, Y be a uniformly chosen node from a uniformly chosen edge $e \in E$. Then,*

$$\mathbb{E}\{d(Y)\} \geq \mathbb{E}\{d(X)\}, \tag{1}$$

where, $d(X)$ denotes the degree of X .

Studying the friendship paradox (Theorem 1) based variants of the SIS model is motivated by the following two assumptions made in most works (for example, see [59, 39, 60, 77, 40]) related to SIS models.

- i. Each node is equally likely to update its state at each time instant i.e. uniform nodes are sampled in the first step of the SIS model.
- ii. Individuals decide whether to get infected or not based only on their (immediate) neighbors' states.

In real world social networks (e.g. Facebook or Twitter), the frequency with which a node updates her state (e.g. opinion) depends on the number of her social interactions (i.e. degree) according to recent findings [35]. This contradicts the assumption i. As a solution, Sec. 3.1 studies the modified SIS model where the state of a random friend (instead of a random node) evolves at each time instant. This modification to the standard SIS model reflects the fact that high degree nodes evolve more often in real world social networks. The main result of Sec. 3.1 shows that this modification results in different dynamics (compared to the standard SIS model) but, with the same critical thresholds (which determine if the information diffusion process will eventually die away or not) on the parameters of the SIS model.

Further, it has been shown in several recent works (e.g. [2, 18]) that the individuals' attributes and decisions in real world social networks are affected by two-hop neighbors (i.e. friends of friends). This two-hop neighbors' effects in real world networks are ignored in the assumption ii of the standard SIS model. As an alternative, Sec. 3.2 considers the case where friends of friends influences the state evolutions instead of friends. We refer to this two-hop influence as *monophilic contagion* since the correlation between two-hop nodes is called *monophily*¹ [2]. Main result of Sec. 3.2 shows that information diffusion processes under monophilic contagion (decision to adopt a product, an idea, etc. is based on two-hop neighbors) spreads more easily (i.e. has a smaller critical threshold) compared to information diffusion under non-monophilic contagion (one-hop influence) as a result of the friendship paradox². This result also suggests that talking to random friends of friends could be more efficient (compared to talking to random friends) in spreading rumors, news, etc. The well known friendship paradox based immunization approach [12] that immunizes random friends (instead of random nodes) relies on a similar argument: random friends have larger degrees (compared to random nodes) and are more critical to the spreading of a disease.

2. SIS Model and Reactive Networks: Collective Dynamics

Modeling a network as a deterministic graph does not capture information diffusion processes in real world networks. Several works proposed and analyzed evolving graph models: [73] studied the adaptive susceptible-infected-susceptible (ASIS) model where susceptible individuals are allowed to temporarily cut edges connecting them to infected nodes in order to prevent the spread of the infection, [76] analyzed the stability of epidemic processes over time-varying networks and provides

¹ The concept of monophily presented in [2] does not give a causal interpretation but only the correlation between two-hop neighbors of an undirected graph. What we consider is monophilic contagion (motivated by monophily): the information diffusion caused by the influence of two hop neighbors in an undirected network.

² Effects of the friendship paradox on information diffusion have been considered in [54, 5, 55]. However, the effect of friendship paradox on information diffusion under monophilic contagion (two-hop influence) has not been explored in the literature to the best of our knowledge.

sufficient conditions for convergence, [74] studied a SIS process over a static contact network where the nodes have partial information about the epidemic state and react by limiting their interactions with their neighbors when they believe the epidemic is currently prevalent. These serve as the motivation for Sec. 4 where the underlying network is modeled as a *reactive network*: a random graph process whose transition probabilities at each time instant depend on the state of the information diffusion process. The main result of Sec. 4 shows that, when the network is a *reactive network* which randomly evolves depending on the state of the information diffusion, the collective dynamics of the network and the diffusion process can be approximated (under some assumptions) by an ordinary differential equation (ODE) with an algebraic constraint. From a statistical modeling and machine learning perspective, the importance of this result relies on the fact that it provides a simple deterministic approximation of the collective stochastic dynamics of a complex system (an SIS process on a random graph, both evolving on the same time scale).

3. Estimating the Population State under Slow and Fast Information Diffusion

Sec. 5 and Sec. 6 deal with estimating the population states induced by the SIS model under two cases:

1. the information diffusion is slow and hence states of nodes can be treated as fixed for the purpose of the estimating the population state
2. the information diffusion is fast and hence, states of nodes cannot be treated as fixed for the purpose of estimation.

Case 1 - Polling under slow information diffusion

Polling is the method of asking a question from randomly (according to some distribution) sampled individuals and averaging their responses [26]. Therefore, the accuracy of a poll depends on two factors: (i) - method of sampling respondents for the poll (ii) - question presented to the sampled individuals. For example, when forecasting the outcome of an election, asking people "Who do you think will win?" (expectation polling) is better compared to "Who will you vote for?" (intent polling) [81]. This is due to the fact that an individual will name the candidate that is most popular among her friends in expectation polling (and thus summarizing a number of individuals in the social network) instead of providing her own voting intention. Motivated by such polling approaches, Sec. 5 presents two friendship paradox based polling algorithms that aim to estimate the fraction of infected individuals by querying random friends instead of random individuals. Since random friends have more friends (and hence, have more observations) than random individuals on average, the proposed methods yield a better (in a mean squared error sense) estimate compared to intent polling as well as expectation polling with random nodes.

Case 2 - Bayesian filtering under fast information diffusion

Friendship paradox based polling algorithms in Sec. 5 assume that information diffusion takes place on a slower time scale compared to the time taken to poll the individuals. Hence, for the purpose of the polling algorithm, the states of the individuals can be treated as fixed. However, such approaches are not applicable in

situations where the information diffusion takes place on the same time scale as the time scale on which individuals are polled i.e. cases where measurement (polling) process takes place on the same time scale as the one on which the information spreads. Further, the information diffusion process constitute a non-linear (in the states) dynamical system as we show subsequently in Sec. 2. Hence applying optimal filtering algorithms such as Kalman filter is not possible. These facts motivate the non-linear filtering algorithm discussed in Sec. 6 which recursively (with each new measurement) computes the conditional mean of the state of the information diffusion (given the observations).

Summary

The main topics explored in this chapter bring together two important aspects related to a stochastic dynamical system mentioned at the beginning of this chapter: dynamic modeling and statistical inference. In terms of the dynamic modeling aspect (which is covered in Sec. 2, Sec. 3 and Sec. 4), we are interested in understanding how changes to the standard SIS-model can result in different dynamics and stationary states. In terms of statistical inference (covered in Sec. 5 and Sec. 6), we are interested in estimating the underlying state of the population induced by the model. Fig. 1 illustrates how these topics are organized in this chapter and are interconnected under the unifying theme. Rather than delving into detailed proofs, our aim in this chapter is to stress several novel insights.

2 Mean-Field Dynamics of SIS Model and Friendship Paradox

Mean-field dynamics refers to a simplified model of a (stochastic) system where the stochastic dynamics are replaced by deterministic dynamics. Much of this research is based on the seminal work of Kurtz [52] on population dynamics models. In this section, we first discuss how mean-field dynamics can be used as a deterministic model of a SIS diffusion process over an undirected network. Since an SIS diffusion over a social network is a Markov process whose state space grows exponentially with the number of individuals, mean-field dynamics offers a deterministic model that is analytically tractable [59, 60, 39, 46]. Then, several recent generalizations of the original version of the friendship paradox are presented. The purpose of mean-field dynamics and the friendship paradox results discussed in this section is to study (in Sec. 3) how friendship paradox based changes to the standard SIS model (e.g. random friends evolving instead of random nodes in the step 1 of SIS model) can result in different mean-field dynamics and critical thresholds.

2.1 Discrete time SIS Model

Consider a social network represented by an undirected graph $G = (V, E)$ where $V = \{1, 2, \dots, M\}$ denotes the set of nodes. At each discrete time instant n , a node $v \in V$ of the network can take the state $s_n^{(v)} \in \{0, 1\}$ where, 0 denotes the susceptible state and 1 denotes the infected state. The degree $d(v) \in \{1, \dots, D\}$ of a node $v \in V$ is the number of nodes connected to v and, $M(k)$ denotes the total number of nodes with degree k . Then, the degree distribution $P(k) = \frac{M(k)}{M}$ is the probability that a randomly selected node has degree k . Further, we also define the population state $\bar{x}_n(k)$ as the fraction of nodes with degree k that are infected (state 1) at time n i.e.

$$\bar{x}_n(k) = \frac{1}{M(k)} \sum_{v \in V} \mathbb{1}_{\{d(v)=k, s_n^{(v)}=1\}}, \quad k = 1, \dots, D. \tag{2}$$

For this setting, we adopt the SIS model used in [46, 47] which is as follows briefly.

Discrete Time SIS Model: At each discrete time instant n ,

Step 1: A node $m \in V$ is chosen with uniform probability $p^X(m) = 1/M$ where, M is the number of nodes in the graph.

Step 2: The state $s_n^{(v)} \in \{0, 1\}$ of the sampled node m (in Step 1) evolves to $s_{n+1}^{(v)} \in \{0, 1\}$ with transition probabilities that depend on the degree of m , number of infected neighbors of m , population state of the network \bar{x}_n ³ and the current state of $s_n^{(m)}$.

Note that the above model is a Markov chain with a state space consisting of 2^M states (since each of the M nodes can be either infected or susceptible at any time instant). Due to this exponentially large state space, the discrete time SIS model is not mathematically tractable. However, we are interested only in the fraction of the infected nodes (as opposed to the exact state out of the 2^M states) and therefore, it is sufficient to focus on the dynamics of the population state \bar{x}_n defined in (2) instead of the exact state of the infection.

2.2 Mean-Field Dynamics Model

Mean-field dynamics has been used in literature (e.g. [59, 60, 39, 52, 46]) as a useful means of obtaining a tractable deterministic model of the dynamics of the population state \bar{x}_n . The following result from [46] shows how mean-field dynamics model closely approximates the stochastic dynamics of the true population state \bar{x}_n .

³ $\bar{x}_n(k)$ is the fraction of infected nodes with degree k i.e. $\bar{x}_n(k) = \frac{M^1(k)}{M(k)}$ where $M^1(k)$ is the number of infected nodes with degree k and $M(k)$ is the number of nodes with degree k .

Theorem 2 (Mean-Field Dynamics). *With M denoting the number of nodes in the network:*

1. *The population state defined in (2) evolves according to the following stochastic difference equation driven by a martingale difference process ζ :*

$$\bar{x}_{n+1}(k) = \bar{x}_n(k) + \frac{1}{M} [P_{01}(k, \bar{x}_n) - P_{10}(k, \bar{x}_n)] + \zeta_n \tag{3}$$

Here,

$$P_{01}(k, \bar{x}_n) = (1 - \bar{x}_n(k)) \mathbb{P}(s_{n+1}^m = 1 | s_n^m = 0, d(m) = k, \bar{x}_n) \tag{4}$$

$$P_{10}(k, \bar{x}_n) = \bar{x}_n(k) \mathbb{P}(s_{n+1}^m = 0 | s_n^m = 1, d(m) = k, \bar{x}_n). \tag{5}$$

are the scaled transition probabilities of the states and, ζ_n is a martingale difference process with $\|\zeta_n\|_2 \leq \frac{\Gamma}{M}$ for some positive constant Γ .

2. *Consider the mean-field dynamics process associated with the population state:*

$$x_{n+1}(k) = x_n(k) + \frac{1}{M} (P_{01}(k, x_n) - P_{10}(k, x_n)) \tag{6}$$

where, $P_{01}(k, x_n)$ and $P_{10}(k, x_n)$ are defined in (4), (5) and initial state $x_0 = \bar{x}_0$. Then, for a time horizon of T points, the deviation between the mean-field dynamics (6) and the actual population state \bar{x}_n of the SIS model satisfies

$$\mathbb{P}\left\{ \max_{0 \leq n \leq T} \|x_n - \bar{x}_n\|_\infty \geq \varepsilon \right\} \leq C_1 \exp(-C_2 \varepsilon^2 M) \tag{7}$$

for some positive constants C_1, C_2 providing $T = O(M)$.

The first part of Theorem 2 is the classical martingale representation of a Markov chain (which is the population state \bar{x}_n). Note from (3) that the dynamics of the population state \bar{x}_n resemble a stochastic approximation recursion (new state is the old state plus a noisy term). Hence, the trajectory of the population state \bar{x}_n should converge (weakly) to the deterministic trajectory given by the ODE corresponding to the mean-field dynamics in (6) as the size of the network M goes to infinity i.e. the step size of the stochastic approximation algorithm goes to zero (for details, see [45, 53]). Second part of the Theorem 2 provides an exponential bound on the deviation of the mean-field dynamics model from the actual population state for a finite length of the sample path. In the subsequent sections of this chapter, the mean-field approximation (6) is utilized to explore the topics outlined in Sec. 1.

2.3 Friendship Paradox

Recall that the friendship paradox (Theorem 1) is a comparison between the average degrees of a random individual X and a random friend Y . This subsection reviews recent generalizations and extensions of the friendship paradox stated in Theorem 1.

The original version of the friendship paradox (Theorem 1) can be described more generally in terms of likelihood ratio ordering as follows:

Theorem 3 (Friendship Paradox - Version 1 [9]). *Let $G = (V, E)$ be an undirected graph, X be a node chosen uniformly from V and, Y be a uniformly chosen node from a uniformly chosen edge $e \in E$. Then,*

$$d(Y) \geq_{lr} d(X), \tag{8}$$

where, \geq_{lr} denotes the likelihood ratio dominance⁴.

Theorem 4 (based on [9]) states that a similar result holds when the degrees of a random node X and a random friend Z of a random node X are compared as well.

Theorem 4 (Friendship Paradox - Version 2 [9]). *Let $G = (V, E)$ be an undirected graph, X be a node chosen uniformly from V and, Z be a uniformly chosen neighbor of a uniformly chosen node from V . Then,*

$$d(Z) \geq_{fbsd} d(X) \tag{9}$$

where, \geq_{fbsd} denotes the first order stochastic dominance⁵.

The intuition behind the two versions of the friendship paradox (Theorems 3 and 4) stems from the fact that individuals with a large number of friends (high degree nodes) appear as the friends of a large number of individuals. Therefore, high degree nodes contributes to an increase in the average number of friends of friends. On the other hand, individuals with smaller number of friends appear as friends of a smaller number of individuals. Hence, they do not cause a significant change in the average number of friends of friends.

Friendship paradox, which in essence is a sampling bias observed in undirected social networks has gained attention as a useful tool for estimation and detection problems in social networks. For example, [16] proposes to utilize friendship paradox as a sampling method for reduced variance estimation of a heavy-tailed degree distribution, [11, 20, 85] explore how the friendship paradox can be used for detecting a contagious outbreak quickly, [83, 54, 37, 42, 51] utilizes friendship paradox for maximizing influence in a social network, [69] proposes friendship paradox based algorithms for efficiently polling a social network (e.g. to forecast an election) in a social network, [38] studies how the friendship paradox in a game theoretic setting can systematically bias the individual perceptions.

⁴ A discrete random variable Y (with a probability mass function f_Y) likelihood ratio dominates a discrete random variable X (with a probability mass function f_X), denoted $Y \geq_{lr} X$ if, $f_Y(n)/f_X(n)$ is an increasing function of n . Further, likelihood ratio dominance implies larger mean. Therefore, Theorem 3 implies that $\mathbb{E}\{d(Y)\} \geq \mathbb{E}\{d(X)\}$ as stated in Theorem 1.

⁵ A discrete random variable Y (with a cumulative distribution function F_Y) first order stochastically dominates a discrete random variable X (with a cumulative distribution function F_X), denoted $Y \geq_{fbsd} X$ if, $F_Y(n) \leq F_X(n)$, for all n . Further, first order stochastic dominance implies larger mean. Hence, Theorem 4 implies that $\mathbb{E}\{d(Z)\} \geq \mathbb{E}\{d(X)\}$.

Several generalizations, extensions and consequences of friendship paradox have also been proposed in the literature. [15] shows how friendship paradox can be generalized to other attributes (apart from the degree) such as income and happiness when there exists a positive correlation between the attribute and the degree. Related to this work, [33] showed that certain other graph based centrality measures such as eigenvector centrality and Katz centrality (under certain assumptions) exhibit a version of the friendship paradox, leading to the statement “your friends are more important than you, on average”. [34] extended the concept of friendship paradox to directed networks and empirically showed that four versions of friendship paradox which compare the expected in- and out- degrees of random friends and random followers to expected degree of a random node can exist in directed social networks such as Twitter. [57] discusses “majority illusion” an observation bias that stems from friendship paradox which makes many individuals in a social network to observe that a majority of their neighbors are in a particular state (e.g. possesses an iPhone), even when that state is globally rare. Similarly, [5, 43, 7, 19] also discuss various other generalizations and consequences of friendship paradox.

3 Effects of Friendship Paradox on SIS Model

Sec. 2 reviewed the discrete time SIS model that involves two steps and, showed how mean-field dynamics can be used as a deterministic model of an SIS information diffusion process. In the context of the SIS model, the aim of this section is to explore how changes (motivated by examples discussed in Sec. 1) to the first step (sampling a node m) and the second step (m updates its state probabilistically based on the states of neighbors) of the standard SIS model are reflected in the deterministic mean-field dynamics model and its critical threshold. The changes to the standard SIS model (Sec. 2.1) that we explore are motivated by friendship paradox in the sense that we consider 1 - random friends (instead of random nodes) are sampled in the first step, 2 - state of the sampled node is updated based on the states of friends of friends (instead of immediate friends).

3.1 *Effect of the Sampling Distribution in the Step 1 of the SIS Model*

Recall from Sec. 2.3 that we distinguished between three sampling methods for a network $G = (V, E)$: a random node X , a random friend Y and, a random friend Z of a random node. Further, recall that in the discrete-time SIS model explained in Sec. 2.1, the node m that whose state evolves is sampled uniformly from V i.e. $m \stackrel{d}{=} X$. This section studies the effect of random friends (Y or Z) evolving at each time

instant instead of random nodes (X) i.e. the cases where $m \stackrel{d}{=} Y$ or $m \stackrel{d}{=} Z$. Following is the main result in this section:

Theorem 5. *Consider the discrete time SIS model presented in Sec. 2.1.*

1. *If the node m is a random end Y of random link i.e. node m with degree $d(m)$ is chosen with probability $p^Y(m) = \frac{d(m)}{\sum_{v \in V} d(v)}$, then the stochastic dynamics of the SIS model can be approximated by,*

$$x_{n+1}(k) = x_n(k) + \frac{1}{M} \frac{k}{\bar{k}} (P_{01}(k, x_n) - P_{10}(k, x_n)), \quad (10)$$

where \bar{k} is the average degree of the graph $G = (V, E)$.

2. *If the node m is a random neighbor Z of a random node X , then the stochastic dynamics of the SIS model can be approximated by,*

$$x_{n+1}(k) = x_n(k) + \frac{1}{M} \left(\sum_{k'} \frac{P(k)}{P(k')} P(k|k') \right) (P_{01}(k, x_n) - P_{10}(k, x_n)), \quad (11)$$

where \bar{k} is the average degree of the graph $G = (V, E)$, P is the degree distribution and $P(k|k')$ is the probability that a random neighbor of a degree k' node is of degree k . Further, if the network is a degree-uncorrelated network i.e. $P(k|k')$ does not depend on k' , then (11) will be the same as (10).

Theorem 5 is proved in [70]. Theorem 5 shows that, if the node m sampled in the step 1 of the SIS model (explained in Sec. 2.1), is chosen to be a random friend or a random friend of a random node, then different elements $x_n(k)$ of the mean-field approximation evolves at different rates. This result allows us to model the dynamics of the population state in the more involved case where, frequency of the evolution of an individual is proportional his/her degree (part 1 - e.g. high degree nodes change opinions more frequently due to higher exposure) and also depends on the degree correlation (part 2 - e.g. nodes being connected to other similar/different degree nodes changes the frequency of changing the opinion).

Remark 1 (Invariance of the critical thresholds to the sampling distribution in step 1). The stationary condition for the mean-field dynamics is obtained by setting $x_{n+1}(k) - x_n(k) = 0$ for all $k \geq 1$. Comparing (6) with (10) and (11), it can be seen that this condition yields the same expression $P_{01}(k, x_n) - P_{10}(k, x_n) = 0$, for all three sampling methods (random node - X , random end of a random link Y and, a random neighbor Z of a random node). Hence, the critical thresholds of the SIS model are invariant to the distribution from which the node m is sampled in step 1. This leads us to Sec. 3.2 where, modifications to the step 2 of the SIS model are analyzed in terms of the critical thresholds.

3.2 Critical Thresholds for Unbiased-degree Networks

In Sec. 3.1 of this paper, we focused on step 1 of the SIS model (namely, distribution with which the node m is drawn at each time instant) and, showed that different sampling methods for selecting the node m result in different mean-field dynamics with the same stationary conditions. In contrast, the focus of this subsection is on the step 2 of the SIS model (namely, the probabilistic evolution of the state of the node m sampled in step 1) and, how changes to step 2 would result in different stationary conditions and critical thresholds. More specifically, we are interested in understanding the effects on the SIS information diffusion process caused by *monophilic contagion*: node m 's state evolves based on the states of random friends of friends (two-hop neighbors). This should be contrasted to the standard SIS information diffusion processes based on non-monophilic contagion where, evolution of node m 's state is based on states of random friends (one-hop neighbors).

3.2.1 Critical Thresholds of Information Diffusion Process under Monophilic and Non-Monophilic Contagion Rules

Recall the SIS model reviewed in Sec. 2.1 again. We limit our attention to the case of *unbiased-degree* networks and viral adoption rules discussed in [61].

Unbiased-degree network: In an unbiased-degree network, neighbors of agent m sampled in the step 1 of the SIS model are $d(m)$ (degree of agent m) number of uniformly sampled agents (similar in distribution to the random variable X) from the network. Therefore, in an unbiased-degree network, any agent is equally likely to be a neighbor of the sampled (in the step 1 of the SIS model) agent m .

Viral adoption rules⁶: If the sampled agent m (in the step 1 of the SIS model) is an infected agent, she becomes susceptible with a constant probability δ . If the sampled agent m (in the step 1 of the SIS model) is a susceptible (state 0) agent, she samples $d(m)$ (degree of m) number of other agents $X_1, X_2, \dots, X_{d(m)}$ (neighbors of m in the unbiased-degree network) from the network and, updates her state (infected or susceptible) based on one of the following rules:

Case 1 - Non-monophilic contagion: For each sampled neighbor X_i , m observes the state of X_i . Hence, agent m observes the states of $d(m)$ number of random nodes. Let a_m^X denote the number of infected agents among $X_1, \dots, X_{d(m)}$. Then, the susceptible agent m becomes infected with probability $\frac{va_m^X}{D}$ where, $0 \leq v \leq 1$ is a constant and D is the largest degree of the network.

Case 2 - Monophilic contagion: For each sampled neighbor X_i , m observes the state of a random friend $Z_i \in \mathcal{N}(X_i)$ of that neighbor. Hence, agent m observes

⁶ The two rules (case 1 and case 2) are called viral adoption rules as they consider the total number of infected nodes (denoted by a_m^X and a_m^Z in case 1 and case 2 respectively) in the sample in contrast to the persuasive adoption rules that consider the fraction of infected nodes in the sample [60].

the states of $d(m)$ number of random friends $Z_1, \dots, Z_{d(m)}$ of random nodes $X_1, \dots, X_{d(m)}$. Let a_m^Z be the number of infected agents among $Z_1, \dots, Z_{d(m)}$.

Then, the susceptible agent m becomes infected with probability $\frac{\nu a_m^Z}{D}$ where, $0 \leq \nu \leq 1$ is a constant and D is the largest degree of the network.

In order to compare the non-monophilic and monophilic contagion rules, we look at the conditions on the model parameters for which, each rule leads to a positive fraction of infected nodes starting from a small fraction of infected nodes i.e. a positive stationary solution to the mean-field dynamics (6). The main result is the following (proof given in [70]):

Theorem 6. Consider the SIS model described in Sec. 2.1. Define the effective spreading rate as $\lambda = \frac{\nu}{\delta}$ and let X be a random node and Z be a random friend of X .

1. Under the non-monophilic contagion rule (Case 1), the mean-field dynamics equation (6) takes the form,

$$x_{n+1}(k) = x_n(k) + \frac{1}{M} \left((1 - x_n(k)) \frac{\nu k \theta_n^X}{D} - x_n(k) \delta \right) \tag{12}$$

where,

$$\theta_n^X = \sum_k P(k) x_n(k) \tag{13}$$

is the probability that a randomly chosen node X at time n is infected. Further, there exists a positive stationary solution to the mean field dynamics (12) for case 1 if and only if

$$\lambda > \frac{D}{\mathbb{E}\{d(X)\}} = \lambda_X^* \tag{14}$$

2. Under the monophilic contagion rule (Case 2), the mean-field dynamics equation (6) takes the form,

$$x_{n+1}(k) = x_n(k) + \frac{1}{M} \left((1 - x_n(k)) \frac{\nu k \theta_n^Z}{D} - x_n(k) \delta \right) \tag{15}$$

where,

$$\theta_n^Z = \sum_k \left(\sum_{k'} P(k') P(k|k') \right) x_n(k) \tag{16}$$

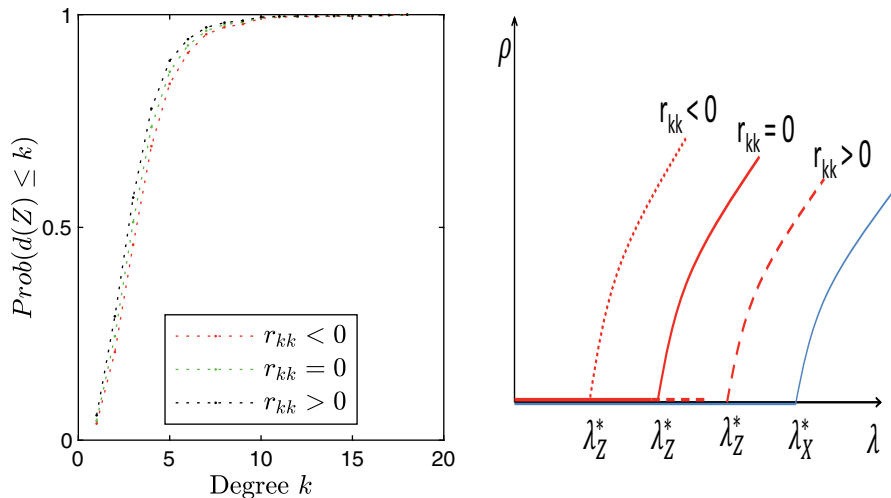
is the probability that a randomly chosen friend Z of a randomly chosen node X at time n is infected⁷. Further, there exists a positive stationary solution to the mean field dynamics (15) if and only if

⁷ We use $P(k|k')$ to denote the conditional probability that a node with degree k' is connected to a node with degree k . More specifically $P(k|k') = \frac{e(k,k')}{q(k)}$ where $e(k,k')$ is the joint degree distribution of the network and $q(k)$ is the marginal distribution that gives the probability of random end (denoted by random variable Y in Theorem 1) of random link having degree k . We also use σ_q to denote the variance of $q(k)$ in subsequent sections.

$$\lambda > \frac{D}{\mathbb{E}\{d(Z)\}} = \lambda_Z^* \tag{17}$$

The proof of the first part of Theorem 6 is inspired by [60, 61] that consider the unbiased degree networks with non-monophilic adoption rules and continuous-time evolutions (as opposed to the discrete time case considered here). The main purpose of the first part is to provide a comparison of the non-monophilic adoption rule with the monophilic adoption rule (part 2) under the same setting. Theorem 6 allows us to analyze the effects of friendship paradox and degree-assortativity on the diffusion process as discussed in the next subsection.

3.2.2 Effects of Friendship Paradox and Degree Correlation on Information Diffusion under Monophilic Contagion



(a) CDFs of the of the degree $d(Z)$ of a random friend Z of a random node for three networks with same degree distribution but different assortativity r_{kk} values. Note that the CDFs are point-wise increasing with r_{kk} showing that $\mathbb{E}\{d(Z)\}$ decreases with r_{kk} .

(b) Variation of the stationary fraction ρ of infected nodes with the effective spreading rate λ for the case 1 (blue) and case 2 (red), illustrating the ordering of the critical thresholds of cases 1,2 and the effect of assortativity.

Fig. 2: Comparison of non-monophilic and monophilic contagion rules and the effect of assortativity on the critical thresholds of the monophilic contagion.

Theorem 6 showed that the critical thresholds of the mean-filed dynamics equation (6) for the two rules (non-monophilic and monophilic contagion) are different.

Following is an immediate corollary of Theorem 6 which gives the ordering of these critical thresholds using the friendship paradox stated in Theorem 1.

Corollary 1. *The critical thresholds λ_X^*, λ_Z^* in (14), (17) for the cases of non-monophilic (case 1) and monophilic (case 2) contagion rules satisfy*

$$\lambda_Z^* \leq \lambda_X^*. \tag{18}$$

Corollary 1 shows that in the case of information diffusion under monophilic contagion rule, it is easier (smaller effective spreading rate) for the information to spread to a positive fraction of the agents as a result of the friendship paradox. Hence, observing random friends of random neighbors makes it easier for the information to spread instead of dying away (in unbiased-degree networks). This shows how friendship paradox can affect information diffusion over a network under monophilic contagion.

Remark 2. If we interpret an individual’s second-hop connections as weak-ties, then Theorem 6 and Corollary 1 can be interpreted as results showing the importance of weak-ties in information diffusion (in the context of a SIS model and an unbiased-degree network). See the seminal works in [78, 28] for the definitions and importance of weak-ties in the sociology context.

The ordering $\lambda_Z^* \leq \lambda_X^*$ of the critical thresholds in Corollary 1 holds irrespective of any other network property. However, the magnitude of the difference of the critical thresholds $\lambda_X^* - \lambda_Z^*$ depends on the neighbor-degree correlation (assortativity) coefficient defined as,

$$r_{kk} = \frac{1}{\sigma_q^2} \sum_{k,k'} kk' (e(k,k') - q(k)q(k')) \tag{19}$$

using the notation defined in Footnote 7. To intuitively understand this, consider a star graph that has a negative assortativity coefficient (as all low degree nodes are connected to the only high degree node). Therefore, a randomly chosen node X from the star graph has a much smaller expected degree $\mathbb{E}\{d(X)\}$ than the expected degree $\mathbb{E}\{d(Z)\}$ of a random friend Z of the random node X compared to the case where the network has a positive assortativity coefficient. This phenomenon is further illustrated in Fig. 2a using three networks with the same degree distribution but different assortativity coefficients obtained using Newman’s edge rewiring procedure [71].

Consider the stationary fraction of the infected nodes

$$\rho = \sum_k P(k)x(k) \tag{20}$$

where $P(k)$ is the degree distribution and $x(k), k = 1, \dots, D$ are the stationary states of the mean-field dynamics in (6). Fig. 2b illustrates how the stationary fraction of the infected nodes varies with the effective spreading rate λ for case 1 and 2, showing the difference between the two cases and the effect of assortativity.

4 Collective Dynamics of SIS-Model and Reactive Networks

So far in Sec. 2 and Sec. 3, the underlying social network on which the information diffusion takes place was treated as a deterministic graph and, the mean-field dynamics equation (6) was used to approximate the SIS-model. This section explores the more general case where the underlying social network also randomly evolve at each time step n (of the SIS-model) in a manner that depends on the population state \bar{x}_n . Our aim is to obtain a tractable model that represents the collective dynamics of the SIS-model and the evolving graph process. As explained in Sec. 1 with examples, the motivation for this problem comes from the real world networks that evolves depending on the state of information diffusion on them. In order to state the main result, we first define a reactive network and state our assumptions.

Definition 1 (Reactive Network). A reactive network is a Markovian graph process $\{G_n\}_{n \geq 0}$ with a state space $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_N\}$ consisting of N graphs and transition probabilities $P_{\bar{x}_n}$ parameterized by the population state \bar{x}_n i.e. $G_{n+1} \sim P_{\bar{x}_n}(\cdot | G_n)$.

In Definition 1, the parameterization of the transition probabilities by the population state \bar{x}_n represents the (functional) dependency of the graph process on the current state of the SIS information diffusion process. The term *reactive network* denotes this functional dependency of the graph evolution on the population state. We assume the following two conditions on the reactive graph process (Definition 1).

Assumption 1 Each graph $\mathcal{G}_i \in \mathcal{G}, i = 1, \dots, N$ has the same number of nodes and the same degree distribution $P(k)$ but different conditional degree distributions $P_{\mathcal{G}_1}(k|k'), \dots, P_{\mathcal{G}_N}(k|k')$.

Assumption 2 The transition probability matrix $P_{\bar{x}_n}$ of the reactive network $\{G_n\}_{n \geq 0}$ (Definition 1) is irreducible and aperiodic with a unique stationary distribution $\pi_{\bar{x}_n}$ for all values of the population state \bar{x}_n .

The first assumption imposes the constraint that each graph in the state space has the same degree distribution $P(k)$ but different conditional degree distributions. Recall (from Footnote 7) that the conditional degree distribution $P_{\mathcal{G}_i}(k|k')$ is the probability that a node (in graph \mathcal{G}_i) with degree k' is connected to a node with degree k . Assumption 1 implies that the state space consists of networks which are different to each other in terms of the higher order properties such as assortativity. Hence, the reactive network (Definition 1) under Assumption 1 represents for example, a network which performs edge re-wiring [71] to change the assortativity depending on the state of a product spreading on it. Under Assumption 1, the number of nodes $M(k)$ with degree k will remain the same at each time instant n and hence, the new population state at each time instant can still be expressed as the old population state plus an update term as in Theorem 2. Assumption 2 is standard in Markov chains and it ensures the convergence to a unique stationary distribution.

The main result of this section is the following (proof is in [70]).

Theorem 7 (Collective Dynamics of SIS-model and Reactive Network). Consider a reactive network $\{G_n\}_{n \geq 0}$ (Definition 1) with state space \mathcal{G} and transition probabilities $P_{\bar{x}_n}(\cdot | G_n)$ (parameterized by the population state \bar{x}_n) satisfying the Assumptions 1 and 2. Let the k^{th} element of the vector $H(x_n, G_n)$ be

$$H_k(x_n, G_n) = (1 - x_n(k)) \frac{vk\theta_n^Z}{D} - x_n(k)\delta \quad \text{where,} \tag{21}$$

$$\theta_n^Z = \sum_k \left(\sum_{k'} P(k') P_{G_n}(k|k') \right) x_n(k). \tag{22}$$

Further, assume that $H(x, \mathcal{G}_i)$ is Lipschitz continuous in x for all $\mathcal{G}_i \in \mathcal{G}$. Then, the sequence of the population state vectors $\{\bar{x}_n\}_{n \geq 0}$ generated by the SIS model under monophilic contagion over the reactive network converges weakly to the trajectory of the deterministic differential equation

$$\frac{dx}{dt} = \mathbb{E}_{G \sim \pi_x} \{H(x, G)\} \quad (\text{ODE}) \tag{23}$$

$$P'_x \pi_x = \pi_x. \quad (\text{algebraic constraint}) \tag{24}$$

Theorem 7 asserts that the dynamics of the population state of the SIS diffusion (under monophilic contagion) on a reactive network can be approximated by an ODE (23) with an algebraic constraint (24). The core idea behind this result (and the proof that leads to it) can also be understood as follows in order to gain some intuition. Due to the Assumption 1, the mean-field dynamics

$$x_{n+1} = x_n + \frac{1}{M} H(x_n, G_n) \tag{25}$$

can be used to model the evolution of the population state of the SIS process over network despite the fact that it is evolving. Then, as the number of nodes M becomes large (i.e. the scaling factor $\frac{1}{M}$ goes to zero), the sequence $\{x_n\}_{n \geq 0}$ evolves on a slow time scale compared to the reactive network $\{G_n\}_{n \geq 0}$. In other words, it will be a system where $\{x_n\}_{n \geq 0}$ evolves on a slow time scale (due to the large M) and $\{G_n\}_{n \geq 0}$ evolves on a fast time scale. Stochastic averaging theory results (used in the proof) for such two time scale problems state that, the fast dynamics of the reactive network $\{G_n\}_{n \geq 0}$ can be approximated by their average on the slow time scale of the population state $\{x_n\}_{n \geq 0}$. In other words, $H(x_n, G_n)$ can be replaced by $\mathbb{E}_{G \sim \pi_{x_n}} \{H(x_n, G)\}$ which is the average of the update term with respect to the stationary distribution π_{x_n} of the Markov chain and thus yielding the ODE (23). The algebraic constraint (24) follows from the fact that π_x is the eigenvector with unit eigenvalue of transpose of the parameterized transition probability matrix P_x .

From a statistical modeling perspective, Theorem 7 provides a useful means of approximating the complex dynamics of two interdependent stochastic processes (information diffusion process and the stochastic graph process) by an ODE (23) whose trajectory $x(t)$ at each time instant $t > 0$ is constrained by the algebraic condition (24). Further, having an algebraic constraint restricts the number of possible

sample paths of the population state vector $\{\bar{x}_n\}_{n \geq 0}$. Hence, from a statistical inference/filtering perspective, this makes estimation/prediction of the population state easier. For example, the algebraic condition can be used in Bayesian filtering algorithms (such as the one discussed in Sec. 6) for the population state to obtain more accurate results.

5 Friendship Paradox based Polling for Networks

In Sec. 3 and Sec. 4, the effects of the friendship paradox on SIS-model and the effects of state dependent network evolutions were discussed. In contrast, this section deals with polling: estimating the fraction of infected (state denoted by 1) individuals

$$\bar{\rho}_n = \sum_k P(k)x_n(k) \quad (26)$$

at a given time instant n , using the responses (to some query) of b sampled individuals from the network. It is assumed that the information diffusion is slow and the states of nodes remain unchanged during the estimation task. In other words, we assume that the information diffusion takes place on a slower time scale compared to the time it take to estimate $\bar{\rho}_n$.

Notation: Since we consider the case of estimating the fraction of infected nodes at a given time instant n , we omit the subscript denoting time and use $\bar{\rho}$ and $s^{(\cdot)}$ to denote the infected fraction of nodes and state of nodes respectively (at the given time instant n) in this section.

Motivation and Related Work: Recall that in intent polling⁸, a set S of nodes are obtained by uniform sampling with replacement and then, the average of the labels $s^{(u)}$ of nodes $u \in S$

$$I^b = \frac{\sum_{u \in S} s^{(u)}}{|S|}, \quad (27)$$

is used as the estimate (called intent polling estimate henceforth) of the fraction $\bar{\rho}$ of infected individuals. The main limitation of intent polling is that the sample size needed to achieve an ε - additive error is $O(\frac{1}{\varepsilon^2})$ [13]. The algorithms presented in this section are motivated by two recently proposed methods, namely “expectation polling” [81] and “social sampling” [13], that attempt to overcome this limitation in intent polling. Firstly, in expectation polling [81], each sampled individual is asked to provide an estimate about the state held by the majority of the individuals in the network (e.g. asking “What do you think the state of the majority is?”). Then, each sampled individual will look at his/her neighbors and provide an answer (1 or 0)

⁸ This method is called intent polling because, in the case of predicting the outcome of an election, this is equivalent to asking the voting intention of sampled individuals i.e. asking “Who are you going to vote for in the upcoming election?” [81].

based on the state held by the majority of them. This method is more efficient (in terms of sample size) compared to the intent polling method since each sample now provides the putative response of a neighborhood^{9,10}. Secondly, in social sampling [13], the response of each sampled individual is a function of the states, degrees and the sampling probabilities of his/her neighbors. [13] provides several unbiased estimators for the fraction \bar{p} using this method and, establishes bounds for their variances. The main limitation of social sampling method (compared to friendship paradox based algorithms in Sec. 5) is that it requires the sampled individuals to know a significant amount of information about their neighbors (apart from just their labels), the graph and the sampling process (employed by the pollster). Therefore, a practical implementation of social sampling in the setting of estimating the fraction of infected individuals at a given time instant is not practically feasible. These facts motivate the polling method called *neighborhood expectation polling (NEP)* [69] which we present next.

In NEP, a set $S \subset V$ of individuals from the social network $G = (V, E)$ are selected and asked,

“What is your estimate of the fraction of people with label 1?”.

When trying to estimate an unknown quantity about the world, any individual naturally looks at his/her neighbors. Therefore, each sampled individual $s \in S$ would provide the fraction of their neighbors $\mathcal{N}(s)$, with label 1. In other words, the response of the individual $s \in S$ for the NEP query would be,

$$q(s) = \frac{|\{u \in \mathcal{N}(s) : s^{(u)} = 1\}|}{|\mathcal{N}(s)|}. \tag{28}$$

Then, the average of all the responses $\frac{\sum_{s \in S} q(s)}{|S|}$ is used as the NEP estimate of the fraction \bar{p} .

Why call it NEP? The term neighborhood expectation polling is derived, from the fact that the response $q(v)$ of each sampled individual $v \in S$ is the expected label value among her neighbors i.e. $q(v) = \mathbb{E}\{s^{(U)}\}$ where, U is a random neighbor of the sampled individual $v \in S$.

Why (not) use NEP? NEP is substantially different to classical intent polling where, each sampled individual is asked “What is your label?”. In intent polling, the response of each sampled individual $v \in S$ is his/her label $s^{(v)}$. In contrast, in NEP, the response $q(v)$ of each sampled individual $v \in S$ is a function of his/her neighborhood (defined by the underlying graph G) as well as the labels of his/her neighbors.

⁹ Intent polling and expectation polling have been considered intensively in literature, mostly in the context of forecasting elections and, it is generally accepted that expectation polling is more efficient compared to intent polling [27, 26, 66, 67, 62].

¹⁰ [47, 45] discuss how expectation polling can give rise to misinformation propagation in social learning and, propose Bayesian filtering methods to eliminate the misinformation propagation.

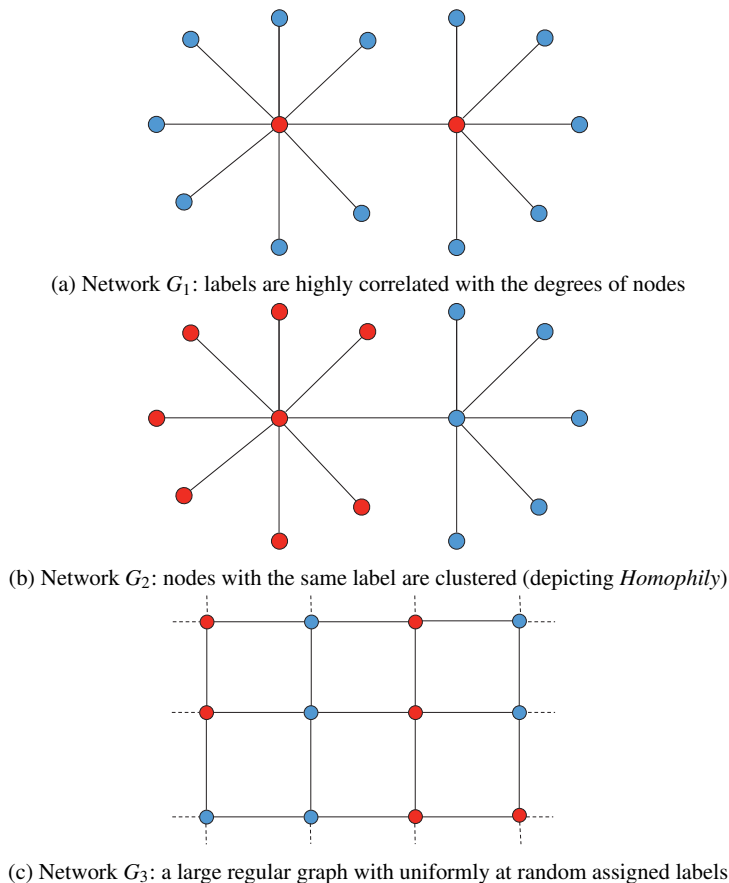


Fig. 3: Consider the case of uniformly sampling nodes and obtaining responses $q(s)$ of sampled nodes $s \in S$ about the fraction of red (i.e. label 1) nodes in the network. In graph G_1 of Fig. 3a, most nodes have their only neighbor to be of color red even though most of the nodes in the network are of color blue. Hence, uniformly sampling nodes for NEP in this case would result in a highly biased estimate. In graph G_2 of Fig. 3b, approximately half the nodes have only a red neighbor and, rest of the nodes have only a blue neighbor. Hence, uniformly sampling nodes for NEP in this case would result in an estimate with a large variance. In graph G_3 of Fig. 3c, average of the NEP responses $q(v)$ of nodes is approximately equal to the fraction of nodes with red labels. Further, $q(v)$ does not vary largely among nodes. Hence, uniformly sampling nodes for NEP in this case would result in an accurate estimate. Similar examples can also be found in [13]. The figure highlights the importance of exploiting network structure and node labels when sampling nodes for NEP.

Therefore, depending on the graph G , function $s^{(\cdot)}$ and the method of obtaining the samples S , NEP might produce either,

- I an estimate with a larger MSE compared to intent polling (e.g. networks in Fig. 3a and Fig. 3b shows when uniform sampling of individuals for NEP might not work), or,
- II an estimate with a smaller MSE compared to intent polling (e.g. network in Fig. 3c shows when uniform sampling of individuals for NEP might work)

These two possible outcomes highlight the importance of using the available information about the graph G and the function $s^{(\cdot)}$ (which represent the states at the time of the estimation), when selecting the set S of individuals in NEP. This lead us to the friendship paradox based NEP algorithms.

5.1 NEP Algorithms Based on Friendship Paradox

In this subsection, we consider randomized methods for selecting individuals for NEP based on the concept of friendship paradox explained in Sec. 2.3.

5.1.1 Case 1 - Sampling friends using random walks

In this section, we consider the case where the graph $G = (V, E)$ is not known initially, but sequential exploration of the graph is possible using multiple random walks over the nodes of the graph. A motivating example is a massive online social network where the fraction of user profiles indicating infection needs to be estimated (e.g. profiles mentioning symptoms of a disease). Web-crawling (using random walks) approaches are widely used to obtain samples from such massive online social networks without requiring the global knowledge of the full network graph [58, 22, 79, 23, 64].

Algorithm 1: NEP with Random Walk Based Sampling

Input: b number of samples $\{v_1, v_2, \dots, v_b\} \subset V$.

Output: Estimate T_{RW}^b of the of the fraction $\bar{\rho}$ of nodes with label 1.

1. Initialize b random walks on the social network starting from v_1, v_2, \dots, v_b .
2. Run each random walk for a N steps and then collect sample $S = \{s_1, \dots, s_b\}$ where, $s_i \in V$ is collected from i^{th} random walk.
3. Query each $s \in S$ to obtain $q(s)$ and, compute the estimate

$$T_{RW}^b = \frac{\sum_{s \in S} q(s)}{b}$$

of the fraction $\bar{\rho}$ of nodes with label 1.

Algorithm 1 was proposed in [69] for estimating the fraction of infected individuals $\bar{\rho}$ defined in (26). The intuition behind Algorithm 1 stems from the fact that the stationary distribution of a random walk on an undirected graph is uniform over the set of neighbors [1]. Therefore, Algorithm 1 obtains a set S of b neighbors independently (for sufficiently large N) from the graph $G = (V, E)$ in step 2. Then, the response $q(s)$ of each sampled individual $s \in S$ for the NEP query is used to compute the estimate T_{RW}^b in step 3. According to the friendship paradox (Theorem 1), using uniformly sampled neighbors is equivalent to using more nodes due to the fact that random neighbors have more neighbors than random nodes on average. Hence, it is intuitive that the performance of this method should have a smaller MSE compared to the method of NEP with uniformly sampled nodes and intent polling method. In Sec. 5.2, we verify this claim theoretically and explore the conditions on the state function $s^{(\cdot)}$ and the properties of the graph G for the estimator T_{RW}^b to be more accurate compared to the intent polling method.

5.1.2 Case 2 - Sampling a Random Friend of a Random Individual

Here we assume that the graph $G = (V, E)$ is not known and it is not possible to crawl the graph (using random walks). It is further assumed that a set of uniform samples $S = \{s_1, \dots, s_b\}$ from the set of nodes V can be obtained and, each sampled individual $s_i \in S$ has the ability to answer the question "What is your (random) friend's estimate of the fraction of individuals with label 1?".

A motivating example for case 2 is the situation where random individuals are requested to answer survey questions for an incentive. In most such cases, the pollster does not have any information about the structural connectivity of the queried individuals and, will only be able to obtain their answer for a question.

For this case, Algorithm 2 was proposed in [69] to obtain an estimate of the fraction $\bar{\rho}$ of individuals with label 1 defined in (26).

Algorithm 2: NEP using Friends of Uniformly Sampled Nodes

Input: b number of uniform samples $S = \{s_1, s_2, \dots, s_b\} \subset V$.

Output: Estimate T_{FN}^b of the of the fraction $\bar{\rho}$ of the individuals with label 1.

1. Ask each $s_i \in S$ to provide $q(u_i)$ for some randomly chosen neighbor $u_i \in \mathcal{N}(s_i)$.
2. Compute the estimate,

$$T_{FN}^b = \frac{\sum_{i=1}^b q(u_i)}{b}$$

of the fraction $\bar{\rho}$ of the individuals with label 1.

In Algorithm 2, each uniformly sampled individual is asked the question "What is your (random) friend's estimate of the fraction of individuals with state 1?". Then, each sampled node $s_i \in S$ would provide $q(u_i)$ for some randomly chosen

$u_i \in \mathcal{N}(s_i)$. The theoretical reasoning behind this method comes from Theorem 4 in Section 2 which states that, a random friend of a randomly chosen individual has more friends than a randomly chosen individual on average¹¹. Therefore, intuitively this method should result in a smaller MSE compared to the method of NEP with uniformly sampled nodes and intent polling method.

5.2 Analysis of the Estimates Obtained via Algorithms 1 and 2

Algorithm 1 and Algorithm 2 presented in Sec. 5.1 query random friends and random friends of random nodes (denoted by Y, Z in Theorem 1 and Theorem 4) respectively, exploiting the friendship paradox.

In this context, the aim of this subsection is to present the following results (proof can be found in [69]):

1. Theorem 8 motivates using friendship paradox based NEP algorithms (as opposed to NEP with uniformly sampled nodes)
2. Theorem 9 relates bias and variance of the estimate T_{RW}^b obtained using Algorithm 1 to the properties of the network. Then, Corollary 2 gives sufficient conditions for T_{RW}^b to be an unbiased estimate with a smaller mean squared error (MSE) compared to intent polling method where, MSE of an estimate T of a parameter $\bar{\rho}$ is defined as

$$\text{MSE}\{T\} = \mathbb{E}\{(T - \bar{\rho})^2\} \tag{29}$$

$$= \text{Bias}\{T\}^2 + \text{Var}\{T\} \tag{30}$$

3. Theorem 10 motivates the use of friendship paradox based sampling methods when the sampling budget b is small

Theorem 8. *If the label $f(v)$ of each node $v \in V$ is independently and identically distributed then,*

$$\text{MSE}\{T_{FN}^b\} \leq \text{MSE}\{T_{UN}^b\} \tag{31}$$

$$\text{MSE}\{T_{RW}^b\} \leq \text{MSE}\{T_{UN}^b\} \tag{32}$$

where, MSE denotes mean square error defined in (30), T_{UN}^b is the NEP estimate with b uniformly sampled nodes and, T_{RW}^b, T_{FN}^b are the estimates obtained using Algorithm 1 and Algorithm 2 respectively.

Theorem 8 shows that friendship paradox based sampling always has a smaller mean squared error when the node labels are independently and identically distributed

¹¹ It should be noted that this does not follow from the original version of friendship paradox (Theorem 1) since the random friend is not a uniformly chosen neighbor from the set of all $2|E|$ neighbors. Instead, the response now comes from a random neighbor conditioned to be a friend of the sampled node.

(iid). This motivates the use of friendship paradox based NEP methods (Algorithm 1 and Algorithm 2) instead of uniform sampling based NEP. In the subsequent results, we show that the superiority of friendship paradox based NEP algorithms over the widely used intent polling method holds for conditions less stringent than the iid assumption.

Next, we formally quantify the bias $\text{Bias}(T_{RW}^b)$ and the variance $\text{Var}(T_{RW}^b)$ of the estimator T_{RW}^b obtained via Algorithm 1 as the random walk length N goes to infinity and then, compare it with the widely used intent polling method.

Theorem 9. *Let X be a random node and (U, Y) be a random link sampled from a connected graph. Then, as N tends to infinity, the bias $\text{Bias}(T_{RW}^b)$ and the variance $\text{Var}(T_{RW}^b)$ of the estimate T_{RW}^b , obtained via Algorithm 1 are given by,*

$$\text{Bias}(T_{RW}^b) = \mathbb{E}\{s^{(Y)}\} - \mathbb{E}\{s^{(X)}\} \tag{33}$$

$$= \frac{\text{Cov}\{s^{(X)}, d(X)\}}{\mathbb{E}\{d(X)\}} \tag{34}$$

$$\text{Var}\{T_{RW}^b\} = \frac{1}{b} \text{Cov}\{s^{(Y)}, q(U)\}. \tag{35}$$

Theorem 9 provides insights into the properties of the networks for which, NEP based Algorithm 2 provides a better estimate compared to the intent polling method. Eq. (33) of Theorem 9 shows that, the bias of the estimate T_{RW}^b is the difference between the expected label value at a random friend, Y and the expected value at a random individual, X . Further, (34) shows that it is proportional to the covariance between the degree $d(X)$ and the state $s^{(X)}$ of a randomly chosen node X . An immediate consequence of this result is the following corollary, which gives a sufficient condition for the estimate T_{RW}^b to be unbiased and, also have a smaller variance (and therefore, a smaller MSE) compared to intent polling.

Corollary 2. *If the label $s^{(X)}$ and the degree $d(X)$ are uncorrelated and the graph is connected, the following statements hold as N tends to infinity:*

1. *The estimate T_{RW}^b , obtained via Algorithm 1 is unbiased for $\bar{\rho}$ i.e.*

$$\mathbb{E}\{T_{RW}^b\} = \bar{\rho} \tag{36}$$

2. *The estimate T_{RW}^b , obtained via Algorithm 1 is more efficient compared to intent polling estimate I in (27) i.e.*

$$\text{MSE}\{T_{RW}^b\} \leq \text{MSE}\{I^b\} \tag{37}$$

where, MSE denotes mean square error defined in (30).

Theorem 9 also shows that the variance of the estimate T_{RW}^b is the covariance of the state $s^{(Y)}$ of a random friend Y and the response $q(U)$ of her random friend U .

The following result gives sufficient conditions for T_{RW}^b to be a more efficient (in an MSE sense) estimator compared to intent polling method (even in the presence of bias) when the sampling budget $b = 1$.

Theorem 10. Assume that the graph is connected and the sampling budget $b = 1$. Then, as N tends to infinity, the estimate T_{RW}^b has a smaller MSE compared to the intent polling estimate I , defined in (27), if

$$\mathbb{E}\{d(X)|s^{(X)} = 1\} \leq \mathbb{E}\{d(X)|s^{(X)} = 0\} \text{ and } \bar{\rho} \leq 0.5 \tag{38}$$

or

$$\mathbb{E}\{d(X)|s^{(X)}\} \geq \mathbb{E}\{d(X)|s^{(X)} = 0\} \text{ and } \bar{\rho} \geq 0.5. \tag{39}$$

Theorem 10 shows that, if the expected degree of an individual with state 1 is larger (smaller) compared to the expected degree of an individual with opinion 0 and, the expected state in the network is above (below) half then, MSE of the estimate T_{RW}^b is smaller than intent polling estimate I in (27) when the pollster can query only one individual. This helps the pollster to incorporate prior knowledge about the current state of the diffusion and the structure of the network to decide whether its suitable to use NEP based Algorithm 1 (over the intent polling method).

5.3 Numerical Examples

In this section, results (based on [69]) illustrating the performance of Algorithms 1 and 2 are provided. The aim of these numerical simulations is to evaluate the dependence of the accuracy (MSE) of the estimate of $\bar{\rho}$ on the following three properties related to the network and the state of the information diffusion:

1. **Degree distribution** $P(k)$ (which is the probability that a randomly chosen node has k neighbors).
2. **Neighbor Degree correlation (assortativity) coefficient** r_{kk} defined in (19)
3. **Degree-label correlation coefficient**

$$p_{ks} = \frac{1}{\sigma_k \sigma_s} \sum_k k \left(\mathbb{P}(s^{(X)} = 1, d(X) = k) - \mathbb{P}(s^X = 1)P(k) \right) \tag{40}$$

where, σ_k, σ_s are the standard deviations of the degree distribution $P(k)$ and the state (label) distribution respectively and, $P(s, k)$ is the joint distribution of the states and degrees of nodes.

A detailed discussion about these metrics and their effects can be found in [57].

Simulation Setup: [69] evaluated Algorithms 1, 2 on networks with 5000 nodes obtained using two models: configuration mode [65] and Erdős-Rényi ($G(n,p)$) model [72] that result in a power-law degree distribution and a Poisson degree distribution respectively. Further, the assortativity coefficient and degree-label correlation coefficient of the networks obtained using these models were modified to different values (while preserving the degree distribution) by using Newman’s edge rewiring

procedure [71] and label swapping procedure [57]. Then, the MSE of the algorithms (on the obtained networks) were estimated by averaging the squared error of the estimates over 500 independent iterations for each value of the sampling budget b from 1 to 50. The resulting MSE values for the network obtained using configuration model (with a power-law degree distribution) are shown in Fig. 4 and Fig. 5 for power-law coefficient values $\alpha = 2.1$ and $\alpha = 2.4$ respectively. Similarly, resulting MSE values for the network obtained from the Erdős-Rényi model (with a Poisson degree distribution) are shown in Fig. 6.

In the case of Erdős-Rényi graphs, only the assortativity coefficient $r_{kk} = 0$ is considered as it cannot be changed significantly due to the homogeneity in the degree distribution (see [69] for more details on the simulation procedure).

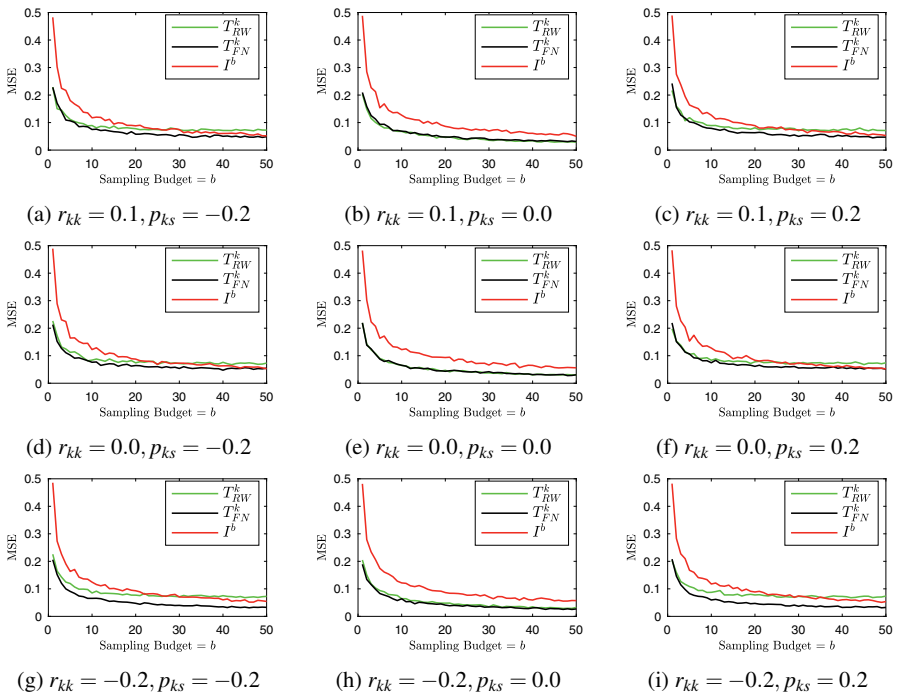


Fig. 4: MSE of the estimates obtained using Algorithm 1 (T_{RW}^b), Algorithm 2 (T_{FN}^b) and intent polling method (I^b) versus the sampling budget b , for a power-law graph with parameter $\alpha = 2.1$ with different values of assortativity coefficient r_{kk} and degree-label correlation coefficient p_{ks} . This figure shows that, for power-law networks, the proposed friendship paradox based NEP methods have smaller mean squared error compared to classical intent polling method under general conditions.

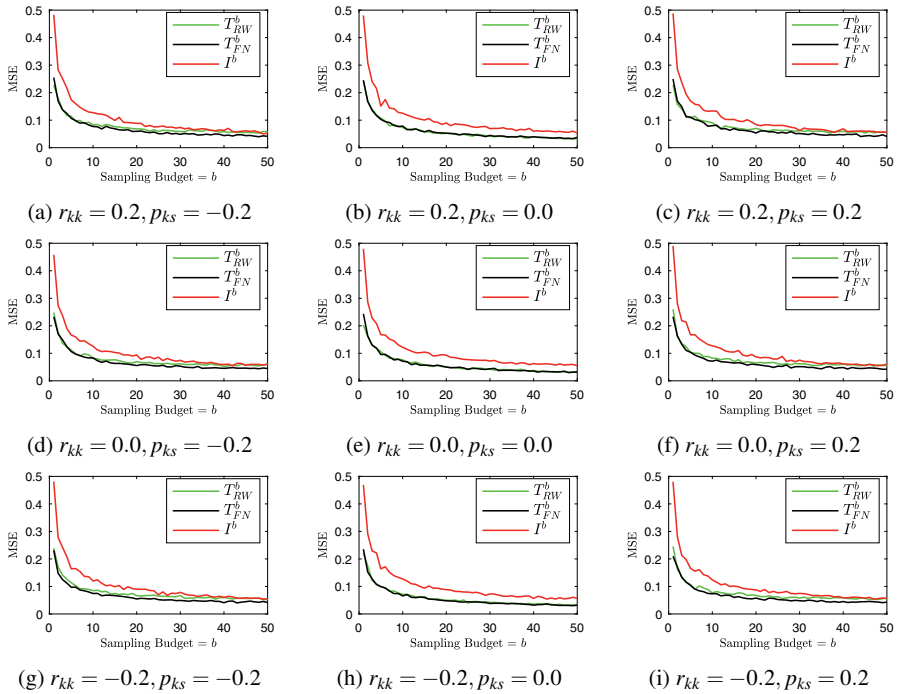


Fig. 5: MSE of the estimates obtained using Algorithm 1 (T_{RW}^b), Algorithm 2 (T_{FN}^b) and intent polling method (I^b) versus the sampling budget b , for a power-law graph with parameter $\alpha = 2.4$ with different values of the assortativity coefficient r_{kk} and degree-label correlation coefficient p_{ks} . This figure shows that, for power-law networks, the proposed friendship paradox based NEP methods have smaller mean squared error compared to classical intent polling method under general conditions.

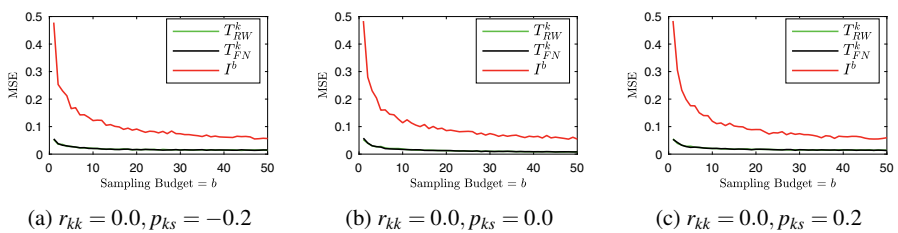


Fig. 6: MSE of the estimates obtained using Algorithm 1 (T_{RW}^b), Algorithm 2 (T_{FN}^b) and intent polling method (I^b) versus the sampling budget b , for a Erdős-Rényi graph with parameter average degree 50 with assortativity coefficient $r_{kk} = 0$ and different values of degree-label correlation coefficient p_{ks} . This figure shows that, for ER graphs, the proposed friendship paradox based NEP method as well as the greedy deterministic sample selection method result in better performance compared to the intent polling method.

5.4 Discussion of the Results

In this subsection, the main findings of the numerical simulation studies (under the setup described in Sec. 5.3) and how they relate to the theoretical results are discussed. Further, how these findings can be useful to identify the best possible algorithm (out of Algorithms 1, Algorithm 2 and the alternative intent polling method) depending on the context is discussed.

5.4.1 Power-law Graphs

Intent Polling vs. Friendship Paradox Based Polling: In the numerical results (shown in Fig. 4, Fig. 5 and Fig. 6), Algorithm 1 and Algorithm 2 outperform the intent polling method (in terms of the MSE) when the degree-label correlation $p_{ks} = 0$. This outcome agrees with Corollary 2. When p_{ks} is non-zero, Algorithm 2 has a smaller MSE than intent polling in the examples we considered while Algorithm 1 has a smaller MSE than intent polling for small sampling budgets ($b \leq 30$). Hence, our numerical results indicate that friendship paradox based polling methods have a smaller MSE in contexts where the sampling budget b is smaller compared to the number of nodes in the network (which is set to 5000 in our simulation).

Effect of the Heavy-Tails: Comparing Fig. 4 with Fig. 5 shows that the MSE of the Algorithm 1 and Algorithm 2 are smaller in the network with power-law coefficient $\alpha = 2.1$ compared to the network with power-law coefficient $\alpha = 2.4$ that we considered in the simulations. The difference of the MSE in the two cases ($\alpha = 2.1$ and $\alpha = 2.4$) is more visible for Algorithm 2 compared to Algorithm 1. Hence, this observation suggests that friendship paradox based algorithms are more suitable to contexts where it is known that the underlying network has a heavy tailed degree distribution.

Effect of the Assortativity of the Network: Many different joint degree distributions $e(k, k')$ can yield the same neighbor degree distribution $q(k)$ (which is the marginal distribution of $e(k, k')$ as defined in Footnote 7). This marginal distribution $q(k)$ does not capture the joint variation of the degrees a random pair of neighbors. In Algorithm 1 (which samples neighbors uniformly), the degree distribution of the samples (i.e. queried nodes) is the neighbor degree distribution $q(k)$. Hence, the performance is not affected by the assortativity coefficient r_{kk} , which captures the joint variation of the degrees of a random pair of neighbors. This is apparent in Fig. 5 where, each column (corresponding to different r_{kk} values) has approximately same MSE for Algorithm 1. However, it can be seen that, the MSE of Algorithm 2 (that samples random friends Z of random nodes) increases with assortativity r_{kk} due to the fact that the distribution of degree $d(Z)$ of a random friend Z of a random node is a function of the joint degree distribution. In order to highlight this further, Fig. 7 illustrates the effect of the neighbor degree correlation r_{kk} on the distribution of $d(Z)$ (and the invariance of the distribution of $d(Y)$ to r_{kk}). This numerical result

indicates that, if it is apriori known that the network is disassortative ($r_{kk} < 0$), the Algorithm 2 is a more suitable choice for polling (compared to Algorithm 1).

When to use friendship paradox based NEP? Both theoretical (Theorem 10) as well as numerical results (Fig. 5, Fig. 6) show that friendship paradox based NEP methods outperform classical intent polling method when the sampling budget is small compared to the size of the network (which is the case in many applications related to polling). Further, the absence of degree-label correlation and the presence of assortativity improves the performance of friendship paradox based polling methods. These analytical results and numerical simulation studies provide the pollster the ability to decide which algorithm to be deployed using the available information about the network and the sampling budget.

5.4.2 Erdős-Rényi Graphs

Erdős-Rényi ($G(n, p)$) model starts with n vertices and then connects any two vertices with probability p resulting in an average degree of $(n - 1)p$. From the Fig. 6, it can be seen that both Algorithm 1 and Algorithm 2 yield a smaller MSE than the intent polling method for the Erdős-Rényi network that we considered (which has $p = 0.01$ and $n = 5000$). Further, Algorithm 1 and Algorithm 2 both have approximately equal MSE in the case of the Erdős-Rényi network we considered. This is a result of the fact that distributions of the degree $d(Y)$ of a random neighbor Y and the distribution of the degree $d(Z)$ of a random neighbor Z of a random node are approximately equal when the neighbor degree correlation is zero.

6 Non-Linear Bayesian Filtering for Estimating Population State

Sec. 5 discussed algorithms to estimate the fraction of infected (state 1) individuals in the case of slow diffusion dynamics where node states can be treated as fixed for the purpose of estimation. However, treating the node states as fixed is not realistic when the diffusion takes place on the same time scale as the one on which individuals are polled (i.e. measurements are collected). Further, the non-linear dynamics (6) of the information diffusion rules out the possibility of applying standard Bayesian filtering methods such as Kalman filter to recursively update the population state estimate with new measurements [45]. This section presents the non-linear Bayesian filtering method proposed in [46] which computes an optimal (in a mean-squared error sense) estimate of the population state with each new measurement. The method consists of two steps at each time instant:

1. sample nodes from the network to obtain a noisy estimate of the population state
2. use the noisy estimate to compute the posterior distribution and then compute the new conditional mean of the estimate.

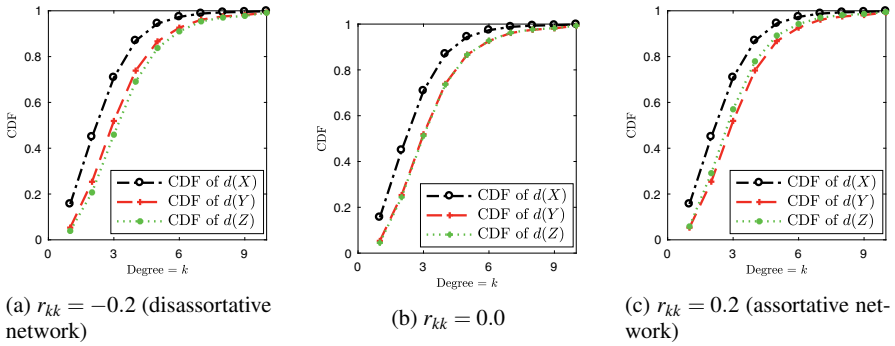


Fig. 7: The cumulative distribution functions (CDF) of the degrees $d(X)$, $d(Y)$, $d(Z)$ of a random node (X), a random friend (Y) and a random friend (Z) of a random node respectively, for three graphs with the same degree distribution (power-law distribution with a coefficient $\alpha = 2.4$) but different neighbor-degree correlation coefficients r_{kk} , generated using the Newman's edge rewiring procedure. This illustrates that $\mathbb{E}\{d(Z)\} \geq \mathbb{E}\{d(Y)\}$ for $r_{kk} \leq 0$ (Fig. 7a) and vice-versa. Further, this figure also shows how the distributions of $d(X)$, $d(Y)$ remain invariant to the changes in the joint degree distribution $e(k, k')$ that preserve the degree distribution $P(k)$.

6.1 Sampling

We first consider sampling the social network $G = (V, E)$ described in Sec. 2.1 for the purpose collecting measurements to estimate the population state x_n at time n . We assume that the degree distribution $P(\cdot)$ of the underlying network is known. Note that friendship paradox based NEP algorithms (presented in Sec. 5 for estimating the scalar valued fraction of infected nodes) can be easily extended for obtaining such (noisy) measurements of population state vector x_n . For example, at each time instant n , a random friend can be sampled and asked to provide an estimate of the population state x_n based on her neighbors. Apart from such extensions of friendship paradox based NEP methods, we discuss two other widely used methods for sampling large networks for the purpose of obtaining an empirical estimate of x_n .

6.1.1 Uniform Sampling

At each time n , $v(k)$ individuals are sampled¹² independently and uniformly from the population $M(k)$ comprising of agents with degree k . Thus, a uniformly distributed independent sequence of nodes, denoted by $\{m_l, l \in \{1, 2, \dots, v(k)\}\}$, is generated from the population $M(k)$. From these independent samples, the empirical infected population state $\hat{x}_n(k)$ of degree k nodes at each time n is

¹² For large population where $M(d)$ is large, sampling with and without replacement are equivalent.

$$\hat{x}_n(k) = \frac{1}{v(k)} \sum_{l=1}^{v(k)} \mathbb{1}(s_n^{(m_l)} = 1). \tag{41}$$

At each time n , \hat{x}_n can be viewed as noisy observation of the infected population state x_n .

6.1.2 MCMC Based Respondent-Driven Sampling (RDS)

Respondent-driven sampling (RDS) was introduced by Heckathorn [29, 30] as an approach for sampling from hidden populations in social networks and has gained enormous popularity in recent years. In RDS sampling, current sample members recruit future sample members. The RDS procedure is as follows: A small number of people in the target population serve as seeds. After participating in the study, the seeds recruit other people they know through the social network in the target population. The sampling continues according to this procedure with current sample members recruiting the next wave of sample members until the desired sampling size is reached.

RDS can be viewed as a form of Markov Chain Monte Carlo (MCMC) sampling. Let $\{m_l, l \in \{1, 2, \dots, v(k)\}\}$ be the realization of an aperiodic irreducible Markov chain with state space $M(k)$ comprising of nodes of degree k . This Markov chain models the individuals of degree k that are sampled, namely, the first individual m_1 is sampled and then recruits the second individual m_2 to be sampled, who then recruits m_3 and so on. Instead of the independent sample estimator (41), an asymptotically unbiased MCMC estimate is computed as

$$\frac{\sum_{l=1}^{v(k)} \frac{\mathbb{1}(s_n^{(m_l)}=1)}{\pi(m_l)}}{\sum_{l=1}^{v(k)} \frac{1}{\pi(m_l)}} \tag{42}$$

where $\pi(m)$, $m \in M(k)$, denotes the stationary distribution of the Markov chain m_l .

In RDS, the transition matrix and, hence, the stationary distribution π in the estimate (42) is specified as follows: Assume that edges between any two nodes i and j have symmetric weights W_{ij} (i.e., $W_{ij} = W_{ji}$). Node i recruits node j with transition probability $W_{ij} / \sum_j W_{ij}$. Then, it can be easily seen that the stationary distribution is $\pi(i) = \sum_{j \in V} W_{ij} / \sum_{i \in V, j \in V} W_{ij}$. Using this stationary distribution along with (42) yields the RDS algorithm. Since a Markov chain over a non-bipartite connected undirected network is aperiodic, the initial seed for RDS can be picked arbitrarily, and the estimate (42) is asymptotically unbiased [24].

The key outcome of this subsection is that by the central limit theorem (for an irreducible aperiodic finite state Markov chain), the estimate of the probability that a node is infected in a large population (given its degree) is asymptotically Gaussian. For a sufficiently large number of samples, observation of the infected population state is approximately Gaussian, and the sample observations can be expressed as

$$y_n = Cx_n + v_n \quad (43)$$

where $v_n \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ is the observation noise with the covariance matrix \mathbf{R} and observation matrix C dependent on the sampling process and $x_n \in \mathbb{R}^D$ is the infected population state and evolves according to the polynomial dynamics (6).

6.2 Non-linear filter and PCRLB for Bayesian Tracking of Infected Populations

In Sec. 2, the mean field dynamics for the population state as a system with polynomial dynamics (6) was discussed. Linear Gaussian observations (43) can be obtained by sampling the network as outlined in Sec. 6.1. In this subsection, we consider Bayesian filtering for recursively estimating the infected population state x_n in large networks. We first describe how to express the mean field dynamics (6) in a form amenable to employing the non-linear filter described in [31].

6.2.1 Mean Field Polynomial Dynamics

Consider a D -dimensional polynomial vector $f(x) \in \mathbb{R}^D$:

$$f(x) = A_0 + A_1x + A_2xx' + A_3xxx' + \dots \quad (44)$$

where the co-coefficients A_0, A_1, \dots, A_i are dimension $1, 2, \dots, (i+1)$ tensors, respectively. Note that $A_i xx \dots x'$ is a vector with r^{th} entry given by

$$A_i xx \dots x'(r) = \sum_{j_1, j_2, j_3, \dots, j_i} A_i(r, j_1, j_2, \dots, j_i) x_{j_1} x_{j_2} \dots x_{j_i}$$

where $A_i(r, j_1, j_2, \dots, j_i)$ is the r, j_1, j_2, \dots, j_i entry of tensor A_i and x_j is the j^{th} entry of x . Because (6) has polynomial dynamics, it can be expressed in the form of (44) by constructing the tensors A_i . We refer the reader to [46] for the exact forms of these equations.

6.2.2 Optimal Filter for Polynomial Dynamics

With the mean-field dynamics (6) expressed in the form (44), we are now ready to describe the optimal filter to estimate the infected population state. Optimal Bayesian filtering refers to recursively computing the conditional density (posterior) $p(x_n | Y_n)$, for $n = 1, 2, \dots$, where Y_n denotes the observation sequence y_1, \dots, y_n . From this posterior density, the conditional mean estimate $\mathbb{E}\{x_n | Y_n\}$ can be computed by integration. (The term optimal refers to the fact that the conditional mean estimate is the minimum variance estimate). In general for nonlinear or non-

Gaussian systems, there is no finite dimensional filtering algorithm, that is, the posterior $p(x_n|Y_n)$ does not have a finite dimensional statistic. However, it is shown in [31] that for Gaussian systems with polynomial dynamics, one can devise a finite dimensional filter (based on the Kalman filter) to compute the conditional mean estimate. That is, Bayes rule can be implemented exactly (without numerical approximation) to compute the posterior, and the conditional mean can be computed from the posterior. Therefore, to estimate the infected population state using the sampled observations (43), we employ this optimal filter.

The non-linear filter prediction and update equations are given as:

Prediction step:

$$\begin{aligned}\hat{x}_n^- &= \mathbb{E}\{x_n|Y_{n-1}\} = \mathbb{E}\{f(x_{n-1})|Y_{n-1}\} \\ H_n^- &= \mathbb{E}\{(x_n - \hat{x}_n)(x_n - \hat{x}_n)'|Y_{n-1}\} \\ &= \mathbb{E}\{(f(x_{n-1}) - \mathbb{E}\{f(x_{n-1})|Y_{n-1}\} + v_{n-1}) \\ &\quad \times (f(x_{n-1}) - \mathbb{E}\{f(x_{n-1})|Y_{n-1}\} + v_{n-1})'|Y_{n-1}\} \\ &= \mathbb{E}\{f(x_{n-1})f(x_{n-1})'|Y_{n-1}\} - \mathbb{E}\{f(x_{n-1})|Y_{n-1}\} \\ &\quad \times \mathbb{E}\{f(x_{n-1})|Y_{n-1}\}' + \mathbf{Q}_{n-1}\end{aligned}\quad (45)$$

where $Y_n = \{Y_{n-1}, y_n\}$ denotes the observation process; H_n^- denotes the priori state co-variance estimate at time n ; and v_n denotes the Gaussian state noise at time n , with covariance \mathbf{Q}_n .

The filter is initialized with mean \hat{x}_0 and covariance H_0^- . The filter relies upon being able to compute the expectation $\mathbb{E}\{f(x_{n-1})f'(x_{n-1})|Y_{n-1}\}$ in terms of \hat{x}_{n-1} and H_n^- . When $f(\cdot)$ is a polynomial, $f(x_{n-1})f'(x_{n-1})'$ is a function of x_{n-1} , and the conditional expectations in (46) can be expressed only in terms of \hat{x}_{n-1} and H_n^- , permitting a closed form¹³ prediction step.

Update step:

$$\begin{aligned}\hat{x}_n &= \mathbb{E}\{x_n|Y_n\} = \hat{x}_n^- + H_n^- C' (\mathbf{R}_n + CH_n^- C')^{-1} (y_n - C\hat{x}_n^-) \\ K_n &= H_n^- C' (\mathbf{R}_n + CH_n^- C')^{-1} \\ H_n &= (I - K_n C) H_n^- (I - K_n C)' + K_n \mathbf{R}_n K_n'\end{aligned}\quad (46)$$

where \hat{x}_n denotes the conditional mean estimate of the state and H_n the associated conditional covariance at time n . C denotes the state observation matrix; \mathbf{R}_n denotes the observation noise co-variance matrix; K_n denotes the filter gain; and I denotes the identity matrix.

Since the dynamics of (6) are polynomial, the prediction and update steps of (45) and (46) can be implemented without approximation. These expressions constitute

¹³ For an explicit implementation of such a filter for a third order system with an exact priori update equation for H_n^- and \hat{x}_n^- , see [31].

the optimal non-linear filter and can be used to track the evolving infected population state.

7 Summary and Discussion

This chapter discussed in detail, three interrelated topics in information diffusion in social networks under the central theme of dynamic modeling and statistical inference of SIS models.

First, Sec. 3 showed that the effect of high degree nodes updating their states (infected or susceptible) more frequently is reflected in the update term of the deterministic mean-field dynamics model and, does not affect the critical thresholds which decide if the information diffusion process will eventually die out or spread to a non-zero fraction of individuals. Secondly, the case where two-hop neighbors are influencing the evolution of states in the SIS model was discussed. This two-hop neighbor influence, called monophilic contagion, was shown to make the SIS information diffusion easier (by lowering the critical thresholds). Sec. 4 extended the mean-field model to the case where the underlying social network randomly evolves depending on the state of the information diffusion. How the collective dynamics of such a process can be modeled by a deterministic ordinary differential equation with an algebraic constraint was discussed.

Related to the statistical inference aspect of the SIS information diffusion processes, how the state of the underlying population (induced by the SIS model) can be estimated was explored under two cases. Firstly, for the case where the dynamics of the SIS model is slower (and hence node states can be treated as fixed for estimation purpose) compared to measurement collection (polling), friendship paradox based polling algorithms to estimate the fraction of infected nodes were discussed. Such algorithms can outperform classical polling methods such as intent polling by lowering the variance of the estimate. Secondly, for the case where the dynamics of the SIS model evolve on the same time scale as the measurement collection process, a non-linear Bayesian filtering algorithm which harvests the polynomial dynamics of the SIS model was discussed. This filtering algorithm is an optimal filter which updates the state estimate with each new measurement.

Future research directions: The topics discussed in this chapter yield interesting directions for future research in information diffusion processes. The nodes were treated as non-strategic decision makers in this chapter i.e. their decisions to update the states are not strategic. The changes in the dynamics and critical thresholds yielded by the case where nodes are strategic utility maximizers is an interesting direction for future research. Also, this chapter focused on the case where the underlying network is an undirected graph. The mean-field dynamics of diffusion based on two-hop contagion, effects of friendship paradox and filtering in directed graphs (such as Twitter) is an interesting research direction to be explored. There is also

substantial motivation to evaluate SIS models using real data; some preliminary results applied to YouTube appear in [36, 49].

Topics beyond the scope of the chapter: Since the current chapter dealt with SIS contagions that are modeled using mean-field dynamics, several important topics related to the diffusion of information in social networks are beyond the scope of the current chapter. These include: social learning [44, 50, 10], data incest [48], influence maximization [41], strategic agents and game theoretic learning [21, 38] and, inferring the network structure using diffusion traces [25].

References

1. Aldous, D., Fill, J.: Reversible Markov chains and random walks on graphs (2002)
2. Altenburger, K.M., Ugander, J.: Monophily in social networks introduces similarity among friends-of-friends. *Nature Human Behaviour* **2**(4), 284 (2018)
3. Aral, S., Muchnik, L., Sundararajan, A.: Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* **106**(51), 21,544–21,549 (2009)
4. Asur, S., Huberman, B.A.: Predicting the future with social media. In: *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 492–499. IEEE Computer Society (2010)
5. Bagrow, J.P., Danforth, C.M., Mitchell, L.: Which friends are more popular than you?: Contact strength and the friendship paradox in social networks. In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 103–108. ACM (2017)
6. Bakshy, E., Rosenn, I., Marlow, C., Adamic, L.: The role of social networks in information diffusion. In: *Proceedings of the 21st international conference on World Wide Web*, pp. 519–528. ACM (2012)
7. Bollen, J., Gonçalves, B., van de Leemput, I., Ruan, G.: The happiness paradox: your friends are happier than you. *EPJ Data Science* **6**(1), 4 (2017)
8. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. *Journal of computational science* **2**(1), 1–8 (2011)
9. Cao, Y., Ross, S.M.: The friendship paradox. *Mathematical Scientist* **41**(1) (2016)
10. Chamley, C.P.: *Rational herds*. Cambridge Books (2004)
11. Christakis, N.A., Fowler, J.H.: Social network sensors for early detection of contagious outbreaks. *PloS one* **5**(9), e12,948 (2010)
12. Cohen, R., Havlin, S., Ben-Avraham, D.: Efficient immunization strategies for computer networks and populations. *Physical review letters* **91**(24), 247,901 (2003)
13. Dasgupta, A., Kumar, R., Sivakumar, D.: Social sampling. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 235–243. ACM (2012)
14. Easley, D., Kleinberg, J.: *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press (2010)
15. Eom, Y.H., Jo, H.H.: Generalized friendship paradox in complex networks: The case of scientific collaboration. *Scientific Reports* **4** (2014)
16. Eom, Y.H., Jo, H.H.: Tail-scope: Using friends to estimate heavy tails of degree distributions in large-scale complex networks. *Scientific reports* **5**, 09,752 (2015)
17. Feld, S.L.: Why your friends have more friends than you do. *American Journal of Sociology* **96**(6), 1464–1477 (1991)

18. Fotouhi, B., Momeni, N., Allen, B., Nowak, M.A.: Conjoining uncooperative societies facilitates evolution of cooperation. *Nature Human Behaviour* **2**(7), 492–499 (2018). DOI 10.1038/s41562-018-0368-6. URL <https://doi.org/10.1038/s41562-018-0368-6>
19. Fotouhi, B., Momeni, N., Rabbat, M.G.: Generalized friendship paradox: An analytical approach. In: *International Conference on Social Informatics*, pp. 339–352. Springer (2014)
20. Garcia-Herranz, M., Moro, E., Cebrian, M., Christakis, N.A., Fowler, J.H.: Using friends as sensors to detect global-scale contagious outbreaks. *PLoS one* **9**(4), e92,413 (2014)
21. Gharehshiran, O.N., Krishnamurthy, V., Yin, G.: Distributed tracking of correlated equilibria in regime switching noncooperative games. *IEEE Transactions on Automatic Control* **58**(10), 2435–2450 (2013)
22. Gjoka, M., Kurant, M., Butts, C.T., Markopoulou, A.: Walking in facebook: A case study of unbiased sampling of OSNs. In: *Infocom, 2010 Proceedings IEEE*, pp. 1–9. IEEE (2010)
23. Gjoka, M., Kurant, M., Butts, C.T., Markopoulou, A.: Practical recommendations on crawling online social networks. *IEEE Journal on Selected Areas in Communications* **29**(9), 1872–1892 (2011)
24. Goel, S., Salganik, M.J.: Respondent-driven sampling as Markov chain Monte Carlo. *Statistics in medicine* **28**(17), 2202–2229 (2009)
25. Gomez-Rodriguez, M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **5**(4), 21 (2012)
26. Graefe, A.: Accuracy of vote expectation surveys in forecasting elections. *Public Opinion Quarterly* **78**(S1), 204–232 (2014)
27. Graefe, A.: Accuracy gains of adding vote expectation surveys to a combined forecast of us presidential election outcomes. *Research & Politics* **2**(1), 2053168015570,416 (2015)
28. Granovetter, M.S.: The strength of weak ties I. *American Journal of Sociology* **78**(6), 1360–1380 (1973)
29. Heckathorn, D.D.: Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems* **44**(2), 174–199 (1997)
30. Heckathorn, D.D.: Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social problems* **49**(1), 11–34 (2002)
31. Hernández-González, M., Basin, M.V.: Discrete-time filtering for nonlinear polynomial systems over linear observations. *International Journal of Systems Science* **45**(7), 1461–1472 (2014)
32. Hethcote, H.W.: The mathematics of infectious diseases. *SIAM review* **42**(4), 599–653 (2000)
33. Higham, D.J.: Centrality–friendship paradoxes: When our friends are more important than us. *arXiv preprint arXiv:1807.01496* (2018)
34. Hodas, N.O., Kooti, F., Lerman, K.: Friendship paradox redux: Your friends are more interesting than you. In: *Seventh International AAAI Conference on Weblogs and Social Media* (2013)
35. Hodas, N.O., Lerman, K.: How visibility and divided attention constrain social contagion. In: *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pp. 249–257. IEEE (2012)
36. Hoiles, W., Aprem, A., Krishnamurthy, V.: Engagement dynamics and sensitivity analysis of youtube videos. *ArXiv e-prints* (2016)
37. Horel, T., Singer, Y.: Scalable methods for adaptively seeding a social network. In: *Proceedings of the 24th International Conference on World Wide Web*, pp. 441–451. International World Wide Web Conferences Steering Committee (2015)
38. Jackson, M.O.: The friendship paradox and systematic biases in perceptions and social norms. *arXiv preprint arXiv:1605.04470* (2016)
39. Jackson, M.O., Rogers, B.W.: Relating network structure to diffusion properties through stochastic dominance. *The BE Journal of Theoretical Economics* **7**(1) (2007)
40. Jackson, M.O., Yariv, L.: Diffusion of behavior and equilibrium properties in network games. *American Economic Review* **97**(2), 92–98 (2007)
41. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146. ACM (2003)

42. Kim, D.A., Hwong, A.R., Stafford, D., Hughes, D.A., O'Malley, A.J., Fowler, J.H., Christakis, N.A.: Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet* **386**(9989), 145–153 (2015)
43. Kramer, J.B., Cutler, J., Radcliffe, A.: The multistep friendship paradox. *The American Mathematical Monthly* **123**(9), 900–908 (2016)
44. Krishnamurthy, V.: Quickest detection POMDPs with social learning: Interaction of local and global decision makers. *IEEE Transactions on Information Theory* **58**(8), 5563–5587 (2012)
45. Krishnamurthy, V.: *Partially Observed Markov Decision Processes*. Cambridge University Press (2016)
46. Krishnamurthy, V., Bhatt, S., Pedersen, T.: Tracking infection diffusion in social networks: Filtering algorithms and threshold bounds. *IEEE Transactions on Signal and Information Processing over Networks* **3**(2), 298–315 (2017)
47. Krishnamurthy, V., Gharehshiran, O.N., Hamdi, M., et al.: Interactive sensing and decision making in social networks. *Foundations and Trends® in Signal Processing* **7**(1-2), 1–196 (2014)
48. Krishnamurthy, V., Hoiles, W.: Online reputation and polling systems: Data incest, social learning and revealed preferences. *IEEE Transactions Computational Social Systems* **1**(3), 164–179 (2015)
49. Krishnamurthy, V., Hoiles, W.: Dynamics of information diffusion and social sensing. In: *Cooperative and Graph Signal Processing*, pp. 525–600. Elsevier (2018)
50. Krishnamurthy, V., Poor, H.V.: Social learning and Bayesian games in multiagent signal processing: How do local and global decision makers interact? *IEEE Signal Processing Magazine* **30**(3), 43–57 (2013)
51. Kumar, V., Krackhardt, D., Feld, S.: Network interventions based on inversivity: Leveraging the friendship paradox in unknown network structures. Tech. rep., Yale University (2018)
52. Kurtz, T.G.: *Approximation of population processes*, vol. 36. SIAM (1981)
53. Kushner, H.J., Yin, G.: *Stochastic approximation and recursive algorithms and applications*. No. 35 in *Applications of mathematics*. Springer, New York (2003)
54. Lattanzi, S., Singer, Y.: The power of random neighbors in social networks. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 77–86. ACM (2015)
55. Lee, E., Lee, S., Eom, Y.H., Holme, P., Jo, H.H.: Impact of perception models on friendship paradox and opinion formation. *arXiv preprint arXiv:1808.04170* (2018)
56. Lerman, K., Ghosh, R.: Information contagion: An empirical study of the spread of news on digg and twitter social networks. (2010)
57. Lerman, K., Yan, X., Wu, X.Z.: The “majority illusion” in social networks. *PloS one* **11**(2), e0147,617 (2016)
58. Leskovec, J., Faloutsos, C.: Sampling from large graphs. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 631–636. ACM (2006)
59. López-Pintado, D.: Diffusion in complex social networks. *Games and Economic Behavior* **62**(2), 573–590 (2008)
60. López-Pintado, D.: Influence networks. *Games and Economic Behavior* **75**(2), 776–787 (2012)
61. López-Pintado, D.: An overview of diffusion in complex networks. In: *Complex Networks and Dynamics*, pp. 27–48. Springer (2016)
62. Manski, C.F.: Measuring expectations. *Econometrica* **72**(5), 1329–1376 (2004)
63. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. *Annual review of sociology* **27**(1), 415–444 (2001)
64. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 29–42. ACM (2007)
65. Molloy, M., Reed, B.: A critical point for random graphs with a given degree sequence. *Random structures & algorithms* **6**(2-3), 161–180 (1995)

66. Murr, A.E.: “wisdom of crowds”? a decentralised election forecasting model that uses citizens local expectations. *Electoral Studies* **30**(4), 771–783 (2011)
67. Murr, A.E.: The wisdom of crowds: Applying Condorcet’s jury theorem to forecasting us presidential elections. *International Journal of Forecasting* **31**(3), 916–929 (2015)
68. Nettasinghe, B., Krishnamurthy, V.: Influence maximization over Markovian graphs: A stochastic optimization approach. *IEEE Transactions on Signal and Information Processing over Networks* (2018)
69. Nettasinghe, B., Krishnamurthy, V.: “What do your friends think?”: Efficient polling methods for networks using friendship paradox. arXiv preprint arXiv:1802.06505 (2018)
70. Nettasinghe, B., Krishnamurthy, V., Lerman, K.: Contagions in social networks: Effects of monophilic contagion, friendship paradox and reactive networks. arXiv preprint arXiv:1810.05822 (2018)
71. Newman, M.E.: Assortative mixing in networks. *Physical review letters* **89**(20), 208,701 (2002)
72. Newman, M.E., Watts, D.J., Strogatz, S.H.: Random graph models of social networks. *Proceedings of the National Academy of Sciences* **99**(suppl 1), 2566–2572 (2002)
73. Ogura, M., Preciado, V.M.: Epidemic processes over adaptive state-dependent networks. *Physical Review E* **93**(6), 062,316 (2016)
74. Paarporn, K., Eksin, C., Weitz, J.S., Shamma, J.S.: Networked SIS epidemics with awareness. *IEEE Transactions on Computational Social Systems* **4**(3), 93–103 (2017)
75. Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* **2**(1–2), 1–135 (2008)
76. Paré, P.E., Beck, C.L., Nedić, A.: Epidemic processes over time-varying networks. *IEEE Transactions on Control of Network Systems* **5**(3), 1322–1334 (2018)
77. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. *Physical review letters* **86**(14), 3200 (2001)
78. Rapoport, A.: Spread of information through a population with socio-structural bias: I. assumption of transitivity. *The bulletin of mathematical biophysics* **15**(4), 523–533 (1953)
79. Ribeiro, B., Towsley, D.: Estimating and sampling graphs with multidimensional random walks. In: *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pp. 390–403. ACM (2010)
80. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In: *Proceedings of the 20th international conference on World wide web*, pp. 695–704. ACM (2011)
81. Rothschild, D.M., Wolfers, J.: Forecasting elections: Voter intentions versus expectations (2011)
82. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: *Proceedings of the 19th international conference on World wide web*, pp. 851–860. ACM (2010)
83. Seeman, L., Singer, Y.: Adaptive seeding in social networks. In: *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pp. 459–468. IEEE (2013)
84. Shalizi, C.R., Thomas, A.C.: Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research* **40**(2), 211–239 (2011)
85. Sun, L., Axhausen, K.W., Lee, D.H., Cebrian, M.: Efficient detection of contagious outbreaks in massive metropolitan encounter networks. *Scientific reports* **4**, 5099 (2014)



Portfolio Optimization Using Regime-Switching Stochastic Interest Rate and Stochastic Volatility Models

R.H. Liu and D. Ren

Abstract This paper considers the continuous-time portfolio optimization problem with both stochastic interest rate and stochastic volatility in regime-switching models, where a regime-switching Vasicek model is assumed for the interest rate and a regime-switching Heston model is assumed for the stock price. We use the dynamic programming approach to solve this stochastic optimal control problem. Under suitable assumptions, we prove a verification theorem. We then derive a closed-form solution of the associated Hamilton-Jacobi-Bellman (HJB) equation for a power utility function and a special choice of some model parameters. We prove the optimality of the closed-form solution by verifying the required conditions as stated in the verification theorem. We present a numerical example to show the optimal portfolio policies and value functions in different regimes.

1 Introduction

Portfolio optimization with regime-switching has attracted much attention in recent years in the financial mathematics community. A continuous-time Markov chain is often embedded into the asset price models in order to capture the dynamical change of the asset prices across different stages of business cycles. The presence of regime-switching in market behavior has been examined by many researchers and well documented in the literature. For example, empirical studies have provided substantial support for including regime-switching in equity models [5], interest rate models [1] and stochastic volatility models [9].

R.H. Liu

Department of Mathematics, University of Dayton, 300 College Park, Dayton, OH 45469-2316
e-mail: rliu01@udayton.edu

D. Ren

Department of Mathematics, University of Dayton, 300 College Park, Dayton, OH 45469-2316
e-mail: dren01@udayton.edu

The problem of portfolio selection via utility maximization in regime-switching models has been studied in a number of papers. For example, [2, 3, 10] considered the problem of maximizing the expected utility from terminal wealth for a market with one stock and one bond, where the interest rate, the return rate and the volatility of the stock are all dependent on an external Markov chain. [11] considered a similar problem with one bond and multiple stocks by assuming that the return rates of the stocks are governed by an unobservable Markov chain which can be estimated by using the method of hidden Markov model (HMM) filtering. [4] added an option into the portfolio and considered the problem of maximizing the expected utility of the terminal wealth with regime-switching. [15] first completed the market by introducing a set of Markovian jump securities and then solved the problem of maximizing the expected utility from terminal wealth in this enlarged market. In a recent paper [13] we considered the optimal portfolio problem with stochastic interest rate where a regime-switching Vasicek model is assumed for the interest rate. [13] extended the model and results developed in [6] to the more general regime-switching models for both asset prices and interest rate.

In this work we continue to study the portfolio optimization problem in regime-switching models by further taking stochastic volatility into consideration. Aiming to explore the impact of transitions in macroeconomic conditions on the optimal decisions of an investor who is facing risks from both stock volatility and interest rate, we assume that the stock price follows a regime-switching Heston model and the interest rate is governed by a regime-switching Vasicek model. We solve this utility maximization control problem by using the dynamic programming approach. We prove a verification theorem under suitable assumptions. We then consider a power utility function and derive a closed-form solution for the optimal portfolio policy and the value function with special choices of some model parameters (see Section 4 for details). We verify that the required conditions in the verification theorem are satisfied by the closed-form solution and therefore establish the optimality of the solution. We note that the portfolio optimization problem using Heston model for stochastic volatility in the absence of regime-switching has been considered in the literature. For example, [7] solved the problem of maximizing utility from terminal wealth with respect to a power utility function using a Heston model for the stock price and a constant interest rate. [8] considered the problem by using a Cox-Ingersoll-Ross (CIR) model for the volatility process and another CIR model for the interest rate and derived a closed-form solution for the case when the stock price and volatility are driven by the same Brownian motion. Our results presented in this paper extend the related studies in the literature to the regime-switching models and provide insights to the portfolio behavior subject to changing macroeconomic conditions.

The paper is organized as follows. The optimal portfolio problem in regime-switching market with both stochastic interest rate and stochastic volatility is formulated in Section 2. The associated Hamilton-Jacobi-Bellman (HJB) equation is presented. A verification theorem is established in Section 3. Closed-form solution is derived in Section 4 for a regime-switching power utility function and a special choice of some model parameters. The conditions required in the verification theo-

rem of Section 3 is verified and hence the optimality of the closed-form solution is proved. A numerical example is provided in Section 5 to show the impact of regime-switching on the optimal portfolio decisions. Section 6 provides further remarks and concludes the paper.

2 Formulation of the Portfolio Optimization Problem

As commonly done in literature, we consider our problem in a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\alpha(t)$ be a continuous-time Markov chain taking values in a finite set $\mathcal{M} := \{1, \dots, m_0\}$ to model the random regime-switching, where $m_0 > 0$ is a fixed integer specifying the total number of market regimes. The intensity matrix (or the generator) $Q = (q_{ij})_{m_0 \times m_0}$ of $\alpha(t)$ is assumed given. It is known that q_{ij} 's satisfy: (I) $q_{ij} \geq 0$ if $i \neq j$; (II) $q_{ii} \leq 0$ and $q_{ii} = -\sum_{j \neq i} q_{ij}$ for each $i = 1, \dots, m_0$. In addition, Let $I(t) = (I_{\{\alpha(t)=1\}}, \dots, I_{\{\alpha(t)=m_0\}})^\top \in \mathbb{R}^{m_0}$, where I_A is the indicator function of the subset A . Then, in view of [14, Lemma 2.4], the process M^α defined by

$$M^\alpha(t) := I(t) - \int_0^t Q^\top I(s) ds \tag{1}$$

is an m_0 -dimensional martingale with respect to the filtration generated by the Markov chain $\{\alpha(t), t \geq 0\}$.

Both stock price and interest rate are assumed to depend on the market regime $\alpha(t)$. Specifically, The interest rate $r(t)$ follows a regime-switching Vasicek model given by

$$dr(t) = [a(\alpha(t)) - b(\alpha(t))r(t)]dt + \sigma_r(\alpha(t))dW^b(t), \tag{2}$$

where the coefficients $a(\alpha(t)), b(\alpha(t))$ and $\sigma_r(\alpha(t))$ are all regime-dependent, and $W^b(t)$ is an one-dimensional standard Brownian motion. We assume that $a(i) > 0, b(i) > 0$ and $\sigma_r(i) > 0$ for all $i \in \mathcal{M}$.

The stock price $S(t)$ follows a regime-switching Heston model given by

$$dS(t) = S(t) \left[(r(t) + \lambda_s(\alpha(t))z(t)) dt + \sigma_s(\alpha(t))\sqrt{z(t)}dW^s(t) \right], \tag{3}$$

$$dz(t) = [\theta(\alpha(t)) - \eta(\alpha(t))z(t)]dt + \sigma_z(\alpha(t))\sqrt{z(t)}dW^z(t), \tag{4}$$

where $\lambda_s(\alpha(t))z(t)$ is the risk premium of the stock, $\sigma_s(\alpha(t))\sqrt{z(t)}$ is the volatility of the stock, and $W^s(t)$ and $W^z(t)$ are two one-dimensional standard Brownian motions. Assume that $\lambda_s(i), \sigma_s(i), \theta(i), \eta(i)$ and $\sigma_z(i)$ are positive for all $i \in \mathcal{M}$.

Let ρ be the correlation coefficient between $W^s(t)$ and $W^z(t)$, i.e., $dW^s(t)dW^z(t) = \rho dt$. Let W^b be uncorrelated to W^s and W^z . We also assume that $\alpha(t)$ is independent of the Brownian motions $W^s(t), W^b(t)$ and $W^z(t)$.

We consider a market with one bond and one stock. The bond price $B(t)$ follows

$$dB(t) = B(t)r(t)dt, \tag{5}$$

where the interest rate $r(t)$ is given by (2).

In this work we aim to tackle the classical problem of finding the optimal allocation of the wealth between the stock and the bond, where the stock and bond are governed by the regime-switching models (3)-(5). Let $\pi(t)$ denote the percentage of total wealth invested in the stock at time t . Then the percentage of wealth invested in the bond is $1 - \pi(t)$. Let $X(t)$ denote the wealth at time t . Then $X(t)$ satisfies the following stochastic differential equation (SDE):

$$dX(t) = X(t)[r(t) + \pi(t)\lambda_s(\alpha(t))z(t)]dt + X(t)\pi(t)\sigma_s(\alpha(t))\sqrt{z(t)}dW^s(t). \quad (6)$$

Let $\mathcal{O} = (\mathbb{R}^+)^2 \times \mathbb{R}$, and $\mathcal{Q} = [t_0, T] \times \mathcal{O}$. Take any initial data $(t_0, X(t_0), z(t_0), r(t_0), \alpha(t_0)) = (t_0, x_0, z_0, r_0, i) \in \mathcal{Q} \times \mathcal{M}$, we introduce the admissible control in the following definition.

Definition 2.1 A stochastic process $\pi(\cdot) := \{\pi(t) : t_0 \leq t \leq T\}$ is an admissible control with respect to the initial data (t_0, x_0, z_0, r_0, i) if the following three conditions are satisfied:

1. $\pi(\cdot)$ is progressively measurable;
2. For this $\pi(\cdot)$, the SDE (6) has a path-wise unique solution $\{X^\pi(t)\}_{t \in [t_0, T]}$;
3. $X^\pi \geq 0$.

Let $\mathcal{A}_{t_0 x_0 z_0 r_0 i}$ denote the set of admissible controls for the initial data (t_0, x_0, z_0, r_0, i) .

We consider a regime-dependent utility function $U(x, i)$ with the properties: for each $i \in \mathcal{M}$, $U(0, i) = 0$, $U'(x, i) > 0$, $U''(x, i) < 0$ for $x > 0$, $\lim_{x \rightarrow 0^+} U'(x, i) = \infty$, $\lim_{x \rightarrow \infty} U'(x, i) = 0$, where U' and U'' denote the first and second order derivatives of U with respect to x .

Define the objective function J by

$$J(t_0, x_0, z_0, r_0, i; \pi(\cdot)) = \mathbb{E}^{t_0 x_0 z_0 r_0 i} [U(X^\pi(T), \alpha(T))], \quad (7)$$

where $T > 0$ is the investment horizon, $\mathbb{E}^{t_0 x_0 z_0 r_0 i}$ denotes the conditional expectation given $X(t_0) = x_0, Z(t_0) = z_0, r(t_0) = r_0$ and $\alpha(t_0) = i$, and $X^\pi(T)$ is the solution of the wealth equation (6) at time T when the control $\pi(\cdot)$ is being used, given by

$$X^\pi(T) = x_0 \exp \left\{ \int_{t_0}^T \left[\lambda_s(\alpha(s))\pi(s)z(s) + r(s) - \frac{1}{2}\sigma_s^2(\alpha(s))\pi_s^2(s)z(s) \right] ds + \int_{t_0}^T \sigma_s(\alpha(s))\pi(s)\sqrt{z(s)}dW^s(s) \right\}. \quad (8)$$

The value function is defined as

$$V(t_0, x_0, z_0, r_0, i) = \sup_{\pi(\cdot) \in \mathcal{A}_{t_0 x_0 z_0 r_0 i}} J(t_0, x_0, z_0, r_0, i; \pi(\cdot)), \quad (9)$$

with the terminal condition: $V(T, x, z, r, i) = U(x, i)$, for all $x \in \mathbb{R}^+, z \in \mathbb{R}^+, r \in \mathbb{R}$ and $i \in \mathcal{M}$.

For notation simplicity, in what follows, we denote $v(t, x, z, r, i)$ by $v(i)$ for $i \in \mathcal{M}$. Define the operator $\mathcal{L}^\pi(\cdot)$ by:

$$\begin{aligned} \mathcal{L}^\pi(v(i)) &= v_t(i) + xrv_x(i) + [a(i) - b(i)r]v_r(i) + \frac{1}{2}\sigma_r^2(i)v_{rr}(i) + [\theta(i) - \eta(i)z]v_z(i) \\ &\quad + \frac{1}{2}\sigma_z^2(i)zv_{zz}(i) + [x\pi\lambda_s(i)zv_x(i) + \frac{1}{2}x^2\pi^2\sigma_s^2(i)zv_{xx}(i) \\ &\quad + \rho\pi\sigma_s(i)\sigma_z(i)xzv_{xz}(i)] + \sum_{j \neq i} q_{ij} [v(j) - v(i)] = 0, \quad i = 1, \dots, m_0. \end{aligned} \quad (10)$$

Then the HJB equation associated with the optimal control problem (7)-(9) is the following system of m_0 coupled nonlinear partial differential equations:

$$\begin{aligned} &\sup_{\pi \in \mathbb{R}} \mathcal{L}^\pi(v(i)) \\ &= v_t(i) + xrv_x(i) + [a(i) - b(i)r]v_r(i) + \frac{1}{2}\sigma_r^2(i)v_{rr}(i) + [\theta(i) - \eta(i)z]v_z(i) \\ &\quad + \frac{1}{2}\sigma_z^2(i)zv_{zz}(i) + \sum_{j \neq i} q_{ij} [v(j) - v(i)] \\ &\quad + \sup_{\pi \in \mathbb{R}} \left\{ x\pi\lambda_s(i)zv_x(i) + \frac{1}{2}x^2\pi^2\sigma_s^2(i)zv_{xx}(i) + \rho\pi\sigma_s(i)\sigma_z(i)xzv_{xz}(i) \right\} \\ &= 0, \quad i = 1, \dots, m_0, \end{aligned} \quad (11)$$

with the terminal condition

$$v(T, x, z, r, i) = U(x, i), \quad \forall (x, z, r, i) \in \mathcal{O} \times \mathcal{M}. \quad (12)$$

Note that if $v_{xx}(i) < 0$, then the maximizer of (11), denoted by $\pi^*(t, i)$, is given by:

$$\pi^*(t, i) = -\frac{\lambda_s(i)v_x(i)}{x\sigma_s^2(i)v_{xx}(i)} - \frac{\rho\sigma_z(i)v_{xz}(i)}{x\sigma_s(i)v_{xx}(i)}. \quad (13)$$

3 Verification Theorem

In this section we establish a verification theorem for the optimal control problem formulated in Section 2.

Theorem 3.1. *Let $v \in C^{1,2}([t_0, T] \times \mathcal{O} \times \mathcal{M})$ be a positive solution of the HJB equation (11) with the terminal condition (12). Then*

- (a) $v(t_0, x_0, z_0, r_0, i) \geq \mathbb{E}^{t_0, x_0, z_0, r_0, i} [U(X^\pi(T), \alpha(T))]$, for all $\pi(\cdot) \in \mathcal{A}_{t_0, x_0, z_0, r_0, i}$.
- (b) Moreover, assume that the control $\pi^*(\cdot)$ given in (13) is an admissible control (i.e., $\pi^*(\cdot) \in \mathcal{A}_{t_0, x_0, z_0, r_0, i}$) and satisfies

$$\mathbb{E}^{t_0, x_0, z_0, r_0, i} \left[\int_{t_0}^T (\pi^*(s, i))^2 ds \right] < \infty, \text{ for all } i \in \mathcal{M}, \tag{14}$$

and for any sequences of stopping times $\{\tau_n\}_{n \in \mathbb{N}^+}$ with $t_0 \leq \tau_n \leq T$, the sequence $v(\tau_n, X^*(\tau_n), z(\tau_n), r(\tau_n), \alpha(\tau_n))$ is uniformly integrable. Then $\pi^*(\cdot)$ is an optimal control to the optimization problem (7)-(9), i.e.,

$$v(t_0, x_0, z_0, r_0, i) = \mathbb{E}^{t_0, x_0, z_0, r_0, i} [U(X^*(T), \alpha(T))]. \tag{15}$$

Here $X^*(\cdot)$ is the unique solution of the equation (6) when the control $\pi^*(\cdot)$ is being used.

Proof of Theorem 3.1. For notation simplicity, in the following proof, we let $\pi(s) = \pi(s, \alpha(s))$ and $\pi^*(s) = \pi^*(s, \alpha(s))$.

Part (a). Let v be a positive solution of the HJB equation (11), and $\pi(\cdot)$ be any admissible controls in $\mathcal{A}_{t_0, x_0, z_0, r_0, i}$. By assumption $v \in C^{1,2}([t_0, T] \times \mathcal{O} \times \mathcal{M})$, we apply the generalized Itô's formula for RCLL semimartingales to the process $v(s, X(s), z(s), r(s), \alpha(s))$ to get

$$\begin{aligned} & dv(s, X(s), z(s), r(s), \alpha(s)) \\ &= \mathcal{L}^\pi v(s, X(s), z(s), r(s), \alpha(s)) ds \\ &\quad + v_x(s, X(s), z(s), r(s), \alpha(s)) X(s) \pi(s) \sigma_s(\alpha(s)) \sqrt{z(s)} dW^s(s) \\ &\quad + v_z(s, X(s), z(s), r(s), \alpha(s)) \sigma_z(\alpha(s)) \sqrt{z(s)} dW^z(s) \\ &\quad + v_r(s, X(s), z(s), r(s), \alpha(s)) \sigma_r(\alpha(s)) dW^b(s) \\ &\quad + \sum_{j=1}^{m_0} v(s, X(s), z(s), r(s), j) dM_j^\alpha(s), \end{aligned} \tag{16}$$

where M_j^α is the j th component of the martingale M^α defined by (1).

Taking any $t \in [t_0, T]$ and integrating (16) from t_0 to t , we have:

$$\begin{aligned} & v(t, X(t), z(t), r(t), \alpha(t)) \\ &= v(t_0, X(t_0), z(t_0), r(t_0), \alpha(t_0)) + \int_{t_0}^t \mathcal{L}^\pi v(s, X(s), z(s), r(s), \alpha(s)) ds \\ &\quad + \int_{t_0}^t v_x(s, X(s), z(s), r(s), \alpha(s)) X(s) \pi(s) \sigma_s(\alpha(s)) \sqrt{z(s)} dW^s(s) \\ &\quad + \int_{t_0}^t v_z(s, X(s), z(s), r(s), \alpha(s)) \sigma_z(\alpha(s)) \sqrt{z(s)} dW^z(s) \\ &\quad + \int_{t_0}^t v_r(s, X(s), z(s), r(s), \alpha(s)) \sigma_r(\alpha(s)) dW^b(s) \\ &\quad + \sum_{j=1}^{m_0} \int_{t_0}^t v(s, X(s), z(s), r(s), j) dM_j^\alpha(s) \\ &\leq v(t_0, X(t_0), z(t_0), r(t_0), \alpha(t_0)) \end{aligned} \tag{17}$$

$$\begin{aligned}
& + \int_{t_0}^t v_x(s, X(s), z(s), r(s), \alpha(s)) X(s) \pi(s) \sigma_s(\alpha(s)) \sqrt{z(s)} dW^s(s) \\
& + \int_{t_0}^t v_z(s, X(s), z(s), r(s), \alpha(s)) \sigma_z(\alpha(s)) \sqrt{z(s)} dW^z(s) \\
& + \int_{t_0}^t v_r(s, X(s), z(s), r(s), \alpha(s)) \sigma_r(\alpha(s)) dW^b(s) \\
& + \sum_{j=1}^{m_0} \int_{t_0}^t v(s, X(s), z(s), r(s), j) dM_j^\alpha(s), \tag{18}
\end{aligned}$$

where the inequality holds because $\mathcal{L}^\pi v(s, X(s), z(s), r(s), \alpha(s)) \leq 0$ by (11).

Denote the right-hand side of the previous inequality by Y_t , then $Y := \{Y_t\}_{t \in [t_0, T]}$ is a local martingale. Moreover, since v is positive, so is Y , then the positive local martingale Y is a supermartingale. Taking expectations both sides of the previous inequality and letting $t = T$, we obtain:

$$\begin{aligned}
\mathbb{E}^{t_0, x_0, z_0, r_0} [v(T, X(T), z(T), r(T), \alpha(T))] & \leq \mathbb{E}^{t_0, x_0, z_0, r_0} [Y_T] \\
& \leq Y_{t_0} = v(t_0, X(t_0), z(t_0), r(t_0), \alpha(t_0)).
\end{aligned}$$

Part (a) then follows by the terminal condition (12).

Part (b). Now we consider the control $\pi^*(\cdot)$ given in (13).

Take an arbitrary integer n such that $0 < \frac{1}{n} < T - t_0$. Define

$$\mathcal{O}_n = \mathcal{O} \cap \left\{ (x, z, r) : \|(x, z, r)\| < n, \text{dist}((x, z, r), \partial \mathcal{O}) > \frac{1}{n} \right\}. \tag{19}$$

Let τ_n be the first exit time of $(t, X^*(t), z(t), r(t))$ from $[t_0, T - 1/n) \times \mathcal{O}_n$, i.e.,

$$\tau_n = \inf\{t \geq t_0 : (t, X^*(t), z(t), r(t)) \notin [t_0, T - 1/n) \times \mathcal{O}_n\}. \tag{20}$$

Note that $\tau_n \leq T - 1/n < T$.

Since $\mathcal{L}^{\pi^*} v(s, X^*(s), z(s), r(s), \alpha(s)) = 0$, by the similar arguments leading to (18), we can get

$$\begin{aligned}
& v(\tau_n, X^*(\tau_n), z(\tau_n), r(\tau_n), \alpha(\tau_n)) \\
& = v(t_0, X^*(t_0), z(t_0), r(t_0), \alpha(t_0)) \\
& + \int_{t_0}^{\tau_n} v_x(s, X^*(s), z(s), r(s), \alpha(s)) X^*(s) \pi^*(s) \sigma_s(\alpha(s)) \sqrt{z(s)} dW^s(s) \\
& + \int_{t_0}^{\tau_n} v_z(s, X^*(s), z(s), r(s), \alpha(s)) \sigma_z(\alpha(s)) \sqrt{z(s)} dW^z(s) \\
& + \int_{t_0}^{\tau_n} v_r(s, X^*(s), z(s), r(s), \alpha(s)) \sigma_r(\alpha(s)) dW^b(s) \\
& + \sum_{j=1}^{m_0} \int_{t_0}^{\tau_n} v(s, X^*(s), z(s), r(s), j) dM_j^\alpha(s). \tag{21}
\end{aligned}$$

Next we show that the expectation of each of the four stochastic integrals in (21) is zero.

For the first stochastic integral, since $v \in C^{1,2}$, then $v_x(s, X^*(s), z(s), r(s), \alpha(s))$ is bounded for all $s \in [t_0, \tau_n]$, so do $X^*(s), \sigma_s(\alpha(s))$ and $z(s)$. Hence for some positive constant C_1 :

$$\begin{aligned} & \mathbb{E}^{t_0, x_0, z_0, r_0, i} \int_{t_0}^{\tau_n} \left[v_x(s, X^*(s), z(s), r(s), \alpha(s)) X^*(s) \pi^*(s) \sigma_s(\alpha(s)) \sqrt{z(s)} \right]^2 ds \\ & \leq C_1 \mathbb{E}^{t_0, x_0, z_0, r_0, i} \int_{t_0}^{\tau_n} |\pi^*(s)|^2 ds \end{aligned} \tag{22}$$

$$\leq C_1 \mathbb{E}^{t_0, x_0, z_0, r_0, i} \int_{t_0}^T |\pi^*(s)|^2 ds \tag{23}$$

$$< \infty. \quad (\text{by assumption (14)}) \tag{24}$$

It follows that

$$\mathbb{E}^{t_0, x_0, z_0, r_0, i} \int_{t_0}^{\tau_n} v_x(s, X^*(s), z(s), r(s), \alpha(s)) X^*(s) \pi^*(s) \sigma_s(\alpha(s)) \sqrt{z(s)} dW^s(s) = 0.$$

Similarly, the expectations of the last three stochastic integrals are also zeros, by the boundedness of $v_z(s, X^*(s), z(s), r(s), \alpha(s)), \sigma_z(\alpha(s)), z(s), v_r(s, X^*(s), z(s), r(s), \alpha(s)), \sigma_r(\alpha(s))$ and $v(s, X^*(s), z(s), r(s), j)$ for $s \in [t_0, \tau_n]$ and all $j \in \mathcal{M}$.

Taking expectation in (21), we obtain

$$\begin{aligned} & \mathbb{E}^{t_0, x_0, z_0, r_0, i} v(\tau_n, X^*(\tau_n), z(\tau_n), r(\tau_n), \alpha(\tau_n)) \\ & = v(t_0, X^*(t_0), z(t_0), r(t_0), \alpha(t_0)). \end{aligned} \tag{25}$$

Now let $n \rightarrow \infty$. By noting that $\tau_n \rightarrow T$ a.s., v is continuous, and the assumption that the sequence $v(\tau_n, X^*(\tau_n), z(\tau_n), r(\tau_n), \alpha(\tau_n))$ is uniformly integrable, it follows that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \mathbb{E}^{t_0, x_0, z_0, r_0, i} v(\tau_n, X^*(\tau_n), z(\tau_n), r(\tau_n), \alpha(\tau_n)) \\ & = \mathbb{E}^{t_0, x_0, z_0, r_0, i} v(T, X^*(T), z(T), r(T), \alpha(T)) \end{aligned} \tag{26}$$

$$= v(t_0, X^*(t_0), z(t_0), r(t_0), \alpha(t_0)). \tag{27}$$

The proof of (15) is completed by the terminal condition (12). □

4 Closed-form Solution for A Power Utility Function

For the purpose of deriving a closed-form solution and verifying the verification Theorem 3.1, in this section, we consider a power utility function and let the parameter b, η and σ_z in the models (2)-(4) be regime independent. That is, for all regime

states $i \in \mathcal{M}$,

$$b(i) = b, \eta(i) = \eta, \sigma_z(i) = \sigma_z, \text{ for some constants } b, \eta \text{ and } \sigma_z.$$

Moreover, let $\sigma_s(i) = k_s \lambda_s(i)$, $i \in \mathcal{M}$ for some positive constant k_s .

The regime-dependent power utility function U takes the form:

$$U(x, i) = \frac{\delta(i)}{\gamma} x^\gamma, \quad (0 < \gamma < 1; \delta(i) > 0, \forall i \in \mathcal{M}), \quad (28)$$

which, obviously, satisfies all the properties assumed in Section 2. Then the corresponding value function (9) for the considered optimization problem becomes:

$$V(t_0, x_0, z_0, r_0, i) = \sup_{\pi(\cdot) \in \mathcal{A}_{t_0, x_0, z_0, r_0, i}} \mathbb{E}^{t_0, x_0, z_0, r_0, i} \left[\frac{\delta(\alpha(T))}{\gamma} X_T^\gamma \right]. \quad (29)$$

We take the following ansatz for the solution v of the HJB equation (11): for each $i \in \mathcal{M}$,

$$\begin{cases} v(i) = x^\gamma g(t, i) e^{\beta_1(t)r + \beta_2(t)z}, \\ g(T, i) = \frac{\delta(i)}{\gamma}, \\ \beta_1(T) = \beta_2(T) = 0, \end{cases} \quad (30)$$

where the function $g(t, i)$, $\beta_1(t)$ and $\beta_2(t)$ are determined next.

Substituting (13) and (30) into the HJB equation (11), we obtain a system of ordinary differential equations (ODE) for the functions $g(t, i)$, $\beta_1(t)$ and $\beta_2(t)$:

$$\begin{aligned} & g_t(t, i) + h(t, i)g(t, i) + rg(t, i) [\beta_1'(t) - b\beta_1(t) + \gamma] \\ & + zg(t, i) \left[\beta_2'(t) + \frac{(1 - \gamma + \gamma\rho^2)\sigma_z^2}{2(1 - \gamma)} \cdot \beta_2^2(t) + \left(\frac{\gamma\rho\sigma_z}{(1 - \gamma)k_s} - \eta \right) \beta_2(t) + \frac{\gamma}{2(1 - \gamma)k_s^2} \right] \\ & + \sum_{j \neq i} q_{ij} [g(t, j) - g(t, i)] = 0 \end{aligned} \quad (31)$$

for $i = 1, \dots, m_0$, where

$$h(t, i) = a(i)\beta_1(t) + \frac{1}{2}\sigma_r^2(i)\beta_1^2(t) + \theta(i)\beta_2(t). \quad (32)$$

Since the ODE (31) holds for all $r \geq 0$ and $z \geq 0$, it is necessary that

$$\beta_1'(t) - b\beta_1(t) + \gamma = 0, \quad (33)$$

$$\beta_2'(t) + \frac{(1 - \gamma + \gamma\rho^2)\sigma_z^2}{2(1 - \gamma)} \cdot \beta_2^2(t) + \left(\frac{\gamma\rho\sigma_z}{(1 - \gamma)k_s} - \eta \right) \beta_2(t) + \frac{\gamma}{2(1 - \gamma)k_s^2} = 0. \quad (34)$$

Using the condition $\beta_1(T) = 0$, the solution $\beta_1(t)$ of (33) is:

$$\beta_1(t) = \frac{\gamma}{b} [1 - e^{-b(T-t)}]. \quad (35)$$

To solve $\beta_2(t)$, rewrite the ODE (34) as

$$\frac{d\beta_2(t)}{\frac{(1-\gamma+\gamma\rho^2)\sigma_z^2}{2(1-\gamma)} \cdot \beta_2^2(t) + \left(\frac{\gamma\rho\sigma_z}{(1-\gamma)k_s} - \eta\right) \beta_2(t) + \frac{\gamma}{2(1-\gamma)k_s^2}} = -dt. \quad (36)$$

Note that the denominator of the left-hand side of (36) is a quadratic form of $\beta_2(t)$, whose determinant is

$$\begin{aligned} \Delta_{\beta_2} &= \left(\frac{\gamma\rho\sigma_z}{(1-\gamma)k_s} - \eta\right)^2 - 4\left(\frac{(1-\gamma+\gamma\rho^2)\sigma_z^2}{2(1-\gamma)}\right) \left(\frac{\gamma}{2(1-\gamma)k_s^2}\right) \\ &= \frac{-[\eta^2k_s^2 + 2\rho\eta k_s\sigma_z + \sigma_z^2]\gamma + \eta^2k_s^2}{(1-\gamma)k_s^2}. \end{aligned} \quad (37)$$

Since $\gamma < 1$, then $\Delta_{\beta_2} > 0$ if and only if

$$-[\eta^2k_s^2 + 2\rho\eta k_s\sigma_z + \sigma_z^2]\gamma + \eta^2k_s^2 > 0. \quad (38)$$

Furthermore, observe that, for $\rho \geq -1$:

$$\eta^2k_s^2 + 2\rho\eta k_s\sigma_z + \sigma_z^2 \geq \eta^2k_s^2 - 2\eta k_s\sigma_z + \sigma_z^2 = (\eta k_s - \sigma_z)^2 \geq 0.$$

Therefore, $\Delta_{\beta_2} > 0$ if and only if

$$\gamma < \min\left\{1, \frac{\eta^2k_s^2}{\eta^2k_s^2 + 2\rho\eta k_s\sigma_z + \sigma_z^2}\right\}. \quad (39)$$

Under the assumption (39), the ODE (36) can be written as:

$$\frac{d\beta_2(t)}{\frac{(1-\gamma+\gamma\rho^2)\sigma_z^2}{2(1-\gamma)}(\beta_2(t) - \kappa_1)(\beta_2(t) - \kappa_2)} = -dt, \quad (40)$$

with

$$\begin{aligned} \kappa_1 &= \frac{-\gamma\rho\sigma_z/k_s + (1-\gamma)\eta + (1-\gamma)\sqrt{\Delta_{\beta_2}}}{(1-\gamma+\gamma\rho^2)\sigma_z^2} > \kappa_2, \\ \kappa_2 &= \frac{-\gamma\rho\sigma_z/k_s + (1-\gamma)\eta - (1-\gamma)\sqrt{\Delta_{\beta_2}}}{(1-\gamma+\gamma\rho^2)\sigma_z^2}. \end{aligned} \quad (41)$$

Furthermore, if

$$\frac{\gamma\rho\sigma_z}{(1-\gamma)k_s} - \eta < 0, \quad (42)$$

then $\kappa_2 > 0$.

The solution $\beta_2(t)$ of (40), combining with $\beta_2(T) = 0$, is:

$$\beta_2(t) = \frac{\gamma}{(1-\gamma+\gamma\rho^2)k_s^2\sigma_z^2} \cdot \frac{e^{\sqrt{\Delta_{\beta_2}}(T-t)} - 1}{e^{\sqrt{\Delta_{\beta_2}}(T-t)}\kappa_1 - \kappa_2}. \quad (43)$$

Note that (31) becomes a system of m_0 ordinary differential equations (ODE) given by

$$g_i(t, i) + h(t, i)g(t, i) + \sum_{j=1}^{m_0} q_{ij}g(t, j) = 0 \tag{44}$$

for $i = 1, \dots, m_0$, where we rewrite $\sum_{j \neq i} q_{ij}[g(t, j) - g(t, i)]$ as $\sum_{j=1}^{m_0} q_{ij}g(t, j)$ by using the property $q_{ii} = -\sum_{j \neq i} q_{ij}$.

To solve for $g(t, i)$, we rewrite the ODE system (44) together with the terminal condition into the following vector form:

$$\begin{cases} G'(t) + H(t)G(t) = 0, \\ G(T) = \frac{1}{\gamma}\Delta, \end{cases} \tag{45}$$

where

$$G(t) = \begin{pmatrix} g(t, 1) \\ g(t, 2) \\ \vdots \\ g(t, m_0) \end{pmatrix}, \Delta = \begin{pmatrix} \delta(1) \\ \delta(2) \\ \vdots \\ \delta(m_0) \end{pmatrix}, \tag{46}$$

and

$$H(t) = \begin{pmatrix} h(t, 1) + q_{11} & q_{12} & \cdots & q_{1m_0} \\ q_{21} & h(t, 2) + q_{22} & \cdots & q_{2m_0} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m_0 1} & \cdots & q_{m_0(m_0-1)} & h(t, m_0) + q_{m_0 m_0} \end{pmatrix}. \tag{47}$$

By the continuity of $\beta_1(t)$, $\beta_2(t)$ and hence the continuity of $h(t, i)$ for each i , the solution of (45) exists and takes the form:

$$G(t) = \frac{1}{\gamma} e^{\int_t^T H(s) ds} \Delta. \tag{48}$$

Using (30) and (43), (13) becomes

$$\pi^*(t, i) = \frac{1}{(1-\gamma)k_s^2 \lambda_s(i)} + \frac{\rho \sigma_z}{(1-\gamma)k_s \lambda_s(i)} \beta_2(t) \tag{49}$$

$$= \frac{1}{(1-\gamma)k_s^2 \lambda_s(i)} + \frac{\rho \gamma}{(1-\gamma)(1-\gamma + \gamma \rho^2)k_s^3 \lambda_s(i) \sigma_z} \cdot \frac{e^{\sqrt{\Delta \beta_2}(T-t)} - 1}{e^{\sqrt{\Delta \beta_2}(T-t)} \kappa_1 - \kappa_2}. \tag{50}$$

In what follows we will show that $\pi^*(t, \alpha(t))$ given in (50) is indeed an optimal control of the considered optimization problem. Before that, it is helpful to provide some interpretations of π^* .

1. π^* depends on time t through the term $(T - t)$, and also on the regime state $\alpha(t)$ at t through the term λ_s .
2. The first term in π^* is of the same form of the stock allocation for a classical Merton's problem for each fixed regime i , with stock's risk premium being

$\lambda_s(i)z(t)$, and stock's volatility being $k_s\lambda_s(i)\sqrt{z(t)}$ (by deliberately ignoring the stochasticity of z).

3. Recalling that $0 < \gamma < 1$, and observing that

$$1 - \gamma + \gamma\rho^2 = 1 - \gamma(1 - \rho^2) \geq 1 - (1 - \rho^2) = \rho^2 \geq 0,$$

(Note: the equalities in the previous two “ \geq ” do not hold at the same time.)

$$e\sqrt{\Delta\beta_2}^{(T-t)} - 1 \geq 0, \text{ and } e\sqrt{\Delta\beta_2}^{(T-t)} \kappa_1 - \kappa_2 > 0,$$

(51)

the second term in π^* may be positive, negative or zero, depending on the sign of ρ – the correlation between W^s and W^z . In particular, when $\rho = 0$, i.e., W^s and W^z are uncorrelated, π^* equals to the Merton's result.

More specifically, if the two Brownian motion W^s and W^z are positively correlated, our model allocates more on stock than the classical Merton's model; while if W^s and W^z are negatively correlated, our model allocates less on stock than the Merton's model.

4. As t approaches to the terminal time T , π^* approaches to the Merton's result.

5. When γ approaches to 0, π^* also approaches to the Merton's result.

Theorem 4.1. For the power utility $U(x, i) = \frac{\delta(i)}{\gamma}x^\gamma$, ($0 < \gamma < 1$; $\delta(i) > 0$), under the assumptions (39) and (42), the control $\pi^*(\cdot)$ given in (50) is an optimal control to the optimization problem (7)-(9), and the value function is given as in (30):

$$V(t_0, x_0, z_0, r_0, i) = x_0^\gamma \cdot g(t_0, i) \exp\{\beta_1(t_0)r_0 + \beta_2(t_0)z_0\},$$

with $\beta_1(\cdot)$ defined in (35), $\beta_2(\cdot)$ defined in (43), and $g(\cdot, i)$ being the i -th component of the vector $G(\cdot)$ in (48).

Proof of Theorem 4.1. To prove this theorem, it suffices to check that all assumptions in Theorem 3.1 are satisfied.

- (i) Based on the previous calculation and the positivity of $g(\cdot, \cdot)$, v defined in (30) is obviously a positive solution of the HJB equation (11) with the terminal condition (12), and $v \in C^{1,2}([t_0, T] \times \mathcal{O} \times \mathcal{M})$.
- (ii) To show that $\pi^*(\cdot)$ given in (50) satisfies (14), i.e.,

$$\mathbb{E}^{t_0, x_0, z_0, r_0, i} \left[\int_{t_0}^T (\pi^*(s, i))^2 ds \right] < \infty, \text{ for all } i \in \mathcal{M}. \tag{52}$$

Straightforward calculation shows that $\beta_2(t) < \hat{\beta}_2 = \frac{\gamma}{(1 - \gamma + \gamma\rho^2)k_s^2\sigma_z^2\kappa_1}$ for all $t \in [t_0, T]$. Then

$$\pi^*(t, i) = \frac{1}{(1 - \gamma)k_s^2\lambda_s(i)} + \frac{\rho\sigma_z}{(1 - \gamma)k_s\lambda_s(i)}\beta_2(t) \tag{53}$$

$$< \frac{1}{(1 - \gamma)k_s^2\lambda_s(i)} + \frac{\rho\sigma_z}{(1 - \gamma)k_s\lambda_s(i)}\hat{\beta}_2, \tag{54}$$

which implies (14) naturally.

(iii) To show that $\pi^*(\cdot)$ given in (50) is an admissible control.

- a. $\pi^*(\cdot)$ is obviously progressively measurable.
- b. For this $\pi^*(\cdot)$, the SDE (6) has a path-wise unique solution X^* , given by:

$$X^*(t) = x_0 \exp \left\{ \int_{t_0}^t \left[\lambda_s(\alpha(s)) \pi^*(s) z(s) + r(s) - \frac{1}{2} \sigma_s^2(\alpha(s)) (\pi^*(s))^2 \right. \right. \\ \left. \left. \times z(s) \right] ds + \int_{t_0}^t \sigma_s(\alpha(s)) \pi^*(s) \sqrt{z(s)} dW^s(s) \right\}, \quad (55)$$

which is obviously nonnegative.

By Definition 2.1, $\pi^*(\cdot)$ is an admissible control.

(iv) To show that for any sequences of stopping times $\{\tau_n\}_{n \in \mathbb{N}^+}$ with $t_0 \leq \tau_n \leq T$, the sequence $v(\tau_n, X^*(\tau_n), z(\tau_n), r(\tau_n), \alpha(\tau_n))$ is uniformly integrable, with

$$v(t, x, z, r, i) = x^\gamma \cdot g(t, i) \exp\{\beta_1(t)r + \beta_2(t)z\}. \quad (56)$$

For brevity, we denote $v(t, X^*(t), z(t), r(t), \alpha(t))$ by v_t^* , $v(\tau_n, X^*(\tau_n), z(\tau_n), r(\tau_n), \alpha(\tau_n))$ by $v_{\tau_n}^*$, and $X^*(\tau_n)$ by X_n^* in the following arguments.

To show the uniform integrability of $v_{\tau_n}^*$, it suffices to show that:

$$\sup_{n \geq 1} \mathbb{E}^{t_0, x_0, z_0, r_0, i} [(v_{\tau_n}^*)^q] < \infty, \text{ for some } q > 1. \quad (57)$$

By (55) and (56), for any $q > 1$, we have:

$$(v_t^*)^q = (x_0)^{\gamma q} \cdot (g(t, \alpha(t)))^q Y(t) \exp\{q\beta_2(t)z(t)\} \\ \times \exp \left\{ \int_{t_0}^t \gamma q \left[\lambda_s(\alpha(s)) \pi^*(s) z(s) - \frac{1}{2} \sigma_s^2(\alpha(s)) (\pi^*(s))^2 z(s) \right] ds \right. \\ \left. + \int_{t_0}^t \gamma q \sigma_s(\alpha(s)) \pi^*(s) \sqrt{z(s)} dW^s(s) \right\} \\ \leq (x_0)^{\gamma q} \cdot (g(t, \alpha(t)))^q Y(t) \exp\{q\hat{\beta}_2 z(t)\} \\ \times \exp \left\{ \int_{t_0}^t \gamma q \left[\lambda_s(\alpha(s)) \pi^*(s) z(s) - \frac{1}{2} \sigma_s^2(\alpha(s)) (\pi^*(s))^2 z(s) \right] ds \right. \\ \left. + \int_{t_0}^t \gamma q \sigma_s(\alpha(s)) \pi^*(s) \sqrt{z(s)} dW^s(s) \right\}, \quad (59)$$

where

$$Y(t) = \exp \left\{ \int_{t_0}^t \gamma q r(s) ds + q \beta_1(t) r(t) \right\},$$

$$\hat{\beta}_2 = \frac{\gamma}{(1 - \gamma + \gamma \rho^2) k_s^2 \sigma_z^2 \kappa_1},$$

and the inequality (59) holds since $\beta_2(t) < \hat{\beta}_2$ and $z(t) > 0$ for all t .
By (4),

$$z(t) = z(t_0) + \int_{t_0}^t [\theta(\alpha(s)) - \eta(\alpha(s))z(s)] ds + \int_{t_0}^t \sigma_z(\alpha(s)) \sqrt{z(s)} dW^z(s). \tag{60}$$

Then (59) can be rewritten as:

$$\begin{aligned} & (x_0)^{\gamma q} \cdot (g(t, \alpha(t)))^q Y(t) \\ & \times \exp \left\{ q \hat{\beta}_2 \left[z(t_0) + \int_{t_0}^t [\theta(\alpha(s)) - \eta(\alpha(s))z(s)] ds \right. \right. \\ & \left. \left. + \int_{t_0}^t \sigma_z(\alpha(s)) \sqrt{z(s)} dW^z(s) \right] \right. \\ & \left. + \int_{t_0}^t \gamma q \left[\lambda_s(\alpha(s)) \pi^*(s) z(s) - \frac{1}{2} \sigma_s^2(\alpha(s)) (\pi^*(s))^2 z(s) \right] ds \right. \\ & \left. + \int_{t_0}^t \gamma q \sigma_s(\alpha(s)) \pi^*(s) \sqrt{z(s)} dW^s(s) \right\} \\ & = (x_0)^{\gamma q} \cdot (g(t, \alpha(t)))^q \exp \left\{ q \hat{\beta}_2 \left[z(t_0) + \int_{t_0}^t \theta(\alpha(s)) ds \right] \right\} \\ & \times Y(t) M(t) \exp \left\{ \int_{t_0}^t q \zeta(s) z(s) ds \right\}, \end{aligned} \tag{61}$$

where

$$\begin{aligned} M(t) = & \exp \left\{ q \hat{\beta}_2 \int_{t_0}^t \sigma_z(\alpha(s)) \sqrt{z(s)} dW^z(s) \right. \\ & \left. + \int_{t_0}^t \gamma q \sigma_s(\alpha(s)) \pi^*(s) \sqrt{z(s)} dW^s(s) \right. \\ & \left. + \int_{t_0}^t \left[-\frac{q^2 \hat{\beta}_2^2}{2} \sigma_z^2(\alpha(s)) z(s) - \frac{\gamma^2 q^2}{2} \sigma_s^2(\alpha(s)) (\pi^*(s))^2 z(s) \right. \right. \\ & \left. \left. - \rho \hat{\beta}_2 \gamma q^2 \sigma_z(\alpha(s)) \sigma_s(\alpha(s)) \pi^*(s) z(s) \right] ds \right\}, \\ \zeta(s) = & \gamma \lambda_s(\alpha(s)) \pi^*(s) + \left(\frac{\gamma^2 q - \gamma}{2} \right) \sigma_s^2(\alpha(s)) (\pi^*(s))^2 \end{aligned} \tag{62}$$

$$\tag{63}$$

$$\begin{aligned}
 & -\hat{\beta}_2\eta(\alpha(s)) + \frac{q\hat{\beta}_2^2}{2}\sigma_z^2(\alpha(s)) + \rho\hat{\beta}_2\gamma q\sigma_z(\alpha(s))\sigma_s(\alpha(s))\pi^*(s) \\
 = & \left[\frac{\gamma^2}{2}\sigma_s^2(\alpha(s))(\pi^*(s))^2 + \frac{\hat{\beta}_2^2}{2}\sigma_z^2(\alpha(s)) \right. \\
 & \left. + \rho\hat{\beta}_2\gamma\sigma_z(\alpha(s))\sigma_s(\alpha(s))\pi^*(s) \right] q \\
 & + \left[\gamma\lambda_s(\alpha(s))\pi^*(s) - \frac{\gamma}{2}\sigma_s^2(\alpha(s))(\pi^*(s))^2 - \hat{\beta}_2\eta(\alpha(s)) \right].
 \end{aligned} \tag{64}$$

When $q = 1$, by plugging the form of $\pi^*(s)$ given in (49), $\zeta(s)$ becomes:

$$\begin{aligned}
 \zeta(s)|_{q=1} = & -\frac{\gamma\rho^2\sigma_z^2(\alpha(s))}{2(1-\gamma)}\beta_2^2(s) + \frac{\gamma\hat{\beta}_2\rho^2\sigma_z^2(\alpha(s))}{1-\gamma}\beta_2(s) \\
 & + \frac{\gamma + 2\gamma\hat{\beta}_2\rho k_s\sigma_z(\alpha(s))}{2(1-\gamma)k_s^2} - \frac{\hat{\beta}_2(2\eta(\alpha(s)) - \hat{\beta}_2\sigma_z^2(\alpha(s)))}{2}.
 \end{aligned} \tag{65}$$

Treating (65) as a quadratic form in $\beta_2(s)$, and recalling that $\beta_2(s) < \hat{\beta}_2$, straightforward calculation implies that

$$\zeta(s)|_{q=1} < \zeta(s)|_{q=1, \beta_2(s)=\hat{\beta}_2} = 0 \tag{66}$$

On the other hand, note that $\zeta(s)$ is linear in q ; and by recalling that $\rho \in [-1, 1]$, the coefficient of q is:

$$\frac{\gamma^2}{2}\sigma_s^2(\alpha(s))(\pi^*(s))^2 + \frac{\hat{\beta}_2^2}{2}\sigma_z^2(\alpha(s)) + \rho\hat{\beta}_2\gamma\sigma_z(\alpha(s))\sigma_s(\alpha(s))\pi^*(s) \tag{67}$$

$$\geq \frac{1}{2} \left[\hat{\beta}_2\sigma_z(\alpha(s)) - \gamma\sigma_s(\alpha(s))|\pi^*(s)| \right]^2 \geq 0, \tag{68}$$

Therefore, there exists some $q > 1$ such that $\zeta(s) < 0$ for all s , implying that the last term $\exp \left\{ \int_{t_0}^t q\zeta(s)z(s)ds \right\}$ in (61) is less than 1. Together with (59) and (61), it follows that,

$$(v_t^*)^q \leq (x_0)^{\gamma q} \cdot (g(t, \alpha(t)))^q \tag{69}$$

$$\times \exp \left\{ q\hat{\beta}_2 \left[z(t_0) + \int_{t_0}^t \theta(\alpha(s))ds \right] \right\} Y(t)M(t) \tag{70}$$

$$\leq C_3(t)Y(t)M(t), \tag{71}$$

where $C_3(t)$ is some positive deterministic term which is bounded on $[t_0, T]$. Note that $Y(t)$ and $M(t)$ are uncorrelated, by (2) and (4), and the assumption that W^b is uncorrelated to W^s and W^z , then

$$\mathbb{E}^{t_0, x_0, z_0, r_0, i} [(v_t^*)^q] \leq C_3(t) \mathbb{E}^{t_0, x_0, z_0, r_0, i} [Y(t)] \mathbb{E}^{t_0, x_0, z_0, r_0, i} [M(t)] \tag{72}$$

In view of [13, Eq. (3.36) and (3.41)], we have $\mathbb{E}^{t_0, x_0, z_0, r_0 i}[\sup_{t_0 \leq t \leq T} Y(t)] < \infty$. On the other hand, note that M is a positive local martingale, hence a supermartingale with $M(t_0) = 1$. By the optional stopping theorem (OS), we have that for all stopping time τ_n with $t_0 \leq \tau_n \leq T$:

$$\sup_{n \geq 1} \mathbb{E}^{t_0, x_0, z_0, r_0 i}[(Y_{\tau_n}^*)^q] \leq \sup_{t \in [t_0, T]} C_3(t) \cdot \sup \mathbb{E}^{t_0, x_0, z_0, r_0 i}[Y(\tau_n)] < \infty. \tag{73}$$

Therefore, (57) follows, and this completes the proof of Theorem 4.1.

5 A Numerical Example

In this section we present a numerical example to show the different optimal portfolio policies and different value functions in different market regimes. The results clearly indicate the important role played by the regime-switching as introduced in the market models.

In the example we consider a market with two regimes ($m_0 = 2$). The generator of the Markov chain $\alpha(\cdot)$ is given by

$$Q = \begin{pmatrix} -q_{12} & q_{12} \\ q_{21} & -q_{21} \end{pmatrix},$$

where q_{12} and q_{21} are the switching rates from regime 1 to regime 2 and from regime 2 to regime 1, respectively. We set $q_{12} = 3$ and $q_{21} = 4$, implying that on average the market switches three times per year from regime 1 to regime 2 and four times from regime 2 to regime 1.

The various model parameters used in the numerical example are chosen as the following: for the interest rate model (2), $a(1) = 0.16$, $a(2) = 0.08$, $b(1) = b(2) = 2$, $\sigma_r(1) = 0.03$, $\sigma_r(2) = 0.05$; for the regime-switching Heston model (3)-(4) for the stock price and volatility, $\sigma_s(1) = 0.3$, $\sigma_s(2) = 0.5$, $\theta(1) = 0.2$, $\theta(2) = 0.4$, $\eta(1) = \eta(2) = 2$, $\sigma_z(1) = \sigma_z(2) = 0.3$, and $[\lambda_s(1), \lambda_s(2)] = [\sigma_s(1), \sigma_s(2)]/k_s$ with $k_s = 7$, the correlation coefficient ρ between $W^s(t)$ and $W^z(t)$ is $\rho = -0.5$. Note that σ_s and λ_s satisfy $0 < \sigma_s(1) < \sigma_s(2)$, $0 < \lambda_s(1) < \lambda_s(2)$ and $\frac{\lambda_s(2)}{\sigma_s^2(2)} < \frac{\lambda_s(1)}{\sigma_s^2(1)}$. In view of the discussions in [12, section 5], we may treat regime 1 as a bull market and regime 2 a bear market. The utility functions for the two regimes are $U(x, 1) = 4x^{0.5}$ and $U(x, 2) = 2x^{0.5}$, respectively (that is, $\gamma = 0.5$, $\delta(1) = 2$, $\delta(2) = 1$ in (28)). The investment horizon is set to $T = 1$ (year). Straightforward calculation shows that the assumptions (39) and (42) are both satisfied.

The optimal stock allocations $\pi^*(t, 1)$ and $\pi^*(t, 2)$ as given in (50) are displayed in Fig. 1. We see from the displayed optimal controls that $\pi^*(t, 1)$ is always bigger than $\pi^*(t, 2)$. This indicates, as expected, that the investor should always invest a larger percentage of his/her wealth in the stock in the bull market (regime 1) and a smaller percentage in the bear market (regime 2). Note that the market regime may change at any time, then the investor should change his/her portfolio accord-

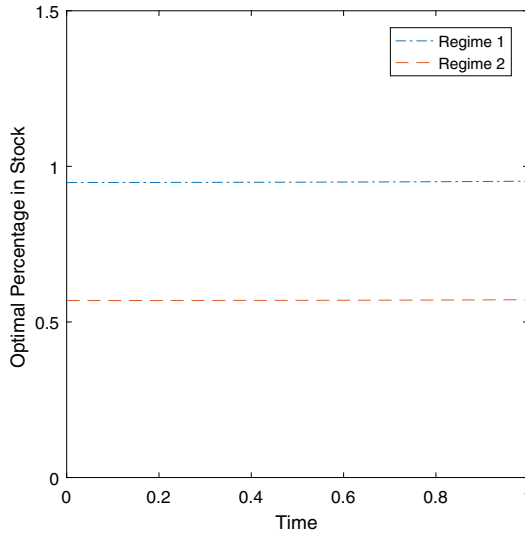


Fig. 1 Optimal percentages of wealth in stock (two regimes). The values of parameters are: $a(1) = 0.16$, $a(2) = 0.08$, $b(1) = b(2) = 2$, $\sigma_r(1) = 0.03$ and $\sigma_r(2) = 0.05$ in equation (2); $\sigma_s(1) = 0.3$, $\sigma_s(2) = 0.5$, $\theta(1) = 0.2$, $\theta(2) = 0.4$, $\eta(1) = \eta(2) = 2$, $\sigma_z(1) = \sigma_z(2) = 0.3$, and $[\lambda_s(1), \lambda_s(2)] = [\sigma_s(1), \sigma_s(2)]/k_s$ with $k_s = 7$, in equations (3) and (4). The correlation coefficient ρ between $W^s(t)$ and $W^z(t)$ is $\rho = -0.5$, and $\gamma = 0.5$, $\delta(1) = 2$, $\delta(2) = 1$ in the utility function (28). The investment horizon is $T = 1$, and the initial time $t_0 = 0$.

ingly at the time when a regime-switching occurs. For example, at $t = 0.5$, we have $\pi^*(0.5, 1) = 0.9492$ and $\pi^*(0.5, 2) = 0.5695$ from the calculation. As a result, if the regime switches from bull to bear market at $t = 0.5$, then the investor should reduce his/her investment in the stock from 94.92% to 56.95% of the total wealth (consequently, his/her investment in the bond would increase from 5.08% to 43.05%). Fig. 1 also shows that the optimal percentages have only slight changes as time t changes. Indeed, $\pi^*(t, 1)$ increases from $\pi^*(0, 1) = 0.9480$ to $\pi^*(1, 1) = 0.9524$ and $\pi^*(t, 2)$ increases from $\pi^*(0, 2) = 0.5688$ to $\pi^*(1, 2) = 0.5714$. Hence, the regime-switching has the major impact on the portfolio allocation.

The value functions $V(t, x, z, r, 1)$ and $V(t, x, z, r, 2)$ are displayed in Fig. 2. The surfaces in the upper panel show V as functions of time t and interest rate r while the wealth and variance are fixed at $x = 1$ and $z = 0.2$; The surfaces in the middle panel show V as functions of time t and variance z while the wealth and interest rate are fixed at $x = 1$ and $r = 0.05$; The surfaces in the lower panel show V as functions of time t and wealth x while the variance and interest rate are fixed at $z = 0.2$ and $r = 0.05$. We can clearly see that the value functions are different in different market regimes.

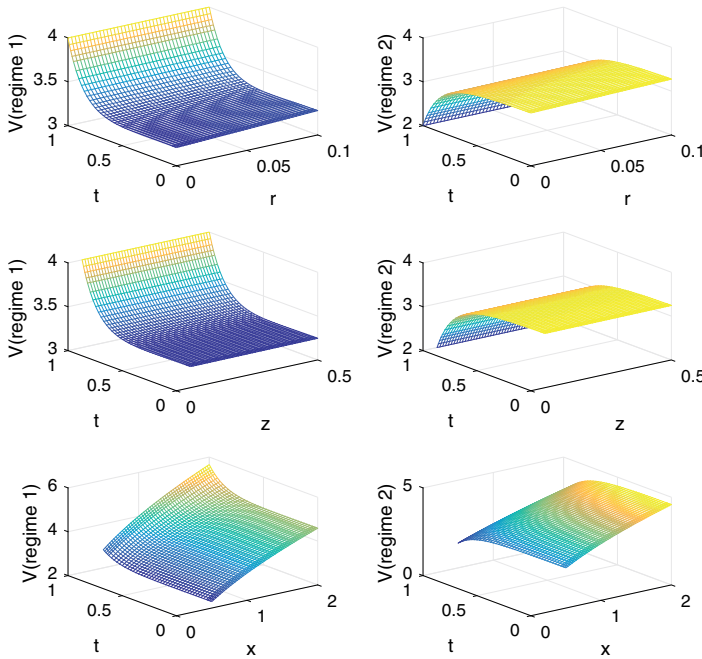


Fig. 2 Value functions (two regimes), with the same parameter values as in Fig.1.

6 Concluding Remarks

We have studied in this paper the portfolio optimization via utility maximization using regime-switching models where a regime-switching Vasicek model is assumed for the interest rate and a regime-switching Heston model is used for the stock price. We have used the dynamic programming approach to obtain a verification result for the formulated stochastic optimal control problem. We have derived a closed-form solution of the associated Hamilton-Jacobi-Bellman (HJB) equation for a power utility function and a special choice of some model parameters, and proved the optimality of the constructed control policy. Including a Markov chain for regime-switching in the market models can better describe the market changes under different macroeconomic conditions and the obtained results can help us better understand the optimal investment decision an investor should follow in different market conditions.

We note that it would be extremely difficult, if not impossible to find a closed-form solution for the general HJB equation (11) and other types of utility function. In fact, to establish the regularity properties of the value function is a very challenging job. For the future research, we are interested in establishing the viscosity

solution property of the value function to the HJB equation and developing convergent numerical algorithm.

References

1. Bansal, R. and Zhou, H. (2002). Term structure of interest rates with regime shifts. *Journal of Finance*, 57, 1997–2043.
2. Bäuerle, N. and Rieder, U. (2004). Portfolio optimization with markov-modulated stock prices and interest rates. *IEEE Trans. on Automatic Control*, 49(3), 442–447.
3. Bäuerle, N. and Rieder, U. (2007). Portfolio optimization with jumps and unobservable intensity process. *Mathematical Finance*, 17(2), 205–224.
4. Fu, J., Wei, J. and Yang, H. (2014). Portfolio optimization in a regime-switching market with derivatives. *European Journal of Operational Research*. 233, 184–192.
5. Hardy, M. (2001). A regime-switching model for long-term stock returns. *North American Actuarial Journal*, 5, 41–53.
6. Korn, R. and Kraft, H. (2001). A stochastic control approach to portfolio problems with stochastic interest rates. *SIAM J. Control and Optimization*, 40(4), 1250–1269.
7. Kraft, H. (2005). Optimal portfolios and Heston's stochastic volatility model: an explicit solution for power utility. *Quantitative Finance*, 5(3), 303–313.
8. Li, J.Z. and Wu, R. (2009). Optimal investment problem with stochastic interest rate and stochastic volatility: Maximizing a power utility. *Appl. Stochastic Models Bus. Ind.*, 25, 407–420.
9. Papanicolaou, A. and Sircar, R. (2014). A Regime-switching Heston model for VIX and S&P 500 implied volatilities. *Quantitative Finance*, 44(10), 1811–1827.
10. Rieder, U. and Bäuerle, N. (2005). Portfolio optimization with unobservable markov-modulated drift process. *Journal of Applied Probability*, 42(2), 362–378.
11. Sass, J. and Haussmann, U. (2004). Optimizing the terminal wealth under partial information: The drift process as a continuous time markov chain. *Finance Stochast.*, 8, 553–577.
12. Sotomayor, L. and Cadenillas, A. (2009). Explicit solutions of consumption-investment problems in financial markets with regime switching, *Mathematical Finance*, 19(2), 251–279.
13. Ye, C., Ren, D. and Liu, R.H. (2018). Optimal asset allocation with stochastic interest rates in regime-switching models. *Int. J. Theor. Appl. Finance*, 21(5), 1-32. DOI: 10.1142/S0219024918500322.
14. Yin, G. and Zhang, Q.: *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*. New York, Springer-Verlag (1998)
15. Zhang, X., Siu, T. and Meng, Q. (2010). Portfolio selection in the enlarged markovian regime-switching market. *SIAM J. Control and Optimization*, 48(5), 3368–3388.



On Optimal Stopping and Impulse Control with Constraint

J.L. Menaldi and M. Robin

Abstract The optimal stopping and impulse control problems for a Markov-Feller process are considered when the controls are allowed only when a signal arrives. This is referred to as control problems with constraint. In [28, 29, 30], the HJB equation was solved and an optimal control (for the optimal stopping problem, the discounted impulse control problem and the ergodic impulse control problem, respectively) was obtained, under suitable conditions, including a setting on a compact metric state space. In this work, we extend most of the results to the situation where the state space of the Markov process is locally compact.

Keywords: Markov-Feller processes, information constraints, impulse control, control by interventions, ergodic control.

AMS Subject Classification: Primary 49J40 and Secondary 60J60, 60J75.

1 Introduction

A considerable literature has been devoted to optimal stopping and impulse control of Markov processes (e.g., see the references in Bensoussan and Lions [3, 4], Bensoussan [2], Davis [10]). A relatively small part of this literature concerns problems where constraints are imposed on the admissible stopping times. In the present paper, we address optimal stopping and impulse control problems of a Markov process x_t when the stopping times must satisfy a constraint, namely, the control is allowed

J.L. Menaldi

Wayne State University, Department of Mathematics, Detroit, MI 48202, USA, e-mail: menaldi@wayne.edu

M. Robin

Fondation Campus Paris-Saclay, 91190 Saint-Aubin, France,
e-mail: maurice.robin@polytechnique.edu

to take place only at the jump times of a given process y_t , these times representing the arrival of a signal.

For instance, the system evolves according to a diffusion process x_t and the signal y_t is a Poisson process as in Dupuis and Wang [11], where an optimal stopping problem is studied, with an application to finance. In this example, thanks to the memoryless property of the exponential distribution, the y_t process does not appear as such. It is interesting to notice that, in the usual (unconstrained) case, the dynamic programming leads to the variational inequality $\max\{-Au + \alpha u, u - \psi\} = 0$, where A is the infinitesimal generator of x_t and ψ is the stopping cost (with running cost $f = 0$). However, in the constrained case, this becomes the equation $-Au + \alpha u + \lambda[u - \psi]^+ = 0$, where λ is the intensity of the Poisson process (which is assumed independent from x_t). As soon as the intervals between the jumps of y_t are not exponentially distributed, the control problem must be formulated with the couple (x_t, y_t) and the generator of this two-component process intervenes in the HJB equation.

Such problems has been studied in [28, 29, 30], when the process x_t takes values in a metric compact space E and $y_t = t - \tau_n$, where $\{\tau_n\}$ is an increasing sequence of instants such that $T_n = \tau_n - \tau_{n-1}$, for $n \geq 1$ are, conditionally to x_t , IID random variables. Using an auxiliary discrete time problem in a systematic way, some results have been obtained for optimal stopping and impulse control (with discounted and ergodic costs). Several applications of optimal stopping with constraint have been studied where the decision times are related to availability of some assets (see Lempa [23] and references therein). More generally, portfolio problems with transaction costs could give rise to impulse control with constraint. Moreover, we can consider applications in simple hybrid models (with the signal being the ‘discrete’ variable, see last section).

The main aim of the present work is to extend the previous results to the case of a locally compact Polish space, considering the three categories of problems: optimal stopping, impulse control with discounted cost as well as ergodic cost. We also mention further extensions and how some generalizations of the present model is related to hybrid models.

Without pretending to be comprehensive, let us mention (a) that references related to optimal stopping with constraint include also Liang [25] who studied particular cases of the model considered here and (b) that other class of (analogue) constraint have been considered, e.g., in Egloff and Leippold [12]. Moreover, for impulse control with constraint, we found only a few references, Brémaud [7, 8], Liang and Wei [26], and Wang [39]. A different kind of constraint is considered in Costa et al. [9], where the constraints are written as infinite horizon expected discounted costs.

The paper is organized as follows. In section 2, we introduce notations, definitions and preliminary properties of the uncontrolled process, which is the two components process (x_t, y_t) . Section 3 presents the definition of the optimal stopping problem and its solution. Section 4 describes the process controlled by impulses and the assumptions, which are used for both discounted cost and ergodic cost. In section 5, the impulse control problem with discounted cost is solved via the HJB

equation. In section 6, we present the ergodic cost problem and its solution. Some extension are mentioned in section 7 and in section 8 we discuss the links with hybrid models.

2 The Uncontrolled Process

Let us begin with some notations, definitions, comments, and preliminary properties.

Basic Notations:

- $\mathbb{R}^+ = [0, \infty[$, E a locally compact, separable and complete metric space (in short, a locally compact Polish space), and also $\mathbb{N}_0 = \{0, 1, \dots\}$ (i.e., natural numbers and 0), $\overline{\mathbb{N}}_0 = \mathbb{N}_0 \cup \{\infty\}$, $\overline{\mathbb{R}}^+ = [0, \infty]$;
- $\mathcal{B}(Z)$ the Borel σ -algebra of sets in Z , $B(Z)$ the space of real-valued Borel and bounded functions on Z , $C_b(Z)$ the space of real-valued continuous and bounded functions on Z , $C_0(Z)$ real-valued continuous functions vanishing at infinity on Z , i.e., a real-valued continuous function v belongs to $C_0(Z)$ if and only if for every $\varepsilon > 0$ there exists a compact set K of Z such that $|v(z)| < \varepsilon$ for every z in $Z \setminus K$ ¹, and also, if necessary, $B^+(Z)$, $C_b^+(Z)$, $C_0^+(Z)$ for non-negative functions; usually either $Z = E$ or $Z = E \times \mathbb{R}^+$;
- the canonical space $D(\mathbb{R}^+, Z)$ of cad-lag functions, with its canonical process $z_t(\omega) = \omega(t)$ for any $\omega \in D(\mathbb{R}^+, Z)$, and its canonical filtration $\mathbb{F}^0 = \{\mathcal{F}_t^0 : t \geq 0\}$, $\mathcal{F}_t^0 = \sigma(z_s : 0 \leq s \leq t)$.

Assumption 2.1 *Let $(\Omega, \mathbb{F}, x_t, y_t, P_{xy})$ be a (realization of a) strong and normal homogeneous Markov process, on $\Omega = D(\mathbb{R}^+, E \times \mathbb{R}^+)$ with its canonical filtration universally completed $\mathbb{F} = \{\mathcal{F}_t : t \geq 0\}$ with $\mathcal{F}_\infty = \mathcal{F}$, where (x_t, y_t) is the canonical process having values in $E \times \mathbb{R}^+$, and \mathbb{E}_{xy} (or $\mathbb{E}_{x,y}$ when a confusion may arrive) denotes the expectation relative to P_{xy} .*

- a) *It is also assumed that x_t is a Markov process by itself (referred as the reduced state), with a C_0 -semigroup $\Phi_x(t)$ (i.e., $\Phi_x(t)C_0(E) \subset C_0(E)$, $\forall t \geq 0$), and infinitesimal generator A_x with domain $\mathcal{D}(A_x) \subset C_0(E)$.*
- b) *The process y_t (referred to as the signal process) has jumps to zero at times $\tau_1, \dots, \tau_n \rightarrow \infty$ and $y_t = t - \tau_n$ for $\tau_n \leq t < \tau_{n+1}$ (i.e., τ_1 is the time of the first jump –to zero– of y_t , each jump is ‘the signal’ and y_t is exactly the ‘time elapsed since the last jump or signal’), and if $y_0 = 0$ and $\tau_0 = 0$ then it is assumed that conditionally to x_t , the intervals between jumps $T_n = \tau_n - \tau_{n-1}$ are independent, identically distributed random variables with a non-negative continuous and bounded intensity function $\lambda(x, y)$, which is such that there exists a constant $K > 0$ satisfying $\mathbb{E}_{x0}\{\tau_1\} \leq K$, for any x in E . \square*

¹ Typically $E = \mathbb{R}^d$ and this means that $v(z) \rightarrow 0$ as $|z| \rightarrow \infty$.

Remark 2.1. Actually, we begin with a realization of the reduced state process x_t on the canonical space $D(\mathbb{R}^+, E)$ and the signal process y_t is constructed based on the given intensity $\lambda(x, y)$, and this procedure yields a $C_0(E \times \mathbb{R}^+)$ -semigroup denoted by $\Phi_{xy}(t)$. Thus, in view of Palczewski and Stettner [34], all this implies that both semigroups $\Phi_x(t)$ and $\Phi_{xy}(t)$ have the Feller property, i.e., $\Phi_{xy}(t)C_b(E) \subset C_b(E)$ and $\Phi_{xy}(t)C_b(E \times \mathbb{R}^+) \subset C_b(E \times \mathbb{R}^+)$, and since only a strong and normal Markov process is assumed, the semigroup $\Phi_{xy}(t)$ is (initially) acting on $B(E \times \mathbb{R}^+)$ and so, weak (or stochastic) continuity is deduced from the assumption of a cad-lag realization, which means that

$$(x, y, t) \mapsto \mathbb{E}_{xy}\{h(x_t, y_t)\} \quad \text{is a continuous function,} \tag{1}$$

for any h in $C_b(E \times \mathbb{R}^+)$. In [28, 29, 30] a probabilistic construction of the signal process y_t was described, but there are other ways to constructing $\Phi_{xy}(t)$. For instances, begin with the process (x_t, \tilde{y}_t) with $\tilde{y}_t = y + t$ having infinitesimal generator $A^0 = A_x + \partial_y$ and a $C_0(E \times \mathbb{R}^+)$ -semigroup. Then, add the perturbation $Bh(x, y) = \lambda(x, y)[h(x, 0) - h(x, y)]$, which is a bounded operator generating a $C_0(E \times \mathbb{R}^+)$ -semigroup, with domain $\mathcal{D}(B) = C_0(E \times \mathbb{R}^+)$. Hence $A_{xy} = A^0 + B$ generates a $C_0(E \times \mathbb{R}^+)$ -semigroup, with $\mathcal{D}(A_{xy}) = \mathcal{D}(A^0)$, e.g., see Ethier and Kurtz [13, Section 1.7, pp. 37–40, Thm 7.1]. Therefore A_{xy} will also denote the weak infinitesimal generator in $C_b(E \times \mathbb{R}^+)$, in several places of the following sections. \square

Remark 2.2. Note that Assumption 2.1 (b) on the signal process y_t means, in particular, that

$$P_{x_0}\{T_n \in (t, t + dt) \mid x_s, 0 \leq s \leq t\} = \lambda(x_t, t) \exp\left(-\int_0^t \lambda(x_s, s) ds\right), \tag{2}$$

and then it is deduced that $\Phi_{xy}(t)$ has an infinitesimal generator $A_{xy} = A_x + A_y$ with

$$A_y \varphi(x, y) = \partial_y \varphi(x, y) + \lambda(x, y)[\varphi(x, 0) - \varphi(x, y)], \tag{3}$$

and recall that ∂_y denotes the derivative with respect to y , and that $\lambda \geq 0$ and $\lambda \in C_b(E \times \mathbb{R}^+)$. Moreover, using the law of T_1 as in (2) and the Feller property of (x_t, y_t) , it is also deduced that

$$(x, y) \mapsto \mathbb{E}_{xy}\{e^{-\alpha \tau_1} g(x_{\tau_1})\} \quad \text{belongs to } C_b(E \times \mathbb{R}^+), \tag{4}$$

for any g in $C_b(E)$ and any $\alpha \geq 0$. Note that if $y_0 = y$ then τ_1 is random variable independent of T_1, T_2, \dots with distribution $P_{x_0}\{T_1 \in \cdot \mid y_0 = y\}$. Furthermore, in turn, by applying Dynkin’s formula to $A_{xy}\varphi(x, y) + \alpha\varphi(x, y) = f(x, y)$, it follows that

$$(x, y) \mapsto \mathbb{E}_{xy}\left\{\int_0^{\tau_1} e^{-\alpha t} f(x_t, y_t) dt\right\} \quad \text{is in } C_b(E \times \mathbb{R}^+), \tag{5}$$

for any f in $C_b(E \times \mathbb{R}^+)$ and any $\alpha > 0$. \square

Remark 2.3. Note that because $\lambda(x, y)$ is bounded (it suffices for y near 0), there exists a constant a such that $P_{x0}\{\tau_1 \geq a > 0\} \geq a > 0$, for any x in E . Moreover, from Assumption 2.1 (b) on the signal process y_t we have

$$\mathbb{E}_{x0}\{\tau_1\} = \mathbb{E}_{x0}\left\{\int_0^\infty t \lambda(x_t, t) \exp\left(-\int_0^t \lambda(x_s, s) ds\right) dt\right\},$$

so if $\lambda(x, y) \leq k_1 < \infty$, for every $y \geq 0$, and $x \in E$, then $\mathbb{E}_{x0}\{\tau_1\} \geq a_1 = 1/k_1$. Also, the condition $\mathbb{E}_{x0}\{\tau_1\} \leq a_2$ is satisfied if, for instance $\lambda(x, y) \geq k_0 > 0$ for $y \geq y_0, x \in E$, then $a_2 = y_0 + 1/k_0$. Moreover, since $\lambda(x, y)$ is a continuous function in $E \times \mathbb{R}^+$, the continuity of $E_{xy}\{\tau_1\}$ follows. \square

Definition 2.1 (with comments). If the evolution $\dot{e} = -\alpha t$ in $[0, 1]$ is added to the homogeneous Markov process $\{(x_t, y_t) : t \geq 0\}$ then the expression

$$\{(X_n, e_n) = (x_{\tau_n}, e^{-\alpha \tau_n}), n = 0, 1, \dots\}, \tag{6}$$

with $e_0 = 1, \tau_0 = 0$ and $X_0 = x$, becomes a *homogeneous Markov chain* in $]0, 1] \times E$ with respect to the filtration $\mathbb{G} = \{\mathcal{G}_n : n = 0, 1, \dots\}$ obtained from \mathbb{F} , namely, $\mathcal{G}_n = \mathcal{F}_{\tau_n}$. Note that $\{x_{\tau_n} : n \geq 0\}$ is also a Markov chain with respect to \mathcal{G}_n . In this context, if

$$\tau = \inf\{t > 0 : y_t = 0\}, \tag{7}$$

is considered as a functional on Ω , then the *sequence of signals* (i.e., the instants of jumps for y_t) is defined by recurrence

$$\tau_{k+1} = \inf\{t > \tau_k : y_t = 0\}, \quad \forall k = 1, 2, \dots, \tag{8}$$

with $\tau_1 = \tau$, and by convenience, set $\tau_0 = 0$. Let us also mention that Remark 2.3 yields: there exists a constant a_1 such that

$$P_{x0}\{\tau \geq a_1 > 0\} \geq a_1 > 0, \quad \forall x \in E, \tag{9}$$

and by Assumption 2.1, there exists another constant $a_2 > 0$ such that

$$\mathbb{E}_{x0}\{\tau\} \leq a_2, \quad \forall x \in E. \tag{10}$$

It is also valid,

$$0 < a_1 \leq \tau(x) := \mathbb{E}_{x0}\{\tau\} \leq a_2, \quad \forall x \in E, \tag{11}$$

for some real numbers a_1, a_2 . An \mathbb{F} -stopping time $\theta > 0$ satisfying $y_\theta = 0$ when $\theta < \infty$ is called an *admissible stopping time*, in other words, if and only if there exists a discrete (i.e., $\overline{\mathbb{N}}_0$ -valued) \mathbb{G} -stopping time η such that $\theta = \tau_\eta$ with the convention that $\tau_\infty = \infty$. Moreover, if the condition $\theta > 0$ (or equivalently $\eta \geq 1$) is dropped then θ is called a *zero-admissible stopping time*. \square

3 Optimal Stopping with Constraint

This section is an extension of [28] to a locally compact space E .

3.1 Setting-up

The usual optimal stopping problems as presented above is well known, but our interest here is to restrict the stopping action (of the controller) to certain instants when a signal arrives. As discussed in the previous section, the state of the dynamic system is a homogeneous Markov process $\{(x_t, y_t) : t \in \mathbb{R}^+\}$ with values in the locally compact Polish space $E \times \mathbb{R}^+$, satisfying the Feller conditions (1). Suppose that

$$f \in C_b(E \times \mathbb{R}^+), \quad \psi \in C_b(E), \quad \alpha > 0, \quad (12)$$

where $f(x, y)$ is the running cost, $\psi(x)$ is the terminal cost, and α is the discount factor.

Thus, for any stopping time θ

$$J_{xy}(\theta, \psi) = \mathbb{E}_{xy} \left\{ \int_0^\theta e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \theta} \psi(x_\theta) \right\}, \quad (13)$$

is the cost function with the optimal cost

$$u(x, y) = \inf \{ J_{xy}(\theta, \psi) : \theta > 0, y_\theta = 0 \}, \quad (14)$$

i.e., θ is any admissible stopping time, as defined in Section 2. Also, it is defined an auxiliary problem with optimal cost

$$u_0(x, y) = \inf \{ J_{xy}(\theta, \psi) : y_\theta = 0 \}, \quad (15)$$

which provides a homogeneous Markovian model. *Since $u(x, y) = u_0(x, y)$ for any $x \in E$ and $y > 0$, it may be convenient to write $u_0(x) = u_0(x, 0)$ as long as no confusion arrives.*

Remark 3.1. Both costs $u(x, y)$ and $u_0(x, y)$ represent the optimization over all stopping times that occur when the signal arrives, the difference is that for $y = 0$ and $t=0$ (i.e., when the first signal arrives at the beginning), the control action is allowed for the optimal cost $u_0(x, 0)$, but it is not allowed for the optimal cost $u(x, 0)$, i.e., one may say that for $u(x, 0)$ the ‘controller is (so to speak) always ‘late’ (at the beginning and arriving simultaneously with the signal) and control is not possible. One may consider even an alternative situation, where with a certain probability (independently of (x_t, y_t) , for instance) the control is allowed, and therefore, the optimal cost (in the simplest case) would be a convex combination of $u(x, 0)$ and $u_0(x, 0)$. Clearly, all this comment will apply later, for the impulse control problem. \square

The Dynamic Programming Principle shows (heuristically) that

$$u(x, y) = \mathbb{E}_{xy} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} \min\{\psi, u\}(x_\tau, y_\tau) \right\}, \tag{16}$$

with $\tau = \inf\{t > 0 : y_t = 0\}$ being the first jump of y_t , and

$$\begin{aligned} u_0(x, y) &= \mathbb{E}_{xy} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} u_0(x_\tau, y_\tau) \right\}, \quad y > 0, \\ u_0(x, 0) &= \min \left\{ \mathbb{E}_{x0} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} u_0(x_\tau, y_\tau) \right\}, \psi(x) \right\}, \end{aligned} \tag{17}$$

are the corresponding Hamilton-Jacobi-Bellman (HJB) equations, which are referred to as variational inequalities (VI) in a weak form. Also, both problems are (logically) related by the condition

$$u(x, y) = \mathbb{E}_{xy} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} u_0(x_\tau, y_\tau) \right\}. \tag{18}$$

Thus, $y_\tau = 0$ implies

$$\begin{aligned} u_0(x) &= \min \left\{ \psi(x), \mathbb{E}_{x0} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} u_0(x_\tau) \right\} \right\}, \\ u(x, y) &= \mathbb{E}_{xy} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} \min\{\psi, u\}(x_\tau, 0) \right\}, \\ u(x, y) &= \mathbb{E}_{xy} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} u_0(x_\tau) \right\}, \end{aligned}$$

i.e., if $u_0(x)$ is known then the above equalities yield $u(x, y)$ and $u_0(x, y)$.

3.2 Solving the VI

By means of (6), the continuous-time cost $J_{x0}(\theta, \psi)$ with $f = 0$ and a stopping time $\theta = \tau_\eta$ can be written as

$$\begin{aligned} J_{x0}(\theta, \psi) &= \mathbb{E}_{x0} \left\{ e^{-\alpha \theta} \psi(x_\theta) \right\} \\ &= \mathbb{E} \left\{ e_\eta \psi(X_\eta) \mid e_0 = 1, X_0 = x \right\} := K_{1x}(\eta, \psi), \end{aligned} \tag{19}$$

for any discrete stopping time η relative to the Markov chain, i.e., where η has values in $\mathbb{N}_0 =$ and the convention $\tau_\infty = \infty$, and the last equality is the definition of the discrete cost $K_{1x}(\eta, \psi)$. This means that the optimal cost $u_0(x)$ is also the optimal cost of a discrete-time stopping time problem relative to the homogeneous Markov chain (certainly, there are several other ways of considering an equivalent problem in discrete-time), i.e., $u_0(x) = \inf\{K_{1x}(\eta, \psi) : \eta \geq 0\}$. This yields

$$u_0(x) = \min \{ \psi(x), \mathbb{E}_{x_0} e^{-\alpha\tau} u_0(x_\tau) \} \tag{20}$$

as the HJB equation for $u_0(x)$, when $f = 0$.

Theorem 3.1. *Under Assumption 2.1 and (12), the VI (17) and (16) have each a unique solution in $C_b(E \times \mathbb{R}^+)$, which are the optimal costs (14) and (15), respectively. Moreover, the first admissible exit time of the continuation region is optimal, i.e., the discrete stopping times*

$$\begin{aligned} \hat{\theta} &= \inf \{ t > 0 : u(x_t, y_t) \leq \psi(x_t, y_t), y_t = 0 \}, \\ \hat{\theta}_0 &= \inf \{ t \geq 0 : u_0(x_t, y_t) = \psi(x_t, y_t), y_t = 0 \} \end{aligned} \tag{21}$$

are optimal, namely, $u(x, y) = J_{xy}(\hat{\theta}, \psi)$ and $u_0(x, y) = J_{xy}(\hat{\theta}_0, \psi)$. Furthermore, the relation (18) holds. \square

Proof. This result is proved in [28] when E is compact, and it is valid under the assumptions in Section 2 with the same arguments, and therefore, only the main idea and comments are presented.

First, let us mention that the translation

$$u \longmapsto u - \mathbb{E}_{xy} \left\{ \int_0^\infty e^{-\alpha t} f(x_t, y_t) dt \right\}$$

(and similarly with u_0) reduces to a zero running cost, i.e., in all this section we may assume $f = 0$ without any loss of generality, only the terminal cost ψ is relevant. Also, Assumption 2.1(b) on the signal and the inequality

$$\begin{aligned} (1 - e^{-\alpha a}) P_{x_0} \{ \tau \geq a \} &= (1 - e^{-\alpha a}) P_{x_0} \{ 1 - e^{-\alpha\tau} \geq 1 - e^{-\alpha a} \} \\ &\leq 1 - \mathbb{E}_{x_0} \{ e^{-\alpha\tau} \}, \quad \forall a > 0, \end{aligned}$$

imply $\mathbb{E}_{x_0} \{ e^{-\alpha\tau} \} \leq 1 - (1 - e^{-\alpha a_1}) a_0 := k_1 < 1$. This is used to solve the VI

$$u_0(x) = \min \left\{ \mathbb{E}_{x_0} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha\tau} u_0(x_\tau) \right\}, \psi(x) \right\},$$

by means of a fixed point for a contraction operator. Then, some martingale arguments are used to establish that $u_0(x)$ is indeed the optimal cost of a discrete-time optimal stopping time problem relative to a Markov chain (6), where the first exit time of the continuation region $\{x : u_0(x) < \psi(x)\}$ is optimal. Next, this is connected with the continuous-time problem and the conclusion follows.

If the function $u_0(x, y)$ belongs to the domain $\mathcal{D}(A_{xy})$ then VI becomes

$$\begin{aligned} A_{xy} u_0(x, y) - \alpha u_0(x) + f(x, y) &= 0, \quad \forall (x, y) \in E \times]0, \infty[, \\ \min \{ A_{xy} u_0(x, y) - \alpha u_0(x) + f(x, y), \psi(x) - u_0(x, y) \} &= 0, \quad \forall (x, y) \in E \times \{0\}, \end{aligned}$$

where A_{xy} is the infinitesimal generator. It may be proved that this is indeed the case when ψ also belongs to $\mathcal{D}(A_{xy})$, but only continuity is usually not sufficient.

However, the optimal cost u given by (14) belongs to $\mathcal{D}(A_{xy})$ and the VI (16) is equivalent to

$$-A_{xy}u(x, y) + \alpha u(x, y) + \lambda(x, y)[u(x, 0) - \psi(x, y)]^+ = f(x, y), \tag{22}$$

for any (x, y) in $E \times [0, \infty[$, where $\lambda(x, y)$ is the jump-intensity as discussed in the previous section. Also remark $u_0 = \min\{u, \psi\}$, which makes clear that $u_0(x, y)$ may not belong to the domain $\mathcal{D}(A_{xy}) \subset C_b(E \times [0, \infty[)$.

Remark 3.2. The VI/HJB equation (22) is similar to the penalized equation of the unconstrained problem, e.g., see Bensoussan and Lions [3]. Similarly, using the same method as in the penalized problem, if λ goes to infinity (uniformly) then the solution u_λ converges to the solution (which is a function of x only) of the classical variational inequality of the unconstrained problem. \square

There are some references regarding the stopping time problem with Poisson constraint (e.g., Dupuis and Wang [11], Lempa [23], Liang and Wei [26]), while there are many more about the usual or standard stopping times problem (e.g., the books by Bensoussan and Lions [3], Peskir and Shiryaev [35], among several others books and papers).

4 Impulse Controlled Process

This section describes the controlled process and assumptions common to both, the discounted problem and the ergodic problem, as treated in the next two sections.

4.1 Controlled Process

For a detailed construction we refer to Bensoussan and Lions [4] (see also Davis [10], Lepeltier and Marchal [24], Robin [36], Stettner [38]).

Let us consider $\Omega^\infty = [D(\mathbb{R}^+; E \times \mathbb{R}^+)]^\infty$, and define $\mathcal{F}_t^0 = \mathcal{F}_t$ and $\mathcal{F}_t^{n+1} = \mathcal{F}_t^n \otimes \mathcal{F}_t$, for $n \geq 0$, where \mathcal{F}_t is the universal completion of the canonical filtration as previously.

An *arbitrary impulse control* v (not necessarily admissible at this stage) is a sequence $(\theta_n, \xi_n)_{n \geq 1}$, where θ_n is a stopping time of \mathcal{F}_t^{n-1} , $\theta_n \geq \theta_{n-1}$, and the impulse ξ_n is $\mathcal{F}_{\theta_n}^{n-1}$ measurable random variable with values in E .

The coordinate in Ω^∞ has the form $(x_t^0, y_t^0, x_t^1, y_t^1, \dots, x_t^n, y_t^n, \dots)$, and for any impulse control v there exists a probability P_{xy}^v on Ω^∞ such that the evolution of the controlled process (x_t^y, y_t^y) is given by the coordinates (x_t^n, y_t^n) of Ω^∞ when $\theta_n \leq t < \theta_{n+1}$, $n \geq 0$ (setting $\theta_0 = 0$), i.e., $(x_t^y, y_t^y) = (x_t^n, y_t^n)$ for $\theta_n \leq t < \theta_{n+1}$. Note that clearly (x_t^y, y_t^y) is defined for any $t \geq 0$, but (x_t^n, y_t^n) is only used for any $t \geq \theta_i$, and $(x_{\theta_i}^{i-1}, y_{\theta_i}^{i-1})$ is the state at time θ_i just before the impulse (or jump) to

$(\xi_i, \mathcal{Y}_{\theta_i}^{i-1}) = (x_{\theta_i}^i, \mathcal{Y}_{\theta_i}^i)$, as long as $\theta_i < \infty$. For the sake of simplicity, we will not always indicate, in the sequel, the dependency of (x_t^v, \mathcal{Y}_t^v) with respect to v . A *Markov impulse control* v is identified by a closed subset S of $E \times \mathbb{R}^+$ and a Borel measurable function $(x, y) \mapsto \xi(x, y)$ from S into $C = E \times \mathbb{R}^+ \setminus S$, with the following meaning: intervene only when the the process (x_t, y_t) is leaving the continuation region C and then apply an impulse $\xi(x, y)$, while in the stopping region S , moving back the process to the continuation region C , i.e., $\theta_{i+1} = \inf\{t > \theta_i : (x_t^i, \mathcal{Y}_t^i) \in S\}$, with the convention that $\inf\{\emptyset\} = \infty$, and $\xi_{i+1} = \xi(x_{\theta_{i+1}}^i, \mathcal{Y}_{\theta_{i+1}}^i)$, for any $i \geq 0$, as long as $\theta_i < \infty$.

Now, the admissible controls are defined as follows, recalling that τ_n are the arrival times of signal

Definition 4.1. (i) As mentioned earlier, a stopping time θ is called ‘admissible’ if almost surely there exists $n = \eta(\omega) \geq 1$ such that $\theta(\omega) = \tau_{\eta(\omega)}(\omega)$, or equivalently if θ satisfies $\theta > 0$ and $y_\theta = 0$ a.s.

(ii) An impulse control $v = \{(\theta_i, \xi_i), i \geq 1\}$ as above is called ‘admissible’, if each θ_i is admissible (i.e., $\theta_i > 0$ and $y_{\theta_i} = 0$), and $\xi_i \in \Gamma(x_{\theta_i}^{i-1})$. The set of admissible impulse controls is denoted by \mathcal{V} .

(iii) If $\theta_1 = 0$ is allowed, then v is called ‘zero-admissible’. The set of zero-admissible impulse controls is denoted by \mathcal{V}_0 .

(iv) An ‘admissible Markov’ impulse control corresponds to a stopping region $S = S_0 \times \{0\}$ with $S_0 \subset E$, and an impulse function satisfying $\xi(x, 0) = \xi_0(x) \in \Gamma(x)$, for any $x \in S_0$, and therefore, $\theta_i = \tau_{\eta_i}^i$ and $\eta_{i+1} = \inf\{k > \eta_i : x_{\tau_k}^i \in S_0\}$, with $\tau_0^0 = 0$, $\tau_k^i = \inf\{t > \tau_{k-1}^i : y_t^i = 0\}$, for any $k \geq i \geq 1$. \square

The discrete time impulse control problem has been consider in Bensoussan [2], Stettner [37]. As seen later, it will be useful to consider an auxiliary problem in discrete time, for the Markov chain $X_n = x_{\tau_n}$, with the filtration $\mathbb{G} = \{\mathcal{G}_n, n \geq 0\}$, $\mathcal{G}_n = \mathcal{F}_{\tau_n}^{n-1}$. The impulses occurs at the stopping times η_k with values in the set $\mathbb{N} = \{0, 1, 2, \dots\}$ and are related to θ_k by $\eta_i = \inf\{k \geq 1 : \theta_k = \tau_k\}$ for admissible controls $\{\theta_k\}$ and similarly for zero-admissible controls. Thus,

Definition 4.2. If $v = \{(\eta_i, \xi_i), i \geq 1\}$ is a sequence of \mathbb{G} -stopping times and \mathcal{G}_{η_i} -measurable random variables ξ_i , with $\xi_i \in \Gamma(x_{\tau_{\eta_i}})$, η_i increasing and $\eta_i \rightarrow +\infty$ a.s., then v is referred to as an ‘admissible discrete time’ impulse control if $\eta_1 \geq 1$. If $\eta_i \geq 0$ is allowed, this is referred as an ‘zero-admissible discrete time’ impulse control. \square

4.2 Common Assumptions

It is assumed that there are a running cost $f(x, y)$ and a cost-per-impulse $c(x, \xi)$ satisfying

$$\begin{aligned}
 f : E \times \mathbb{R}^+ &\rightarrow \mathbb{R}^+ \text{ bounded and continuous, } \alpha > 0, \\
 c : E \times E &\rightarrow [c_0, +\infty[, c_0 > 0, \text{ bounded and continuous,}
 \end{aligned}
 \tag{23}$$

where the discount factor is not used within the ergodic contest. Moreover, for any $x \in E$, the possible impulses must be in $\Gamma(x) = \{\xi \in E : (x, \xi) \in \Gamma\}$, where Γ is a given analytic set in $E \times E$ such that for every x in E the following properties hold true

$$\begin{aligned} \emptyset \neq \Gamma(x) \text{ is compact}^2, \quad \forall \xi \in \Gamma(x), \Gamma(\xi) \subset \Gamma(x), \quad \text{and} \\ c(x, \xi) + c(\xi, \xi') \geq c(x, \xi'), \quad \forall \xi \in \Gamma(x), \forall \xi' \in \Gamma(\xi) \subset \Gamma(x). \end{aligned} \tag{24}$$

Finally, defining the operator M

$$Mv(x) = \inf_{\xi \in \Gamma(x)} \{c(x, \xi) + v(\xi)\}, \tag{25}$$

it is assumed that

$$\begin{aligned} M \text{ maps } C_b(E) \text{ into } C_b(E), \text{ and there exists a measurable} \\ \text{selector } \hat{\xi}(x) = \hat{\xi}(x, v) \text{ realizing the infimum in } Mv(x), \forall x, v. \end{aligned} \tag{26}$$

Remark 4.1. (a) The last condition in (24) is to ensure that simultaneous impulses is never optimal. (b) (26) requires some regularity property of $\Gamma(x)$, e.g., see Davis [10]. (c) It is possible (but not necessary) that x belongs to $\Gamma(x)$, actually, even $\Gamma(x) = E$ whenever E is compact. However, an impulse occurs when the system moves from a state x to another state $\xi \neq x$, i.e., it suffices to avoid (or not to allow) impulses that moves x to itself, since they have a higher cost. \square

5 Discounted Cost

This section is an extension of [29] to a locally compact space E .

5.1 HJB Equation

The *discounted* cost of an impulse control (or policy) $v = \{(\theta_i, \xi_i) : i \geq 1\}$ is given by

$$J_{x,y}(v) = \mathbb{E}_{x,y}^v \left\{ \int_0^\infty e^{-\alpha t} f(x_t, y_t) dt + \sum_{i=0}^\infty e^{-\alpha \theta_i} c(x_{\theta_i}^{i-1}, \xi_i) \right\}, \tag{27}$$

where $\mathbb{E}_{x,y}^v$ is the $P_{x,y}^v$ -expectation of the process under the impulse control v with initial conditions $(x_0, y_0) = (x, y)$, and $x_{\theta_i}^{i-1}$ is the value of the process just before the impulse. Note that the process $\{y_t : t \geq 0\}$ is not subject to any impulse, and the condition $y_\theta = 0$ determines admissibility of the impulse time θ .

² compactness is not really necessary, but it is convenient

Thus, the optimal cost is defined by

$$u(x, y) = \inf \{ J_{x,y}(v) : v \in \mathcal{V} \}, \quad \forall (x, y) \in E \times [0, \infty[, \quad (28)$$

and its associated auxiliary impulse control problem (referred to as the ‘time-homogeneous’ impulse control) with an optimal cost given by

$$u_0(x, y) = \inf \{ J_{x,y}(v) : v \in \mathcal{V}_0 \}, \quad \forall (x, y) \in E \times [0, \infty[. \quad (29)$$

As with the optimal stopping time problems, since $u(x, y) = u_0(x, y)$ for any $x \in E$ and $y > 0$, it may be convenient to write $u_0(x) = u_0(x, 0)$ as long as no confusion arrives.

The Dynamic Programming Principle shows (heuristically), see [29, Section 3] that

$$u(x, y) = \mathbb{E}_{xy} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} \min \{ Mu, u \}(x_\tau, y_\tau) \right\}, \quad (30)$$

and

$$\begin{aligned} u_0(x, y) &= \mathbb{E}_{xy} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} u_0(x_\tau, y_\tau) \right\}, \quad y > 0, \\ u_0(x) &= \min \left\{ \mathbb{E}_{x0} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} u_0(x_\tau) \right\}, Mu_0(x) \right\}, \end{aligned} \quad (31)$$

are the corresponding Hamilton-Jacobi-Bellman (HJB) equations, which are referred to as quasi-variational inequalities (QVI) in a weak form. Note that M is an operator in the variable x alone, so that $Mu(x, y) = [Mu(\cdot, y)](x)$. In any case, $\min \{ Mu, u \}(x_\tau, y_\tau) = \min \{ Mu, u \}(x_\tau, 0)$, because $y_\tau = 0$. Also, both problems are related (logically) by the condition

$$u(x, y) = \mathbb{E}_{xy} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} u_0(x_\tau) \right\}, \quad (32)$$

and so, if $u_0(x)$ is known then the last equality yields $u(x, y)$ and $u_0(x, y)$. The optimal cost $u_0(x)$ can be expressed as a discrete-time optimal impulse control similar to Bensoussan [2, Chapter 8, 89–132] (ignoring the constraint), but this not necessary for the analysis, since everything is based on the results obtained for the optimal stopping time problems discussed in section 3.

5.2 Solving the QVI

Define

$$u^0(x, y) = \mathbb{E}_{xy} \left\{ \int_0^\infty e^{-\alpha t} f(x_t, y_t) dt \right\}, \quad \forall (x, y) \in E \times \mathbb{R}^+, \quad (33)$$

This function u^0 is the cost of no intervention, i.e., when the controller choose not to apply any impulse to the system. Since all cost are supposed nonnegative, the interval

$$C_b(u^0, Z) = \{v \in C_b(E \times \mathbb{R}^+) : 0 \leq v \leq u^0\}, \tag{34}$$

for either $Z = E \times \mathbb{R}^+$ or $Z = E$, contains the optimal cost either u or u_0 , given by either (28) or (29).

To find a solution to the QVIs (30) and (31) set $u^0 = u_0^0 = u^0$ and consider the schemes

$$u^n(x, y) = \mathbb{E}_{xy} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} \min\{Mu^{n-1}, u^n\}(x_\tau, 0) \right\},$$

$$u_0^n(x) = \min \left\{ \mathbb{E}_{x0} \left\{ \int_0^\tau e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \tau} u_0^n(x_\tau) \right\}, Mu_0^{n-1}(x) \right\},$$

for $n \geq 1$, i.e., a sequence of optimal stopping times problems with constraint. Based on Theorem 3.1, each VI has a unique solution either $u(x, y)$ in $C_b(E \times \mathbb{R}^+)$ or u_0^n in $C_b(E)$ satisfying either/or

$$u^n(x, y) = \inf_\theta \mathbb{E}_{xy} \left\{ \int_0^\theta e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \theta} Mu^{n-1}(x_\theta, 0) \right\}, \tag{35}$$

$$u_0^n(x) = \inf_\theta \mathbb{E}_{x0} \left\{ \int_0^\theta e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \theta} Mu_0^{n-1}(x_\theta) \right\},$$

where the minimization is over all admissible (or zero-admissible) stopping times θ .

As in [29, Thms 4.2 and 4.3], we have

Theorem 5.1. *Let us suppose Assumption 2.1 and (23), (24), (26). Then each of the sequences of functions $\{u_0^n\}$ and $\{u^n\}$ defined above, is monotone decreasing to the unique solution u in $C_b(u^0, E \times \mathbb{R}^+)$ and the solution u_0 in $C_b(u^0, E)$, of the QVIs (30) and (31). Moreover, the estimate: there exist constants $C > 0$, $0 < r < 1$ such that*

$$|u^n(x, y) - u(x, y)| + |u_0^n(x, y) - u_0(x, y)| \leq Cr^n, \quad \forall (x, y) \in E \times \mathbb{R}^+,$$

for all $n \geq 1$, as well as the relations (32),

$$u(x, y) = \inf_\theta \mathbb{E}_{xy} \left\{ \int_0^\theta e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \theta} Mu(x_\theta, 0) \right\},$$

$$u_0(x) = \inf_\theta \mathbb{E}_{x0} \left\{ \int_0^\theta e^{-\alpha t} f(x_t, y_t) dt + e^{-\alpha \theta} Mu_0(x_\theta) \right\},$$

hold true, where the minimization is over (zero-)admissible stopping times θ . Furthermore, u^n and u belong to the domain $\mathcal{D}(A_{x,y}) \subset C_b(E \times [0, \infty[)$ of the infinitesimal generator $A_{x,y}$, and $u(x, y)$

$$\begin{aligned}
 -A_{x,y}u(x,y) + \alpha u(x,y) + \lambda(x,y) [u(x,0) - (Mu(\cdot,0))(x)]^+ &= \\
 &= f(x,y), \quad \forall (x,y) \in E \times \mathbb{R}^+, \\
 -A_{x,y}u^n(x,y) + \alpha u(x,y) + \lambda(x,y) [u^n(x,0) - (Mu^{n-1}(\cdot,0))(x)]^+ &= \\
 &= f(x,y), \quad \forall (x,y) \in E \times \mathbb{R}^+, \forall n \geq 1,
 \end{aligned}$$

are equivalent to the corresponding QVI and VI.

Proof. Only a short idea of the main points in the proof are mentioned. First, a decreasing and concave mapping is defined with the expressions in (35), and following an argument similar to the one used in Hanouzet and Joly [16], the exponential convergence/estimate is proved and a fixed point (solving the QVIs) is obtained. At this point, the remaining assertions are obtained with a little more work.

In the following Theorem, all assertions are written for the optimal cost (28), but a similar result holds true for the other optimal cost (29), with the zero-admissible impulse controls.

Theorem 5.2. *Under the assumptions as in Theorem 5.1, the unique solution of the QVI equation (30) is the optimal cost (28), i.e., $u(x,y) = \inf \{J_{x,y}(v) : v \in \mathcal{V}\}$, for every (x,y) in $E \times \mathbb{R}^+$. Moreover, the first admissible exit time of the continuation region provides an optimal impulse control.*

Proof. The arguments are the same as in [29, Thms 4.4 & 4.5], there are no changes in assuming only E locally compact (instead of compact), only the compactness of $\Gamma(x)$ is necessary. Most of the discussion involves some martingale properties.

Note that if u is the optimal cost then (1) the continuation region $[u < Mu]$ is defined as all (x,y) in $E \times \mathbb{R}^+$ such that $u(x,y) < Mu(x,0)$, (2) the optimal jump-to is a Borel minimizer $\hat{\xi}(x)$ of $Mu(x,0)$, i.e., $x \mapsto \hat{\xi}(x)$ is a Borel functions from E into $\Gamma(x)$ and $c(x, \hat{\xi}(x)) + u(\hat{\xi}(x), 0) = Mu(x,0)$, for every x in E ., and (3) the first exit time of $[u < Mu]$ is defined as

$$\hat{\theta}(x,y,s) = \inf \{t > s : u(x_{t-s}, y_{t-s}) = Mu(x_{t-s}, 0), \quad y_{t-s} = 0\},$$

and $\hat{\theta}(x,y,s) = \infty$ if $u(x_{t-s}, y_{t-s}) < Mu(x_{t-s}, 0)$ for every $t > s$ such that $y_t = 0$. Note that the Markov process $t \mapsto (x_{t-s}, y_{t-s})$, for $t \geq s$, represents the initial condition $(x_s, y_s) = (x,y)$. Moreover, the continuity ensures that

$$u(x_{\hat{\theta}(x,y,s)-s}, 0) = c(x_{\hat{\theta}(x,y,s)-s}, \hat{\xi}(x_{\hat{\theta}(x,y,s)-s})) + u(\hat{\xi}(x_{\hat{\theta}(x,y,s)-s}), 0),$$

whenever $\hat{\theta}(x,y,s) < \infty$.

Therefore, the evolution under the above feedback (or Markov impulse control as in Definition 4.1-iv) and initial conditions (x,y) is as follows:

- (1) first $\theta_1 = \hat{\theta}(x,y,0)$ and $\xi_1 = \hat{\xi}(x_{\theta_1})$ when $\theta_1 < \infty$ (we may use an isolated ‘coffin’ state ∂ to set $x_\infty = \partial$ and $\hat{\xi}(\partial) = \partial$),
- (2) next $\theta_{k+1} = \hat{\theta}(\xi_k, 0, \vartheta_k)$, for any $k \geq 1$.

This is an *optimal* admissible impulse control $\hat{v} = \{(\theta_k, \xi_k) : k \geq 1\}$, which is proved in the same way as for the case E compact.

6 Ergodic Cost

This section is an extension of [30] to a locally compact space E .

6.1 Setting-up

We define the average cost to be minimized, as

$$\begin{aligned}
 J^T(0, x, y, v) &= \mathbb{E}_{xy}^v \left\{ \int_0^T f(x_s^v, y_s^v) ds + \sum_i \mathbb{1}_{\theta_i \leq T} c(x_{\theta_i}^{i-1}, \xi_i) \right\}, \\
 J(x, y, v) &= \liminf_{T \rightarrow \infty} \frac{1}{T} J^T(0, x, y, v),
 \end{aligned}
 \tag{36}$$

the ergodic control problem is to characterize

$$\mu(x, y) = \inf_{v \in \mathcal{V}} J(x, y, v),
 \tag{37}$$

and to find an optimal control. The auxiliary problem is concerned with

$$\begin{aligned}
 \mu_0(x, y) &= \inf_{v \in \mathcal{V}_0} \tilde{J}(x, y, v), \quad \text{with} \\
 \tilde{J}(x, y, v) &= \liminf_{n \rightarrow \infty} \frac{1}{\mathbb{E}_{xy}^v \{\tau_n\}} J^{\tau_n}(0, x, y, v),
 \end{aligned}
 \tag{38}$$

and $J^{\tau_n}(0, x, y, v)$ as in (36) with $T = \tau_n$. Actually, as seen later, $\mu(x, y) = \mu_0(x, y)$ is a constant.

The Dynamic Programming Principle shows (heuristically, see [30, Section 3]) that, with $w_0(x) = w_0(x, 0)$,

$$\begin{aligned}
 w_0(x) &= \min \left\{ \mathbb{E}_{x0} \left\{ \int_0^\tau [f(x_t, y_t) - \mu_0] dt + w_0(x_\tau) \right\}, M w_0(x) \right\}, \\
 w_0(x, y) &= \mathbb{E}_{xy} \left\{ \int_0^\tau [f(x_t, y_t) - \mu_0] dt + w_0(x_\tau) \right\},
 \end{aligned}
 \tag{39}$$

are the corresponding Hamilton-Jacobi-Bellman (HJB) equations in a weak form with two unknowns μ_0 and w_0 . Note that M is an operator in the variable x alone, so that $M w_0(x, y) = [M w_0(\cdot, y)](x)$ as given by (25). Also, both problems are related (logically) by the condition

$$w(x, y) = \mathbb{E}_{xy} \left\{ \int_0^\tau [f(x_t, y_t) - \mu_0] dt + w_0(x_\tau) \right\}, \quad (40)$$

and so, if $w_0(x)$ is known then the last/first equality yields $w(x, y)$ and $w_0(x, y)$. Recall that τ is defined by (7) and that since $w(x, y) = w_0(x, y)$ for any $x \in E$ and $y > 0$, it may be convenient to write $w_0(x) = w_0(x, 0)$ as long as no confusion arrives. Note that the functions $w(x, y)$ and $w_0(x)$ may be called *potentials*, and a priori, they are not *costs*, but they are used to determine an optimal control.

6.2 Solving the HJB

An important point to mention is to remark that the HJB equation (39) is equivalent to

$$w_0(x) = \min \{ M w_0(x), \ell(x) - \mu_0 \tau(x) + P w_0(x) \}, \quad (41)$$

where

$$\ell(x) = \mathbb{E}_{x0} \left\{ \int_0^\tau f(x_s, y_s) ds \right\}, \quad \tau(x) = \mathbb{E}_{x0} \{ \tau \}, \quad (42)$$

with τ as in (7), and in view of the property (4),

$$P h(x) = \mathbb{E}_{x0} \{ h(x_\tau) \}, \quad (43)$$

defines the operator P on $C_b(E)$. Note that (10) yields

$$0 \leq \ell(x) \leq a_2 \|f\|. \quad (44)$$

Moreover, from the Feller property of x_t and the law of τ , it follows that $\ell(x)$ is continuous.

In addition to the hypotheses of Sections 2 and 4, we assume that there exists a positive measure m on E such that

$$m(E) > 0 \quad \text{and} \quad P(x, U) \geq m(U), \quad \forall U \in \mathcal{B}(E), \quad (45)$$

where $P(x, U) = \mathbb{E}_{x0} \mathbb{1}_U(x_\tau)$, with τ defined by (7), and $\mathcal{B}(E)$ is the Borel σ -algebra on E .

Remark 6.1. From

$$P(x, U) = \mathbb{E}_{x0} \left\{ \int_0^\infty \lambda(x_t, t) \exp \left(- \int_0^t \lambda(x_s, s) ds \right) \mathbb{1}_U(x_t) dt \right\}.$$

and Remark 2.3, one can check that (45) is satisfied when the transition probability of x_t has a density with respect to a probability on E satisfying: for every $\varepsilon > 0$ there exists $k(\varepsilon)$ such that

$$p(x, t, x') \geq k(\varepsilon) > 0, \text{ on } E \times [\varepsilon, \infty[\times E. \tag{46}$$

This is the case, for instance, for periodic diffusion processes, see Bensoussan [1], and for reflected diffusion processes with jumps, see Garroni and Menaldi [14, 15] (which is also valid for reflected diffusion processes without jumps). Furthermore, a simple example for E locally compact is provided by a pure jump process with generator

$$A_x g(x) = b(x) \left\{ \int_E g(z) q(x, dz) - g(x) \right\}.$$

One can check that (45) is satisfied if, for instance, $0 < k_0 \leq \lambda(x, y) \leq k_1$, $0 < b_1 \leq b(x) \leq b_2$, $q(x, B) \geq m_0(B)$ for a positive measure m_0 , with $m_0(E) > 0$. \square

Lemma 6.1. *Under assumption (45), there exist a positive measure γ on E , and a constant $0 < \beta < 1$ such that $P(x, B) \geq \tau(x)\gamma(B)$, for every $B \in \mathcal{B}(E)$, any $x \in E$, with $\tau(x)\gamma(E) > 1 - \beta$. \square*

Theorem 6.1. *Under Assumption 2.1 and (23), (24), (26), as well as (45), there exists a solution (μ_0, w_0) in $\mathbb{R}^+ \times C_b(E)$ of (41), and therefore, of (39). \square*

For details of the Lemma 6.1 and Theorem 6.1 proofs, note that Kurano [21, 22] results hold true for a locally compact space E , and refer to [30, Lem 4.1 and Thm 4.2]. For instance, the assumptions (45) and (11) imply

$$P\mathbb{1}_B(x) =: P(x, B) \geq \tau(x)\gamma(x), \quad \forall B \in \mathcal{B}(E),$$

with $\gamma(B) = m(B)/a_2$ and any β in $]0, 1[$ such that $1 - \beta < m(E)a_1/a_2$. Now, the HJB equation (41) can be written as

$$w_0(x) = \inf_{\xi \in \Gamma(x) \cup \{x\}} \left\{ \ell(\xi) + \mathbb{1}_{\xi \neq x} c(x, \xi) - \mu_0 \tau(\xi) + Pw_0(\xi) \right\}.$$

Since $P'(x, dz) := P(x, dz) - \tau(x)\gamma(dz)$ satisfies $P'(x, E) < \beta < 1$, the operator

$$Rv(x) = \inf_{\xi \in \Gamma(x) \cup \{x\}} \left\{ \ell(\xi) + \mathbb{1}_{\xi \neq x} c(x, \xi) + Pw_0(\xi) - \tau(\xi) \int_E v(z)\gamma(dz) \right\}$$

is a contraction on $C_b(E)$ having a unique fixed point w_0 , and moreover, $w_0 \geq 0$ because $\ell(x) \geq 0$ and $c(x, \xi) > 0$. Thus, (μ_0, w_0) is a solution, where $\mu_0 := \gamma(w_0)$, the integral of w_0 with respect to $\gamma(\cdot)$ on E .

Remark 6.2. When λ does not depends on x , the function $\tau(x)$ is constant and (41) is the HJB equation of a standard discrete time impulse control problem as studied in Stettner [37, Section 4] for $\Gamma(x) = \Gamma$ fixed. \square

Then, we have

Theorem 6.2. *Under the assumptions as in Theorem 6.1, the constant μ_0 obtained in Theorem 6.1 satisfies*

$$\mu_0 = \inf \{ \tilde{J}(x, 0, \nu) : \nu \in \mathcal{V}_0 \}$$

and there exists an optimal feedback control based on the exit times of the continuation region $[w_0 < Mw_0]$.

Proof. First, by means of Theorem 6.1 when $\Gamma(x) = \{x\}$ (which means ‘no control’), we show that there exists $(j, h) \in \mathbb{R}^+ \times C_b(E)$ solution of

$$h(x) = \ell(x) - j\tau(x) + Ph.$$

Note that assumption (45) implies that P has a unique invariant probability denoted by $\zeta_0(dx)$, see the book Hernández-Lerma [17, Section 3.3, pp. 56–61].

Thus, there are two cases: $\mu_0 = j$ and $\mu_0 < j$. First, for $\mu_0 = j$, from the equation for h and the fact that $X_n = x_{\tau_n}$ is a Markov chain, we have

$$\begin{aligned} j &= \liminf_n \frac{1}{\mathbb{E}_{x_0}\{\tau_n\}} \mathbb{E}_{x_0} \left\{ \sum_{i=0}^{n-1} \ell(X_i) \right\} \\ &= \liminf_n \frac{1}{\mathbb{E}_{x_0}\{\tau_n\}} \mathbb{E}_{x_0} \left\{ \int_0^{\tau_n} f(x_t, y_t) dt \right\} = \tilde{J}(x, 0, \nu), \end{aligned}$$

with $\nu = 0$, i.e., no impulse at all. Then, as in [30, Thm 5.1] we have $\mu_0 \leq \tilde{J}(x, 0, \nu)$, for every ν in \mathcal{V}_0 , i.e., $\mu_0 \leq j$. Therefore, if $\mu_0 = j$ then

$$\mu_0 = \inf \{ \tilde{J}(x, 0, \nu) : \nu \in \mathcal{V}_0 \} = j = \tilde{J}(x, 0, 0),$$

and $\nu = 0$, i.e., ‘no impulses at all’, is optimal.

Next, the case $\mu_0 < j$ is treated as in [30, Thm 5.1], with $\tilde{w}(x) = w_0(x) - h(x)$, $\tilde{\ell}(x) = (j - \mu_0)\tau(x)$, $\tilde{w} = \min\{M\tilde{w}, \tilde{\ell} + P\tilde{w}\}$. Indeed, using the results in Bensoussan [2, Section 7.4, pp. 74–77], we show that this discrete time problem has an optimal control $\hat{\nu} = \{(\hat{\eta}_i, \hat{\xi}_i) : i \geq 1\}$ given by

$$\hat{\eta}_i = \inf \{ n \geq \hat{\eta}_{i-1} : w_0(X_n) = Mw_0(X_n) \},$$

where X_n is the controlled Markov chain and $\hat{\xi}_i = \hat{\xi}(X_{\hat{\eta}_i})$ with a measurable selector $\hat{\xi}(x)$ realizing the infimum in $Mw_0(x)$. This is translated in continuous time as $\hat{\theta}_i = \tau_{\hat{\eta}_i}$ and $\hat{\xi}_i = \hat{\xi}(x_{\hat{\theta}_i})$.

Remark 6.3. It is clear that the previous argument about (j, h) shows that the hypothesis (5.1) in our previous paper [30] is not really necessary, and therefore, it is a small improvement on it. \square

Theorem 6.3. *Under the assumptions as in Theorem 6.1, the constant μ_0 obtained in Theorem 6.1 satisfies*

$$\mu_0 = \inf \{ J(x, 0, \nu) : \nu \in \mathcal{V} \} = J(x, y, \hat{\nu}),$$

where $\hat{\nu}$ is obtained by τ -translations from the optimal control in Theorem 6.2.

Proof. (sketch) The first step is to show that $w(x, y)$ defined by (40) satisfied

$$-A_{xy}w(x, y) + \lambda(x, y)[w(x, 0) - Mw(x, 0)]^+ = f(x, y) - \mu_0$$

Actually, this is not surprising in view of the results for the discounted case, but the proof is somewhat cumbersome, see [30, Proposition 5.5].

This implies that the process

$$M_T = \int_0^T [f(x_t, y_t) - \mu_0] dt + w(x_T, y_T), \quad T \geq 0$$

is a submartingale, and the argument is completed as in [30].

Remark 6.4 (Ergodic cost: A more general ergodic assumption). The assumption (45) is not satisfied, in general, for diffusion processes in the whole space, and thus, it is perhaps, relatively restrictive. A ‘localized’ substitute for (45) could be the assumption:

(i) there exist a closed set C , an open set D , $C \subset D \subset E$, and a constant $\beta_0 \in]0, 1[$ as well as a probability m satisfying $0 < m(C) < 1 = m(D)$ and such that $P(x, B) \geq \beta_0 m(B)$, for every $B \in \mathcal{B}(E)$, any $x \in E$; and

(ii) there exist a continuous function $W : E \rightarrow [1, \infty[$, and constant $\beta \in]0, 1[$ such that PW is continuous and $PW(x) \leq \beta W(x) + \beta_0 \mathbb{1}_C \int_C W(z) m(dz)$, for every $x \in E$.

An adaptation of Jaskiewicz [19] allows us to obtain a solution (μ_0, w_0) of (41), with w_0 in the weighted-space

$$C_w(E) = \left\{ g \text{ continuous and } \sup_x \left\{ \frac{|g(x)|}{W(x)} \right\} < \infty \right\},$$

and to obtain Theorem 6.2, under some additional technical assumptions. Also, Theorem 6.3 can be obtained under the additional assumption

$$\mathbb{E}\{e^{-k_0 t} W(x_t)\} \leq W(x), \quad \forall x \in E, t > 0,$$

where $\lambda(x, y) \geq k_0 > 0$ for every x, y . A detailed analysis will be in a paper in preparation [31] together with examples satisfying the various assumptions. This analysis is based on several references (e.g., Hernández-Lerma and Lasserre [18], Meyn and Tweedie [33, 32], among others). \square

7 Extension

As in [28, 29, 30] let us mention some possible extensions:

- A variable discount factor $\alpha(x, y)$ instead of α constant, as well as a finite-horizon cost.

- Letting the discount factor $\alpha \rightarrow 0$ in the optimal discounted costs $u^\alpha(x, y)$ and $u^\alpha(x) = u^\alpha(x, 0) = u_0^\alpha(x, 0)$ we expect to obtain ergodic costs, e.g., if $\mu_\alpha = \alpha u^\alpha(x)$ and $w_0^\alpha(x) = u^\alpha(x) - u^\alpha(x_0)$ then $\mu_\alpha \rightarrow \mu_0$ and $w_0^\alpha(x) \rightarrow w_0(x)$, but this is still something to be properly shown, when $\Gamma(x)$ is not reduced to a fixed compact.
- A quantify signal, e.g, y_t has jumps back to $\{0, 1, 2\}$ instead of only $\{0\}$ with the following meaning: there are three classes of impulse controls $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \mathcal{V}_2$ that are enabled only and accordingly to the value of y_t (some more details are necessary for a convenient example). In this case, instead (7), the signals are given by the functional

$$\tau = \inf\{t > 0 : y_t \in I\}, \quad (47)$$

where a prototype is $I = \{0, 1, 2\}$. In this case, the Markov chain will include also y_{τ_n} , i.e., $(Z_n, e_n) = (x_{\tau_n}, y_{\tau_n}, e^{-\alpha\tau_n})$. We may think that as the waiting-time passes (indicated or represented by the process y_t) the necessity of ‘controlling’ increases and impulses to other regions (that previously were not allowed) becomes enabled, i.e., when $i < y_t < i + 1$ then only the class \mathcal{V}_i of impulse controls is available, which produces an impulse back to some $y = j < i + 1$. Actually, a detailed example may be needed, and this is not discussed here.

In this case, jumps should be always backward, i.e., y_t may jumps only to the values 0, 1 or 2 that are smaller that the value of y_t . Certainly, what is accomplished for three values could be applied for any finite number of values, and perhaps ‘extrapolate’ to infinite many values (as long as they are isolated values). Thus, $\psi(x, y)$ makes sense for the optimal stopping time problem (without any changes!) but the analysis within the impulse control could give some interesting surprises.

- For stopping time problems, recall that several extensions are possible, in particular the use of data with polynomial growth (instead of bounded). However, there are some extra complications for the impulse control problems.

8 Hybrid Models

The state of a continuous-time hybrid model has a continuous-type variable x (with cad-lag paths) and a discrete-type variable n (with cad-lag piecewise constant paths). The ‘signal’ is represented by the ‘jumps’ of n_t , and in general, this signal enable any possible change in the setting of the model, not only the ‘possibility of controlling’ as studied in this paper (an others). The general idea is that the usual evolution of the system is described by the component x_t , and ‘once in a while’ (or under some specific conditions) a discrete transition (i.e., a jump of n_t occurs) and everything may change, and the evolution continues thereafter. With this in mind, the signal (to act, e.g., to control the system as in our model) is given by the ‘hitting time’ of a set of states S , i.e., $\tau = \inf\{t > 0 : (x_t, n_t) \in S\}$, and this set S plays the

role of a ‘set-interface’, where the continuous-type and discrete-type variables exchange information. This set-interface may be given a priori or used as part of the parameters of control. In our previous ‘control with constraint’ presentation, the discrete-type component n_t was ignored (because there are only on/off possibilities) and the continuous-type component x_t is actually composed by two parts (x_t, y_t) , as they were called, the reduced state x_t and the signal process y_t . Thus, in our model, the set-interface S is $E \times \{0\}$, the same for every n (which is ignored, as mentioned earlier).

To present the problem studied in this paper as a hybrid model the ‘details’ (a) and (b) of Assumption 2.1 are not mentioned, and instead, assumptions directly on the functional (7) and the signal (8) are imposed, e.g., at least it is assumed (9), but for ergodic cost, the condition (10) is required. Also, some continuity is needed, i.e.,

$$(x, y) \mapsto \mathbb{E}_{xy}\{e^{-\alpha\tau}\varphi(x_\tau)\} \text{ and } (x, y) \mapsto \mathbb{E}_{xy}\left\{\int_0^\tau e^{-\alpha t}f(x_t, y_t)\right\}dt \tag{48}$$

are continuous functions, for every φ in $C_b(E)$ and f in $C_b(E \times [0, \infty[)$. Most of the results in previous section are valid under these ‘more general’ assumptions, except those involving the specific form of infinitesimal generator A_y (3). To be more specific, the following results can be extended under these more general hypotheses: Theorem 3.1, without (22), for optimal stopping; Theorem 5.1 (without the formula regarding the generator), and Theorem 5.2 for the discounted cost; Theorem 6.1 and Theorem 6.2 (but not Theorem 6.3) for the ergodic cost. For instance, if we assume (9) and that signals given by (8) then define the time-interval between jumps $T_n = \tau_n - \tau_{n-1}$, which (conditionally to x_t) forms an independent, identically distributed sequence of random variables with a non-negative and bounded intensity $\Lambda(x, y)$. Hence, the initial ‘signal process’ (which is not necessarily equal to the time elapsed since the last signal) can be replaced to obtain an equivalent (in most aspects) model as the one in this paper.

Indeed, let us make an example of a similar situation, i.e., a signal process \tilde{y}_t which is not equal to the process y_t , the ‘time elapsed since the last signal’. In this example, the state is (x, \tilde{y}) , the controller is allowed to ‘control’ (via an impulse) when $\tilde{y} = 0$, however, \tilde{y}_t has

$$A_z\varphi(z) = \partial_z\varphi(z) + \Lambda(z)[q\varphi(0) + (1 - q)\varphi(z/2) - \varphi(z)],$$

as its infinitesimal generator, with $0 < q < 1$, i.e., the process z_t jumps at s_n , $\sigma_n = s_{n+1} - s_n$ are IID having an intensity $\Lambda(z)$, and at the jump-times, $z_{s_n} = 0$ with probability q and $z_{s_n} = z_{s_n-}/2$ with probability $1 - q$. In this case, the functional of interest is always the same (7), namely, $\tau = \inf\{t > 0 : \tilde{y}_t = 0\}$, with the sequence of signals $\tau_{k+1} = \inf\{t > \tau_k : \tilde{y}_t = 0\}$, $\tau_0 = 0$, which are not necessarily the sequence of jump-times $\{s_n\}$ of the process \tilde{y}_t . However, if

$$F(t) = 1 - \exp\left(-\int_0^t \Lambda(s)ds\right)$$

is the law of σ_1 for P_0 then the convolution $F^{*n}(t)$ is the law of $s_n = \sum_n \sigma_n$ and the law of $\tau = \tau_1$ (when $\tilde{y}_0 = 0$) is given by

$$P_0\{\tau \leq t\} = \sum_n F^{*n}(t)q(1-q)^{n-1} := G(t),$$

and the sequence of signals $\{\tau_k\}$ define another sequence $T_k = \tau_{k+1} - \tau_k$ of IID random variables with law $G(t)$. Hence, if we take $\lambda(t) = G'(t)/(1-G(t))$ then the control problem for (x_t, \tilde{y}_t) and $f(x)$ (i.e., independent of \tilde{y}) should be equivalent to the problem (x_t, y_t) , with y_t constructed from $\lambda(y)$, since the discrete problems are identical for (x_t, \tilde{y}_t) and (x_t, y_t) . It is clear that these considerations can be extended to a similar model (x_t, z_t)

$$A_z \varphi(z) = b(z) \partial_z \varphi(z) + \Lambda(z) \int_{\mathbb{R}^+} (\varphi(\zeta) - \varphi(z)) m(z, d\zeta),$$

under suitable assumptions on the drift b and the probability kernel $m(d\zeta, z)$.

Another kind of problem could have the constraint ‘control is allowed at any jump of z_t ’, with x_t as the reduced state process and z_t as the signal process. For this model, the condition, ‘when the process z_t jumps’ is not exactly the same as ‘when z_t vanishes’. In other words, technically speaking, the full state of the system needs something else than the knowledge of (x, z) , i.e., we need to know z_t and z_{t-} to check if a jump has really occurred. Thus, if $z_t = \tilde{y}_t$ above, then we would have $\tau_n = s_n$, the jump-times of z_t . For the (x_t, y_t) model (as well as for the hybrid model) presented in the above sections, the constraint ‘control is allowed only ...’ ‘when y_t jumps’ is exactly the same as saying ‘when y_t vanishes’. Nevertheless, we may have an infinitesimal generator like A_z (of the piecewise deterministic process z_t – or something else–) with a $b(z) > 0$ and $m(\varphi, z) = \varphi(0)$, which is not exactly the process y_t (the time elapsed since the last signal), but it has the property that $z_t = 0$ iff $y_t = 0$. Thus, for those type of processes, the constraint ‘control is allowed only when z_t vanishes’ is equivalent to ‘control is allowed only when y_t vanishes’.

Because of the particular meaning of our signal process y_t as the ‘time elapsed since last signal’, we obtain more detailed results than in the general hybrid model. Therefore, there are many generalization in various directions, e.g., in between to consecutive signals some other type of control could be allowed, signals of various types enabling particular types of controls may be given, and many other ways on how a continuous-type and a discrete-type variables may interact. Actually, much more details on the (hybrid) model are necessary to advance further in this discussion, and this is part of our book in preparation Jasso-Fuentes et al. [20], which follows some the problems discussed in Bensoussan and Menaldi [5, 6] and [27].

References

1. A. Bensoussan. *Perturbation methods in optimal control*. Gauthier-Villars, Montrouge, 1988.
2. A. Bensoussan. *Dynamic programming and inventory control*. IOS Press, Amsterdam, 2011.

3. A. Bensoussan and J.L. Lions. *Applications des inéquations variationnelles en contrôle stochastique*. Dunod, Paris, 1978.
4. A. Bensoussan and J.L. Lions. *Contrôle impulsif et inéquations quasi-variationnelles*. Gauthier-Villars, Paris, 1982.
5. A. Bensoussan and J.L. Menaldi. Hybrid control and dynamic programming. *Dynam. Contin. Discrete Impuls. Systems*, 3(4):395–442, 1997.
6. A. Bensoussan and J.L. Menaldi. Stochastic hybrid control. *J. Math. Anal. Appl.*, 249(1):261–288, 2000. Special issue in honor of Richard Bellman.
7. P. Brémaud. *Point Processes and Queues*. Springer-Verlag, New York, 1981. Martingale dynamics.
8. P. Brémaud. *Markov Chains*. Springer-Verlag, New York, 1999. Gibbs fields, Monte Carlo simulation, and queues.
9. O.L.V. Costa, F. Dufour, and A.B. Piunovskiy. Constrained and unconstrained optimal discounted control of piecewise deterministic Markov processes. *SIAM J. Control Optim.*, 54(3):1444–1474, 2016.
10. M.H.A. Davis. *Markov models and optimization*. Chapman & Hall, London, 1993.
11. P. Dupuis and H. Wang. Optimal stopping with random intervention times. *Adv. in Appl. Probab.*, 34(1):141–157, 2002.
12. D. Egloff and M. Leippold. The valuation of American options with stochastic stopping time constraints. *Appl. Math. Finance*, 16(3-4):287–305, 2009.
13. S.N. Ethier and T.G. Kurtz. *Markov processes*. John Wiley & Sons Inc., New York, 1986. Characterization and convergence.
14. M.G. Garroni and J.L. Menaldi. *Green functions for second order parabolic integro-differential problems*. Longman Scientific & Technical, Harlow, 1992.
15. M.G. Garroni and J.L. Menaldi. *Second order elliptic integro-differential problems*. Chapman & Hall/CRC, Boca Raton, FL, 2002.
16. B. Hanouzet and J.L. Joly. Convergence uniforme des itérés définissant la solution d’une inéquation quasi variationnelle abstraite. *C. R. Acad. Sci. Paris Sér. A-B*, 286(17):A735–A738, 1978.
17. O. Hernández-Lerma. *Adaptive Markov control processes*. Springer-Verlag, New York, 1989.
18. O. Hernández-Lerma and J.B. Lasserre. *Further topics on discrete-time Markov control processes*. Springer-Verlag, New York, 1999.
19. A. Jaskiewicz. A fixed point approach to solve the average cost optimality equation for semi-Markov decision processes with Feller transition probabilities. *Comm. Statist. Theory Methods*, 36(13-16):2559–2575, 2007.
20. H. Jasso-Fuentes, J.L. Menaldi, and M. Robin. *Hybrid Control for Markov-Feller Processes*. To appear, 2016.
21. M. Kurano. Semi-Markov decision processes and their applications in replacement models. *Journal of the Operations Research, Society of Japan*, 28(1):18–29, 1985.
22. M. Kurano. Semi-Markov decision processes with a reachable state-subset. *Optimization*, 20(3):305–315, 1989.
23. J. Lempa. Optimal stopping with information constraint. *Appl. Math. Optim.*, 66(2):147–173, 2012.
24. J.P. Lepeltier and B. Marchal. Théorie générale du contrôle impulsif markovien. *SIAM J. Control Optim.*, 22(4):645–665, 1984.
25. G. Liang. Stochastic control representations for penalized backward stochastic differential equations. *SIAM J. Control Optim.*, 53(3):1440–1463, 2015.
26. G. Liang and W. Wei. Optimal switching at poisson random intervention times. *Discrete and Continuous Dynamical Systems, Series B*. To appear, see arXiv:1309.5608v2 [math.PR].
27. J.L. Menaldi. Stochastic hybrid optimal control models. In *Stochastic models, II (Guanajuato, 2000)*, volume 16 of *Aportaciones Mat. Investig.*, pages 205–250. Soc. Mat. Mexicana, México, 2001.
28. J.L. Menaldi and M. Robin. On some optimal stopping problems with constraint. *SIAM J. Control Optim.*, 54(5):2650–2671, 2016.

29. J.L. Menaldi and M. Robin. On some impulse control problems with constraint. *SIAM J. Control Optim.*, 55(5):3204–3225, 2017.
30. J.L. Menaldi and M. Robin. On some ergodic impulse control problems with constraint. *SIAM J. Control Optim.*, 56(4):2690–2711, 2018.
31. J.L. Menaldi and M. Robin. Remarks on ergodic impulse control problems with constraint. *In preparation*, 2019.
32. S.P. Meyn and R.L. Tweedie. Stability of Markovian processes iii. *Advances in Applied Probability*, 25(3):518–548, 1993. Foster-Lyapunov criteria for continuous time processes.
33. S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, London, 2nd edition, 2009.
34. J. Palczewski and L. Stettner. Finite horizon optimal stopping of time-discontinuous functionals with applications to impulse control with delay. *SIAM J. Control Optim.*, 48(8):4874–4909, 2010.
35. G. Peskir and A. Shiryaev. *Optimal stopping and free-boundary problems*. Birkhäuser Verlag, Basel, 2006.
36. M. Robin. Contrôle impulsionnel des processus de Markov. Thèse d'état, 1978, 353 pp. Link at <https://hal.archives-ouvertes.fr/tel-00735779/document>.
37. L. Stettner. Discrete time adaptive impulsive control theory. *Stochastic Process. Appl.*, 23(2):177–197, 1986.
38. L. Stettner. On ergodic impulsive control problems. *Stochastics*, 18(1):49–72, 1986.
39. H. Wang. Some control problems with random intervention times. *Advances in Applied Probability*, 33(2):404–422, 2001.



Linear-Quadratic McKean-Vlasov Stochastic Differential Games*

Enzo Miller Huyên Pham

Abstract We consider a multi-player stochastic differential game with linear McKean-Vlasov dynamics and quadratic cost functional depending on the variance and mean of the state and control actions of the players in open-loop form. Finite and infinite horizon problems with possibly some random coefficients as well as common noise are addressed. We propose a simple direct approach based on weak martingale optimality principle together with a fixed point argument in the space of controls for solving this game problem. The Nash equilibria are characterized in terms of systems of Riccati ordinary differential equations and linear mean-field backward stochastic differential equations: existence and uniqueness conditions are provided for such systems. Finally, we illustrate our results on a toy example.

1 Introduction

1.1 General introduction-Motivation

The study of large population of interacting individuals (agents, computers, firms) is a central issue in many fields of science, and finds numerous relevant applications in economics/finance (systemic risk with financial entities strongly interconnected), sociology (regulation of a crowd motion, herding behavior, social networks), physics, biology, or electrical engineering (telecommunication). Rationality in the behavior of the population is a natural requirement, especially in social sci-

Enzo Miller

LPSM, University Paris Diderot, e-mail: enzo.miller@polytechnique.edu

Huyên Pham

University Paris Diderot and CREST-ENSAE, e-mail: pham@lpsm.paris

* This work is supported by FiME (Finance for Energy Market Research Centre) and the “Finance et Développement Durable - Approches Quantitatives” EDF - CACIB Chair.

ences, and is addressed by including individual decisions, where each individual optimizes some criterion, e.g. an investor maximizes her/his wealth, a firm chooses how much to produce outputs (goods, electricity, etc) or post advertising for a large population. The criterion and optimal decision of each individual depend on the others and affect the whole group, and one is then typically looking for an equilibrium among the population where the dynamics of the system evolves endogenously as a consequence of the optimal choices made by each individual. When the number of indistinguishable agents in the population tend to infinity, and by considering cooperation between the agents, we are reduced in the asymptotic formulation to a McKean-Vlasov (McKV) control problem where the dynamics and the cost functional depend upon the law of the stochastic process. This corresponds to a Pareto-optimum where a social planner/influencer decides of the strategies for each individual. The theory of McKV control problems, also called mean-field type control, has generated recent advances in the literature, either by the maximum principle [5], or the dynamic programming approach [14], see also the recent books [3] and [6], and the references therein, and linear quadratic (LQ) models provide an important class of solvable applications studied in many papers, see, e.g., [15], [11], [10], [2].

In this paper, we consider multi-player stochastic differential games for McKean-Vlasov dynamics. This corresponds and is motivated by the competitive interaction of multi-population with a large number of indistinguishable agents. In this context, we are then looking for a Nash equilibrium among the multi-class of populations. Such problem, sometimes refereed to as mean-field-type game, allows to incorporate competition and heterogeneity in the population, and is a natural extension of McKean-Vlasov (or mean-field-type) control by including multiple decision makers. It finds natural applications in engineering, power systems, social sciences and cybersecurity, and has attracted recent attention in the literature, see, e.g., [1], [7], [8], [4]. We focus more specifically on the case of linear McKean-Vlasov dynamics and quadratic cost functional for each player (social planner). Linear Quadratic McKean-Vlasov stochastic differential game has been studied in [9] for a one-dimensional state process, and by restricting to closed-loop control. Here, we consider both finite and infinite horizon problems in a multi-dimensional framework, with random coefficients for the affine terms of the McKean-Vlasov dynamics and random coefficients for the linear terms of the cost functional. Moreover, controls of each player are in open-loop form. Our main contribution is to provide a simple and direct approach based on weak martingale optimality principle developed in [2] for McKean-Vlasov control problem, and that we extend to the stochastic differential game, together with a fixed point argument in the space of open-loop controls, for finding a Nash equilibrium. The key point is to find a suitable ansatz for determining the fixed point corresponding to the Nash equilibria that we characterize explicitly in terms of systems of Riccati ordinary differential equations and linear mean-field backward stochastic differential equations: existence and uniqueness conditions are provided for such systems.

The rest of this paper is organized as follows. We continue Section 1 by formulating the Nash equilibrium problem in the linear quadratic McKean-Vlasov finite horizon framework, and by giving some notations and assumptions. Section 2

presents the verification lemma based on weak submartingale optimality principle for finding a Nash equilibrium, and details each step of the method to compute a Nash equilibrium. We give some extensions in Section 3 to the case of infinite horizon and common noise. Finally, we illustrate our results in Section 4 on some toy example.

1.2 Problem formulation

Let $T > 0$ be a finite given horizon. Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a fixed filtered probability space where $\mathbb{F} = (\mathcal{F}_t)_{t \in [0, T]}$ is the natural filtration of a real Brownian motion $W = (W_t)_{t \in [0, T]}$. In this section, for simplicity, we deal with the case of a single real-valued Brownian motion, and the case of multiple Brownian motions will be addressed later in Section 3. We consider a multi-player game with n players, and define the set of admissible controls for each player $i \in \llbracket 1, n \rrbracket$ as:

$$\mathcal{A}^i = \left\{ \alpha_i : \Omega \times [0, T] \rightarrow \mathbb{R}^{d_i} \text{ s.t. } \alpha_i \text{ is } \mathbb{F}\text{-adapted and } \int_0^T e^{-\rho t} \mathbb{E}[\|\alpha_{i,t}\|^2] dt < \infty \right\},$$

where ρ is a nonnegative constant discount factor. We denote by $\mathcal{A} = \mathcal{A}^1 \times \dots \times \mathcal{A}^n$, and for any $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathcal{A}$, $i \in \llbracket 1, n \rrbracket$, we set $\alpha^{-i} = (\alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_n) \in \mathcal{A}^{-i} = \mathcal{A}^1 \times \dots \times \mathcal{A}^{i-1} \times \mathcal{A}^{i+1} \times \dots \times \mathcal{A}^n$.

Given a square integrable measurable random variable X_0 and control $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathcal{A}$, we consider the controlled linear mean-field stochastic differential equation in \mathbb{R}^d :

$$\begin{cases} dX_t &= b(t, X_t^\alpha, \mathbb{E}[X_t^\alpha], \alpha_t, \mathbb{E}[\alpha_t])dt + \sigma(t, X_t^\alpha, \mathbb{E}[X_t^\alpha], \alpha_t, \mathbb{E}[\alpha_t])dW_t, \quad 0 \leq t \leq T, \\ X_0^\alpha &= X_0, \end{cases} \tag{1}$$

where for $t \in [0, T]$, $x, \bar{x} \in \mathbb{R}^d$, $a_i, \bar{a}_i \in \mathbb{R}^{d_i}$:

$$\begin{cases} b(t, x, \bar{x}, \alpha, \bar{\alpha}) &= \beta_t + b_{x,t}x + \tilde{b}_{x,t}\bar{x} + \sum_{i=1}^n b_{i,t}\alpha_i + \tilde{b}_{i,t}\bar{\alpha}_i \\ &= \beta_t + b_{x,t}x + \tilde{b}_{x,t}\bar{x} + B_t\alpha + \tilde{B}_t\bar{\alpha} \\ \sigma(t, x, \bar{x}, \alpha, \bar{\alpha}) &= \gamma_t + \sigma_{x,t}x + \tilde{\sigma}_{x,t}\bar{x} + \sum_{i=1}^n \sigma_{i,t}\alpha_i + \tilde{\sigma}_{i,t}\bar{\alpha}_i \\ &= \gamma_t + \sigma_{x,t}x + \tilde{\sigma}_{x,t}\bar{x} + \Sigma_t\alpha + \tilde{\Sigma}_t\bar{\alpha}. \end{cases} \tag{2}$$

Here all the coefficients are deterministic matrix-valued processes except β and σ which are vector-valued \mathbb{F} -progressively measurable processes.

The goal of each player $i \in \llbracket 1, n \rrbracket$ during the game is to minimize her cost functional over $\alpha_i \in \mathcal{A}^i$, given the actions α^{-i} of the other players:

$$\begin{aligned}
 J^i(\alpha_i, \alpha^{-i}) &= \mathbb{E} \left[\int_0^T e^{-\rho t} f^i(t, X_t^\alpha, \mathbb{E}[X_t^\alpha], \alpha_t, \mathbb{E}[\alpha_t]) dt + g^i(X_T^\alpha, \mathbb{E}[X_T^\alpha]) \right], \\
 V^i(\alpha^{-i}) &= \inf_{\alpha_i \in \mathcal{A}^i} J^i(\alpha_i, \alpha^{-i}),
 \end{aligned}$$

where for each $t \in [0, T]$, $x, \bar{x} \in \mathbb{R}^d$, $a_i, \bar{a}_i \in \mathbb{R}^{d_i}$, we have set the running cost and terminal cost for each player:

$$\begin{cases}
 f^i(t, x, \bar{x}, a, \bar{a}) &= (x - \bar{x})^\top Q_t^i (x - \bar{x}) + \bar{x}^\top [Q_t^i + \tilde{Q}_t^i] \bar{x} \\
 &+ \sum_{k=1}^n a_k^\top I_{k,t}^i (x - \bar{x}) + \bar{a}_k^\top (I_{k,t}^i + \tilde{I}_{k,t}^i) \bar{x} \\
 &+ \sum_{k=1}^n (a_k - \bar{a}_k)^\top N_{k,t}^i (a_k - \bar{a}_k) + \bar{a}_k (N_{k,t}^i + \tilde{N}_{k,t}^i) \bar{a}_k \\
 &+ \sum_{0 \leq k \neq l \leq n} (a_k - \bar{a}_k)^\top G_{k,l,t}^i (a_l - \bar{a}_l) + a_k^\top (G_{k,l,t}^i + \tilde{G}_{k,l,t}^i) a_l \\
 &+ 2[L_{x,t}^{iT} x + \sum_{k=1}^n L_{k,t}^{i\top} a_k] \\
 g^i(x, \bar{x}) &= (x - \bar{x})^\top P^i (x - \bar{x}) + \bar{x}^\top (P^i + \tilde{P}^i) \bar{x} + 2r^{i\top} x.
 \end{cases} \tag{3}$$

Here all the coefficients are deterministic matrix-valued processes, except L_x^i, L_k^i, r^i which are vector-valued \mathbb{F} -progressively measurable processes, and \top denotes the transpose of a vector or matrix.

We say that $\alpha^* = (\alpha_1^*, \dots, \alpha_n^*) \in \mathcal{A}$ is a Nash equilibrium if for any $i \in \llbracket 1, n \rrbracket$,

$$J^i(\alpha^*) \leq J^i(\alpha_i, \alpha^{*,-i}), \quad \forall \alpha_i \in \mathcal{A}^i, \text{ i.e. } , J^i(\alpha^*) = V^i(\alpha^{*,-i}).$$

As it is well-known, the search for a Nash equilibrium can be formulated as a fixed point problem as follows: first, each player i has to compute its best response given the controls of the other players: $\alpha_i^* = BR_i(\alpha^{-i})$, where BR_i is the best response function defined (when it exists) as:

$$\begin{aligned}
 BR_i: \mathcal{A}^{-i} &\rightarrow \mathcal{A}^i \\
 \alpha^{-i} &\mapsto \operatorname{argmin}_{\alpha \in \mathcal{A}^i} J^i(\alpha, \alpha^{-i}).
 \end{aligned}$$

Then, in order to ensure that $(\alpha_1^*, \dots, \alpha_n^*)$ is a Nash equilibrium, we have to check that this candidate verifies the fixed point equation: $(\alpha_1^*, \dots, \alpha_i^*) = BR(\alpha_1^*, \dots, \alpha_i^*)$ where $BR := (BR_1, \dots, BR_n)$.

The main goal of this paper is to state a general martingale optimality principle for the search of Nash equilibria and to apply it to the linear quadratic case. We first obtain best response functions (or optimal control of each agent conditioned to the control of the others) of each player i of the following form:

$$\alpha_{i,t} = -(S_{i,t}^i)^{-1} U_{i,t}^i (X_t - \mathbb{E}[X_t]) - (S_{i,t}^i)^{-1} (\xi_{i,t}^i - \bar{\xi}_{i,t}^i) - (\hat{S}_{i,t}^i)^{-1} (V_{i,t}^i \mathbb{E}[X_t] + O_{i,t}^i)$$

where the coefficients in the r.h.s., defined in (5) and (6), depend on the actions α^{-i} of the other players. We then proceed to a fixed point search for best response function in order to exhibit a Nash equilibrium.

1.3 Notations and Assumptions

Given a normed space $(\mathbb{K}, |\cdot|)$, and for $T \in \mathbb{R}_+^*$, we set:

$$\begin{aligned}
 L^\infty([0, T], \mathbb{K}) &= \left\{ \phi : [0, T] \rightarrow \mathbb{K} \text{ s.t. } \phi \text{ is measurable and } \sup_{t \in [0, T]} |\phi_t| < \infty \right\} \\
 L^2([0, T], \mathbb{K}) &= \left\{ \phi : [0, T] \rightarrow \mathbb{K} \text{ s.t. } \phi \text{ is measurable and } \int_0^T e^{-\rho u} |\phi_t|^2 du < \infty \right\} \\
 L^2_{\mathcal{F}_T}(\mathbb{K}) &= \{ \phi : \Omega \rightarrow \mathbb{K} \text{ s.t. } \phi \text{ is } \mathcal{F}_T\text{-measurable and } \mathbb{E}[|\phi|^2] < \infty \} \\
 \mathbb{S}_{\mathbb{F}}^2(\Omega \times [0, T], \mathbb{K}) &= \{ \phi : \Omega \times [0, T] \rightarrow \mathbb{K} \text{ s.t. } \phi \text{ is } \mathbb{F}\text{-adapted and} \\
 &\quad \mathbb{E}[\sup_{t \in [0, T]} |\phi_t|^2] < \infty \} \\
 L^2_{\mathbb{F}}(\Omega \times [0, T], \mathbb{K}) &= \{ \phi : \Omega \times [0, T] \rightarrow \mathbb{K} \text{ s.t. } \phi \text{ is } \mathbb{F}\text{-adapted and} \\
 &\quad \int_0^T e^{-\rho u} \mathbb{E}[|\phi_u|^2] du < \infty \}.
 \end{aligned}$$

Note that when we will tackle the infinite horizon case we will set $T = \infty$. To make the notations less cluttered, we sometimes denote $X = X^\alpha$ when there is no ambiguity. If C and \tilde{C} are coefficients of our model, either in the dynamics or in a cost function, we note: $\hat{C} = C + \tilde{C}$. Given a random variable Z with a first moment, we denote by $\bar{Z} = \mathbb{E}[Z]$. For $M \in \mathbb{R}^{n \times n}$ and $X \in \mathbb{R}^n$, we denote by $M.X^{\otimes 2} = X^\top M X \in \mathbb{R}$. We denote by \mathbb{S}^d the set of symmetric $d \times d$ matrices and by \mathbb{S}_+^d the subset of non-negative symmetric matrices.

Let us now detail here the assumptions on the coefficients.

(H1) The coefficients in the dynamics (2) satisfy:

- a) $\beta, \gamma \in L^2_{\mathbb{F}}(\Omega \times [0, T], \mathbb{R}^d)$,
- b) $b_x, \tilde{b}_x, \sigma_x, \tilde{\sigma}_x \in L^\infty([0, T], \mathbb{R}^{d \times d})$; $b_i, \tilde{b}_i, \sigma_i, \tilde{\sigma}_i \in L^\infty([0, T], \mathbb{R}^{d \times d_i})$.

(H2) The coefficients of the cost functional (3) satisfy:

- a) $Q^i, \tilde{Q}^i \in L^\infty([0, T], \mathbb{S}_+^d)$, $P^i, \tilde{P}^i \in \mathbb{S}^d$, $N_k^i, \tilde{N}_k^i \in L^\infty([0, T], \mathbb{S}_+^{d_k})$, $I_k^i, \tilde{I}_k^i \in L^\infty([0, T], \mathbb{R}^{d_k \times d})$,
- b) $L_x^i \in L^2_{\mathbb{F}}(\Omega \times [0, T], \mathbb{R}^d)$, $L_k^i \in L^2_{\mathbb{F}}(\Omega \times [0, T], \mathbb{R}^{d_k})$, $r^i \in L^2_{\mathcal{F}_T}(\mathbb{R}^d)$,
- c) $\exists \delta > 0 \forall t \in [0, T]$:
 $N_{i,t}^i \geq \delta \mathbb{I}_{d_k}$ $P^i \geq 0$ $Q_t^i - I_{i,t}^{i\top} (N_{i,t}^i)^{-1} I_{i,t}^i \geq 0$,
- d) $\exists \delta > 0 \forall t \in [0, T]$:
 $\hat{N}_{i,t}^i \geq \delta \mathbb{I}_{d_k}$ $\hat{P}^i \geq 0$ $\hat{Q}_t^i - \hat{I}_{i,t}^{i\top} (\hat{N}_{i,t}^i)^{-1} \hat{I}_{i,t}^i \geq 0$.

Under the above conditions, we easily derive some standard estimates on the mean-field SDE:

- By **(H1)** there exists a unique strong solution to the mean-field SDE (1), which verifies:

$$\mathbb{E} \left[\sup_{t \in [0, T]} |X_t^\alpha|^2 \right] \leq C_\alpha (1 + \mathbb{E}(|X_0|^2)) < \infty \tag{4}$$

where C_α is a constant which depending on α only through $\int_0^T e^{-\rho t} \mathbb{E}[|\alpha_t|^2] dt$.

- By **(H2)** and (4) we have:

$$J^i(\alpha) \in \mathbb{R} \text{ for each } \alpha \in \mathcal{A},$$

which means that the optimisation problem is well defined for each player.

2 A Weak submartingale optimality principle to compute a Nash-equilibrium

2.1 A verification Lemma

We first present the lemma on which the method is based.

Lemma 1 (Weak submartingale optimality principle). *Suppose there exists a couple*

$(\alpha^*, (\mathcal{W}^{\cdot, i})_{i \in \llbracket 1, n \rrbracket})$, where $\alpha^* \in \mathcal{A}$ and $\mathcal{W}^{\cdot, i} = \{\mathcal{W}_t^{\alpha^*, i}, t \in [0, T], \alpha \in \mathcal{A}\}$ is a family of adapted processes indexed by \mathcal{A} for each $i \in \llbracket 1, n \rrbracket$, such that:

- (i) For every $\alpha \in \mathcal{A}$, $\mathbb{E}[\mathcal{W}_0^{\alpha, i}]$ is independent of the control $\alpha_i \in \mathcal{A}^i$;
- (ii) For every $\alpha \in \mathcal{A}$, $\mathbb{E}[\mathcal{W}_T^{\alpha, i}] = \mathbb{E}[g^i(X_T^\alpha, \mathbb{P}_{X_T^\alpha})]$;
- (iii) For every $\alpha \in \mathcal{A}$, the map $t \in [0, T] \mapsto \mathbb{E}[\mathcal{S}_t^{\alpha, i}]$, with $\mathcal{S}_t^{\alpha, i} = e^{-\rho t} \mathcal{W}_t^{\alpha, i} + \int_0^t e^{-\rho u} f^i(u, X_u^\alpha, \mathbb{P}_{X_u^\alpha}, \alpha_u, \mathbb{P}_{\alpha_u}) du$ is well defined and non-decreasing;
- (iv) The map $t \mapsto \mathbb{E}[\mathcal{S}_t^{\alpha^*, i}]$ is constant for every $t \in [0, T]$.

Then α^* is a Nash equilibrium and $J^i(\alpha^*) = \mathbb{E}[\mathcal{W}_0^{\alpha^*, i}]$. Moreover, any other Nash-equilibrium $\tilde{\alpha}$ such that $\mathbb{E}[\mathcal{W}_0^{\tilde{\alpha}, i}] = \mathbb{E}[\mathcal{W}_0^{\alpha^*, i}]$ and $J^i(\tilde{\alpha}) = J^i(\alpha^*)$ for any $i \in \llbracket 1, n \rrbracket$ satisfies the condition (iv).

Proof. Let $i \in \llbracket 1, n \rrbracket$ and $\alpha_i \in \mathcal{A}^i$. From (ii), we have immediately $J^i(\alpha) = \mathbb{E}[\mathcal{S}_T^\alpha]$ for any $\alpha \in \mathcal{A}$. We then have:

$$\begin{aligned} \mathbb{E}[\mathcal{W}_0^{(\alpha_i, \alpha^{*, -i}), i}] &= \mathbb{E}[\mathcal{S}_0^{(\alpha_i, \alpha^{*, -i}), i}] \\ &\leq \mathbb{E}[\mathcal{S}_T^{(\alpha_i, \alpha^{*, -i}), i}] = J^i(\alpha_i, \alpha^{*, -i}). \end{aligned}$$

Moreover for $\alpha_i = \alpha_i^*$ we have:

$$\begin{aligned} \mathbb{E}[\mathcal{W}_0^{(\alpha_i^*, \alpha^{*, -i}), i}] &= \mathbb{E}[\mathcal{S}_0^{(\alpha_i^*, \alpha^{*, -i}), i}] \\ &= \mathbb{E}[\mathcal{S}_T^{(\alpha_i^*, \alpha^{*, -i}), i}] = J^i(\alpha_i^*, \alpha^{*, -i}), \end{aligned}$$

which proves that α^* is a Nash equilibrium and $J^i(\alpha^*) = \mathbb{E}[\mathcal{W}_0^{\alpha^*,i}]$. Finally, let us suppose that $\tilde{\alpha} \in \mathcal{A}$ is another Nash equilibrium such that $\mathbb{E}[\mathcal{W}_0^{\tilde{\alpha},i}] = \mathbb{E}[\mathcal{W}_0^{\alpha^*,i}]$ and $J^i(\tilde{\alpha}) = J^i(\alpha^*)$ for any $i \in \llbracket 1, n \rrbracket$. Then, for $i \in \llbracket 1, n \rrbracket$ we have:

$$\mathbb{E}[\mathcal{S}_0^{\tilde{\alpha},i}] = \mathbb{E}[\mathcal{W}_0^{\tilde{\alpha},i}] = \mathbb{E}[\mathcal{W}_0^{\alpha^*,i}] = \mathbb{E}[\mathcal{S}_T^{\alpha^*,i}] = J^i(\alpha^*) = J^i(\tilde{\alpha}) = \mathbb{E}[\mathcal{S}_T^{\tilde{\alpha},i}].$$

Since $t \mapsto \mathbb{E}[\mathcal{S}_t^{\tilde{\alpha},i}]$ is nondecreasing for every $i \in \llbracket 1, n \rrbracket$, this implies that the map is actually constant and (iv) is verified.

2.2 The method

Let us now apply the optimality principle in Lemma 1 in order to find a Nash equilibrium. In the linear-quadratic case the laws of the state and the controls intervene only through their expectations. Thus we will use a simplified optimality principle where \mathbb{P} is simply replaced by \mathbb{E} in conditions (ii) and (iii) of Lemma 1. The general procedure is the following:

Step 1. We guess a candidate for $\mathcal{W}^{\alpha,i}$. To do so we suppose that $\mathcal{W}_t^{\alpha,i} = w_t^i(X_t^\alpha, \mathbb{E}[X_t^\alpha])$ for some parametric adapted random field $\{w_t^i(x, \bar{x}), t \in [0, T], x, \bar{x} \in \mathbb{R}^d\}$ of the form $w_t^i(x, \bar{x}) = K_t^i \cdot (x - \bar{x})^{\otimes 2} + \Lambda_t^i \cdot \bar{x}^{\otimes 2} + 2Y_t^{i\top} x + R_t^i$.

Step 2. We set $\mathcal{S}_t^{\alpha,i} = e^{-\rho t} w_t^i(X_t^\alpha, \mathbb{E}[X_t^\alpha]) + \int_0^t e^{-\rho u} f^i(u, X_u^\alpha, \mathbb{E}[X_u^\alpha], \alpha_u, \mathbb{E}[\alpha_u]) du$ for $i \in \llbracket 1, n \rrbracket$ and $\alpha \in \mathcal{A}$. We then compute $\frac{d}{dt} \mathbb{E}[\mathcal{S}_t^{\alpha,i}] = e^{-\rho t} \mathbb{E}[D_t^{\alpha,i}]$ (with Itô's formula) where the drift $D^{\alpha,i}$ takes the form:

$$\begin{aligned} \mathbb{E}[D_t^{\alpha,i}] = & \mathbb{E} \left[-\rho w_t^i(X_t^\alpha, \mathbb{E}[X_t^\alpha]) + \frac{d}{dt} \mathbb{E} [w_t^i(X_t^\alpha, \mathbb{E}[X_t^\alpha])] \right. \\ & \left. + f^i(t, X_t^\alpha, \mathbb{E}[X_t^\alpha], \alpha_t, \mathbb{E}[\alpha_t]) \right]. \end{aligned}$$

Step 3. We then constrain the coefficients of the random field so that the conditions of Lemma 1 are satisfied. This leads to a system of backward ordinary and stochastic differential equations for the coefficients of w^i .

Step 4. At time t , given the state and the controls of the other players, we seek the action α_i cancelling the drift. We thus obtain the best response function of each player.

Step 5. We compute the fixed point of the best response functions in order to find an open loop Nash equilibrium $t \mapsto \alpha_t^*$.

Step 6. We check the validity of our computations.

2.2.1 Step 1: guess the random fields

The process $t \mapsto \mathbb{E} [w_t^i(X_t^\alpha, \mathbb{E}[X_t^\alpha])]$ is meant to be equal to $\mathbb{E} [g^i(X_T^\alpha, \mathbb{E}[X_T^\alpha])]$ at time T , where $g(x, \bar{x}) = P^i \cdot (x - \bar{x})^{\otimes 2} + (P^i + \tilde{P}^i) \cdot \bar{x}^{\otimes 2} + r^{i\top} x$ with $(P, \tilde{P}, r^i) \in (\mathbb{S}^d)^2 \times$

$L^2_{\mathcal{F}_T}(\mathbb{R}^d)$. It is then natural to search for a field w^i of the form $w^i_t(x, \bar{x}) = K^i_t \cdot (x - \bar{x})^{\otimes 2} + \Lambda^i_t \cdot \bar{x}^{\otimes 2} + 2Y^i_t{}^\top x + R^i_t$ with the processes $(K^i, \Lambda^i, Y^i, R^i)$ in $(L^\infty([0, T], \mathbb{S}^d_+)^2 \times \mathbb{S}^2_{\mathbb{F}}(\Omega \times [0, T], \mathbb{R}^d) \times L^\infty([0, T], \mathbb{R}))$ and solution to:

$$\begin{cases} dK^i_t = \dot{K}^i_t dt, & K^i_T = P^i \\ d\Lambda^i_t = \dot{\Lambda}^i_t dt, & \Lambda^i_T = P^i + \tilde{P}^i \\ dY^i_t = \dot{Y}^i_t dt + Z^i_t dW_t, & 0 \leq t \leq T, \quad Y^i_T = r^i \\ dR^i_t = \dot{R}^i_t dt, & R^i_T = 0, \end{cases}$$

where $(\dot{K}^i, \dot{\Lambda}^i, \dot{R}^i)$ are deterministic processes valued in $\mathbb{S}^d \times \mathbb{S}^d \times \mathbb{R}$ and (\dot{Y}^i, Z^i) are adapted processes valued in \mathbb{R}^d .

2.2.2 Step 2: derive their drifts

For $i \in \llbracket 1, n \rrbracket$, $t \in [0, T]$ and $\alpha \in \mathcal{A}$, we set:

$$S_t^{\alpha, i} := e^{-\rho t} w^i_t(X_t, \mathbb{E}[X_t]) + \int_0^t e^{-\rho u} f^i(u, X_u^\alpha, \mathbb{E}[X_u^\alpha], \alpha_u, \mathbb{E}[\alpha_u]) du$$

and then compute the drift of the deterministic function $t \mapsto \mathbb{E}[S_t^{\alpha, i}]$:

$$\begin{aligned} \frac{d\mathbb{E}[S_t^{\alpha, i}]}{dt} &= e^{-\rho t} \mathbb{E}[D_t^{\alpha, i}] \\ &= e^{-\rho t} \mathbb{E}[(X_t - \bar{X}_t)^\top (K^i_t + \Phi^i_t)(X_t - \bar{X}_t) + \bar{X}_t^\top (\dot{\Lambda}^i_t + \Psi^i_t) \bar{X}_t + 2[\dot{Y}^i_t + \Lambda^i_t]^\top X_t \\ &\quad + \dot{R}^i_t - \rho R^i_t + \bar{\Gamma}^i_t + \chi^i_t(\alpha_{i, t})], \end{aligned}$$

where we have defined:

$$\begin{aligned} \chi^i_t(\alpha_{i, t}) &:= (\alpha_{i, t} - \bar{\alpha}_{i, t})^\top S_{i, t}^i(\alpha_i - \bar{\alpha}_{i, t}) + \bar{\alpha}_{i, t}^\top \hat{S}_{i, t}^i \bar{\alpha}_{i, t} \\ &\quad + 2[U_{i, t}^i(X_t - \bar{X}_t) + V_{i, t}^i \bar{X}_t + O_{i, t}^i + \xi_{i, t}^i - \bar{\xi}_{i, t}^i]^\top \alpha_{i, t} \end{aligned}$$

with the following coefficients:

$$\left\{ \begin{aligned} \Phi^i_t &= Q^i_t + \sigma_{x, t}^\top K^i_t \sigma_{x, t} + K^i_t b_{x, t} + b_{x, t}^\top K^i_t - \rho K^i_t \\ \Psi^i_t &= \dot{Q}^i_t + \dot{\sigma}_{x, t}^\top K^i_t \dot{\sigma}_{x, t} + \Lambda^i_t \hat{b}_{x, t} + \hat{b}_{x, t}^\top \Lambda^i_t - \rho \Lambda^i_t \\ \Delta^i_t &= L^i_{x, t} + b_{x, t}^\top Y^i_t + \bar{b}_{x, t}^\top \bar{Y}^i_t + \sigma_{x, t}^\top Z^i_t + \bar{\sigma}_{x, t}^\top \bar{Z}^i_t + \Lambda^i_t \bar{\beta}_t \\ &\quad + \sigma_{x, t}^\top K^i_t \gamma_t + \bar{\sigma}_{x, t}^\top K^i_t \bar{\gamma}_t + K^i_t (\beta_t - \bar{\beta}_t) - \rho Y^i_t \\ &\quad + \sum_{k \neq i} U_{k, t}^{i\top} (\alpha_{k, t} - \bar{\alpha}_{k, t}) + V_{k, t}^{i\top} \bar{\alpha}_{k, t} \\ \Gamma^i_t &= \gamma_t^\top K^i_t \gamma_t + 2\beta_t^\top Y^i_t + 2\gamma_t^\top Z^i_t \\ &\quad + \sum_{k \neq i} (\alpha_{k, t} - \bar{\alpha}_{k, t})^\top S_{k, t}^i(\alpha_{k, t} - \bar{\alpha}_{k, t}) + \bar{\alpha}_{k, t}^\top \hat{S}_{k, t}^i \bar{\alpha}_{k, t} + \\ &\quad 2[O_{k, t}^i + \xi_{k, t}^i - \bar{\xi}_{k, t}^i]^\top \alpha_{k, t} - \rho R^i_t, \end{aligned} \right. \tag{5}$$

and

$$\left\{ \begin{aligned} S_{k,t}^i &= N_{k,t}^i + \sigma_{k,t}^\top K_t^i \sigma_{k,t} \\ \hat{S}_{k,t}^i &= \hat{N}_{k,t}^i + \hat{\sigma}_{k,t}^\top K_t^i \hat{\sigma}_{k,t} \\ U_{k,t}^i &= I_{k,t}^i + \sigma_{k,t}^\top K_t^i \sigma_{x,t} + b_{k,t}^\top K_t^i \\ V_{k,t}^i &= \hat{I}_{k,t}^i + \hat{\sigma}_{k,t}^\top K_t^i \hat{\sigma}_{x,t} + \hat{b}_{k,t}^\top \Lambda_t^i \\ O_{k,t}^i &= \bar{L}_{k,t}^i + \hat{b}_{k,t}^\top \bar{Y}_t^i + \hat{\sigma}_{k,t}^\top \bar{Z}_t^i + \hat{\sigma}_{k,t}^\top K_t^i \bar{Y}_t^i \\ &\quad + \frac{1}{2} \sum_{k \neq i} (\hat{J}_{i,k,t}^i + \hat{J}_{k,i,t}^{i\top}) \bar{\alpha}_{k,t} \\ J_{k,l,t}^i &= G_{k,l,t}^i + \sigma_{k,t}^\top K_t^i \sigma_{l,t} \\ \hat{J}_{k,l,t}^i &= \hat{G}_{k,l,t}^i + \hat{\sigma}_{k,t}^\top K_t^i \hat{\sigma}_{l,t} \\ \xi_{k,t}^i &= L_{k,t}^i + b_{k,t}^\top Y_t^i + \sigma_{k,t}^\top Z_t^i + \sigma_{k,t}^\top K_t^i \gamma_t \\ &\quad + \frac{1}{2} \sum_{k \neq i} (J_{i,k,t}^i + J_{k,i,t}^{i\top}) \alpha_{k,t}. \end{aligned} \right. \tag{6}$$

2.2.3 Step 3: constrain their coefficients

Now that we have computed the drift, we need to constrain the coefficients so that $S^{\alpha,i}$ satisfies the condition of Lemma 1. Let us assume for the moment that $S_{i,t}^i$ and $\hat{S}_{i,t}^i$ are positive definite matrices (this will be ensured by the positive definiteness of K). That implies that there exists an invertible matrix θ_t^i such that $\theta_t^i S_{i,t}^i \theta_t^{i\top} = \hat{S}_{i,t}^i$ for all $t \in [0, T]$. We can now rewrite the drift as: "a square in α_i " + "other terms not depending in α_i ". Since we can form the following square:

$$\mathbb{E}[\chi_t^i(\alpha_{i,t})] = \mathbb{E}[(\alpha_{i,t} - \bar{\alpha}_{i,t} + \theta_t^{i\top} \bar{\alpha}_{i,t} - \eta_t^i) S_{i,t}^i (\alpha_{i,t} - \bar{\alpha}_{i,t} + \theta_t^{i\top} \bar{\alpha}_{i,t} - \eta_t^i) - \zeta_t^i]$$

with:

$$\left\{ \begin{aligned} \eta_t^i &= a_t^{i,0}(X_t, \bar{X}_t) + \theta_t^{i\top} a_t^{i,1}(\bar{X}_t) \\ a_t^{i,0}(x, \bar{x}) &= -(S_{i,t}^i)^{-1} U_{i,t}^i (x - \bar{x}) - (S_{i,t}^i)^{-1} (\xi_{i,t}^i - \bar{\xi}_{i,t}^i) \\ a_t^{i,1}(\bar{x}) &= -(\hat{S}_{i,t}^i)^{-1} (V_{i,t}^i \bar{x} + O_{i,t}^i) \\ \zeta_t^i &= (X_t - \bar{X}_t)^\top U_{i,t}^{i\top} (S_{i,t}^i)^{-1} U_{i,t}^i (X_t - \bar{X}_t) + \bar{X}_t^\top V_{i,t}^{i\top} (\hat{S}_{i,t}^i)^{-1} V_{i,t}^i \bar{X}_t \\ &\quad + 2(U_{i,t}^{i\top} (S_{i,t}^i)^{-1} (\xi_{i,t}^i - \bar{\xi}_{i,t}^i) + V_{i,t}^i (\hat{S}_{i,t}^i)^{-1} O_{i,t}^i) X_t \\ &\quad + (\xi_{i,t}^i - \bar{\xi}_{i,t}^i)^\top (S_{i,t}^i)^{-1} (\xi_{i,t}^i - \bar{\xi}_{i,t}^i) + O_{i,t}^{i\top} (\hat{S}_{i,t}^i)^{-1} O_{i,t}^i, \end{aligned} \right.$$

we can then rewrite the drift in the following form:

$$\begin{aligned} \mathbb{E}[D_t^{\alpha,i}] &= \mathbb{E}[(X_t - \bar{X}_t)^\top [K_t^i + \Phi_t^{i0}](X_t - \bar{X}_t) + \bar{X}_t^\top (\Lambda_t^i + \Psi_t^{i0}) \bar{X}_t + 2[\dot{Y}_t^i + \Delta_t^{i0}]^\top X_t \\ &\quad + \dot{R}_t^i + \bar{\Gamma}_t^{i0} \\ &\quad + (\alpha_{i,t} - \bar{\alpha}_{i,t} + \theta_t^{i\top} \bar{\alpha}_{i,t} - \eta_t^i) S_{i,t}^i (\alpha_{i,t} - \bar{\alpha}_{i,t} + \theta_t^{i\top} \bar{\alpha}_{i,t} - \eta_t^i), \end{aligned}$$

where

$$\begin{cases} \Phi_t^{i0} &= \Phi_t^i - U_{i,t}^{i\top} (S_{i,t}^i)^{-1} U_{i,t}^i \\ \Psi_t^{i0} &= \Psi_t^i - V_{i,t}^{i\top} (\hat{S}_{i,t}^i)^{-1} V_{i,t}^i \\ \Delta_t^{i0} &= \Delta_t^i - U_{i,t}^{i\top} (S_{i,t}^i)^{-1} (\xi_{i,t}^i - \bar{\xi}_{i,t}^i) - V_{i,t}^{i\top} (\hat{S}_{i,t}^i)^{-1} O_{i,t}^i \\ \Gamma_t^{i0} &= \Gamma_t^i - (\xi_{i,t}^i - \bar{\xi}_{i,t}^i)^\top (S_{i,t}^i)^{-1} (\xi_{i,t}^i - \bar{\xi}_{i,t}^i) - O_{i,t}^{i\top} (\hat{S}_{i,t}^i)^{-1} O_{i,t}^i. \end{cases} \quad (7)$$

We can finally constrain the coefficients. By choosing the coefficients K^i, Γ^i, Y^i and R^i so that only the square remains, the drift for each player $i \in \llbracket 1, n \rrbracket$ can be rewritten as a square only (in the next step we will verify that we can indeed choose such coefficients). More precisely we set K^i, Γ^i, Y^i and R^i as the solution of:

$$\begin{cases} dK_t^i = -\Phi_t^{i0} dt & K_T^i = P^i \\ d\Lambda_t^i = -\Psi_t^{i0} dt & \Lambda_T^i = P^i + \bar{P}^i \\ dY_t^i = -\Delta_t^{i0} dt + Z_t^i dW_t & Y_T^i = r^i \\ dR_t^i = -\Gamma_t^{i0} dt & R_T^i = 0, \end{cases} \quad (8)$$

and stress the fact that Y^i, Z^i, R^i depend on α^{-i} , which appears in the coefficients Δ^{i0} , and Γ^{i0} . With such coefficients the drift takes now the form:

$$\begin{aligned} \mathbb{E}[D_t^{\alpha^{-i}}] &= \mathbb{E}[(\alpha_{i,t} - \bar{\alpha}_{i,t} + \theta_t^{i\top} \bar{\alpha}_{i,t} - \eta_t^i) S_{i,t}^i (\alpha_{i,t} - \bar{\alpha}_{i,t} + \theta_t^{i\top} \bar{\alpha}_{i,t} - \eta_t^i)] \\ &= \mathbb{E}[(\alpha_{i,t} - \bar{\alpha}_{i,t} - a_t^{i,1} + \theta_t^{i\top} (\bar{\alpha}_{i,t} - a_t^{i,0})) S_{i,t}^i (\alpha_{i,t} - \bar{\alpha}_{i,t} - a_t^{i,1} + \theta_t^{i\top} (\bar{\alpha}_{i,t} - a_t^{i,0}))] \end{aligned}$$

and thus satisfies the nonnegativity constraint: $\mathbb{E}[D_t^{\alpha^{-i}}] \geq 0$, for all $t \in [0, T]$, $i \in \llbracket 1, n \rrbracket$, and $\alpha \in \mathcal{A}$.

2.2.4 Step 4: find the best response functions

Proposition 1 *Assume that for all $i \in \llbracket 1, n \rrbracket$, $(K^i, \Lambda^i, Y^i, Z^i, R^i)$ is a solution of (8) given $\alpha^{-i} \in \mathcal{A}^{-i}$. Then the set of processes*

$$\begin{aligned} \alpha_{i,t} &= a_t^{i,0}(X_t, \mathbb{E}[X_t]) + a_t^{i,1}(\mathbb{E}[X_t]) \\ &= -(S_{i,t}^i)^{-1} U_{i,t}^i (X_t - \mathbb{E}[X_t]) - (S_{i,t}^i)^{-1} (\xi_{i,t}^i - \bar{\xi}_{i,t}^i) - (\hat{S}_{i,t}^i)^{-1} (V_{i,t}^i \mathbb{E}[X_t] + O_{i,t}^i) \end{aligned} \quad (9)$$

(depending on α^{-i}) where X is the state process with the feedback controls $\alpha = (\alpha_1, \dots, \alpha_n)$, are best-response functions, i.e., $J^i(\alpha_i, \alpha^{-i}) = V^i(\alpha^{-i})$ for all $i \in \llbracket 1, n \rrbracket$. Moreover we have

$$\begin{aligned} V^i(\alpha^{-i}) &= \mathbb{E}[W_0^{i,\alpha}] \\ &= \mathbb{E}[K_0^i \cdot (X_0 - \bar{X}_0)^{\otimes 2} + \Lambda_0^i \cdot \bar{X}_0^{\otimes 2} + 2Y_0^{i\top} X_0 + R_0^i]. \end{aligned}$$

Proof. We check that the assumptions of Lemma 1 are satisfied. Since $\mathcal{W}^{\alpha,i}$ is of the form $\mathcal{W}_t^{\alpha,i} = w_t^i(X_t^\alpha, \mathbb{E}[X_t^\alpha])$, condition (i) is verified. The condition (ii) is satisfied thanks to the terminal conditions imposed on the system (8). Since $(K^i, \Lambda^i, Y^i, Z^i, R^i)$ is solution to (8), the drift of $t \mapsto \mathbb{E}[S^{\alpha,i}]$ is positive for all $i \in \llbracket 1, n \rrbracket$ and all $\alpha \in \mathcal{A}$, which implies condition (iii). Finally, for $\alpha \in \mathcal{A}$, we see that $\mathbb{E}[D_t^{\alpha,i}] \equiv 0$ for $t \in [0, T]$ and $i \in \llbracket 1, n \rrbracket$ if and only if:

$$\alpha_{i,t} - \bar{\alpha}_{i,t} - a_t^{i,1}(X_t^\alpha, \mathbb{E}[X_t^\alpha]) + \theta_t^{i\top}(\bar{\alpha}_{i,t} - a_t^{i,0}(\mathbb{E}[X_t^\alpha])) = 0 \quad a.s. \quad t \in [0, T].$$

Since θ_t^i is invertible, we get $\bar{\alpha}_{i,t} = a_t^{i,0}$ by taking the expectation in the above formula. Thus $\mathbb{E}[D_t^{\alpha,i}] \equiv 0$ for every $i \in \llbracket 1, n \rrbracket$ and $t \in [0, T]$ if and only if $\alpha_{i,t} = \bar{\alpha}_{i,t} + a_t^{i,1} = a_t^{i,1} + a_t^{i,0}$ for every $i \in \llbracket 1, n \rrbracket$ and $t \in [0, T]$. For such controls for the players, the condition (iv) is satisfied. We now check that $\alpha_i \in \mathcal{A}^i$ for every $i \in \llbracket 1, n \rrbracket$ (i.e. it satisfies the square integrability condition). Since X is solution to a linear McKean-Vlasov dynamics and satisfies the square integrability condition $\mathbb{E}[\sup_{0 \leq t \leq T} |X_t|^2] < \infty$, it implies that $\alpha_i \in L^2_{\mathbb{F}}(\Omega \times [0, T], \mathbb{R}^{d_i})$ since $S_t^i, U_t^i, \hat{S}_t^i, V_t^i$ are bounded and $(O_t^i, \xi_t^i) \in L^2([0, T], \mathbb{R}^{d_i}) \times L^2(\Omega \times [0, T], \mathbb{R}^{d_i})$. Therefore $\alpha_i \in \mathcal{A}^i$ for every $i \in \llbracket 1, n \rrbracket$.

2.2.5 Step 5: search for a fixed point

We now find semi-explicit expressions for the optimal controls of each player. The issue here is the fact that the controls of the other players appear in the best response functions of each player through the vectors $(Y^1, Z^1), \dots, (Y^n, Z^n)$. To solve this fixed point problem, we first rewrite (9) and the backward equations followed by $(Y, Z) = ((Y^1, Z^1), \dots, (Y^n, Z^n))$ in the following way (note that we omit the time dependence of the coefficients to make the notations less cluttered):

$$\begin{cases} \alpha_t^* - \bar{\alpha}_t^* &= S_x(X_t - \bar{X}_t) + S_y(Y_t - \bar{Y}_t) + S_z(Z_t - \bar{Z}_t) + H - \bar{H} \\ \bar{\alpha}_t^* &= \hat{S}_x \bar{X}_t + \hat{S}_y \bar{Y}_t + \hat{S}_z \bar{Z}_t + \bar{H} \\ dY_t &= (P_y(Y_t - \bar{Y}_t) + P_z(Z_t - \bar{Z}_t) + P_\alpha(\alpha_t - \bar{\alpha}_t) + F - \bar{F} \\ &\quad + \hat{P}_y \bar{Y}_t + \hat{P}_z \bar{Z}_t + \hat{P}_\alpha \bar{\alpha}_t + \bar{F}) dt \\ &\quad + Z_t dW_t, \end{cases} \quad (10)$$

where we define

$$\left\{ \begin{array}{l}
S = ((S_i^i)^{-1} \mathbf{1}_{i=j})_{i,j \in \llbracket 1, n \rrbracket} \\
\hat{S} = ((\hat{S}_i^i)^{-1} \mathbf{1}_{i=j})_{i,j \in \llbracket 1, n \rrbracket} \\
J = \frac{1}{2} \left((J_{ij}^i + J_{ji}^i) \mathbf{1}_{i \neq j} \right)_{i,j \in \llbracket 1, n \rrbracket} \\
\hat{J} = \frac{1}{2} \left((\hat{J}_{ij}^i + \hat{J}_{ji}^i) \mathbf{1}_{i \neq j} \right)_{i,j \in \llbracket 1, n \rrbracket} \\
\mathcal{J} = -(I_d + SJ)^{-1} S \\
\hat{\mathcal{J}} = -(I_d + \hat{S}\hat{J})^{-1} \hat{S} \\
S_x = \mathcal{J} (U_i^i)_{i \in \llbracket 1, n \rrbracket} \\
\hat{S}_x = \hat{\mathcal{J}} (V_i^i)_{i \in \llbracket 1, n \rrbracket} \\
S_y = \mathcal{J} (\mathbf{1}_{i=j} b_i^\top)_{i,j \in \llbracket 1, n \rrbracket} \\
\hat{S}_y = \hat{\mathcal{J}} (\mathbf{1}_{i=j} \hat{b}_i^\top)_{i,j \in \llbracket 1, n \rrbracket} \\
S_z = \mathcal{J} (\sigma_i^\top)_{i \in \llbracket 1, n \rrbracket} \\
\hat{S}_z = \hat{\mathcal{J}} (\hat{\sigma}_i^\top)_{i \in \llbracket 1, n \rrbracket} \\
H = \mathcal{J} (L_i^i + \sigma_i^\top K^i \gamma)_{i \in \llbracket 1, n \rrbracket} \\
\hat{H} = \hat{\mathcal{J}} (L_i^i + \hat{\sigma}_i^\top K^i \gamma)_{i \in \llbracket 1, n \rrbracket}
\end{array} \right. \left\{ \begin{array}{l}
P_y = (\mathbf{1}_{i=j} (U_i^i (S_i^i)^{-1} b_i^\top - b_x^\top + \rho))_{i,j \in \llbracket 1, n \rrbracket} \\
\hat{P}_y = (\mathbf{1}_{i=j} (V_i^i (\hat{S}_i^i)^{-1} \hat{b}_i^\top - \hat{b}_x^\top + \rho))_{i,j \in \llbracket 1, n \rrbracket} \\
P_z = (\mathbf{1}_{i=j} (U_i^i (S_i^i)^{-1} \sigma_i^\top - \sigma_x^\top))_{i,j \in \llbracket 1, n \rrbracket} \\
\hat{P}_z = (\mathbf{1}_{i=j} (V_i^i (\hat{S}_i^i)^{-1} \hat{\sigma}_i^\top - \hat{\sigma}_x^\top))_{i,j \in \llbracket 1, n \rrbracket} \\
P_\alpha = -(\mathbf{1}_{i \neq j} (U_j^i \\
\quad + U_i^i (S_i^i)^{-1} (J_{ij}^i + J_{ji}^i)))_{i,j \in \llbracket 1, n \rrbracket} \\
\hat{P}_\alpha = -(\mathbf{1}_{i \neq j} (V_j^i \\
\quad + V_i^i (\hat{S}_i^i)^{-1} (\hat{J}_{ij}^i + \hat{J}_{ji}^i)))_{i,j \in \llbracket 1, n \rrbracket} \\
F = (K^i \beta + \sigma_x^\top K^i \gamma)_{i \in \llbracket 1, n \rrbracket} \\
\hat{F} = (U_i^i (S_i^i)^{-1} (L_i + \sigma_i^\top K^i \gamma) - L_x - \\
\quad \sigma_x^\top K^i \gamma - K^i \beta)_{i \in \llbracket 1, n \rrbracket}.
\end{array} \right. \quad (11)$$

Now, the strategy is to propose an ansatz for $t \in [0, T] \mapsto Y_t$ in the form:

$$Y_t = \pi_t (X_t - \bar{X}_t) + \hat{\pi}_t \bar{X}_t + \eta_t \quad (12)$$

where $(\pi, \hat{\pi}, \eta) \in L^\infty([0, T], \mathbb{R}^{nd \times d}) \times L^\infty([0, T], \mathbb{R}^{nd \times d}) \times \mathcal{S}_{\mathbb{F}}^2(\Omega \times [0, T], \mathbb{R}^{nd})$ satisfy:

$$\left\{ \begin{array}{l}
d\eta_t = \psi_t dt + \phi_t dW_t, \quad \eta_T = r = (r^i)_{i \in \llbracket 1, n \rrbracket} \\
d\pi_t = \hat{\pi}_t dt, \quad \pi_T = 0 \\
d\hat{\pi}_t = \hat{\hat{\pi}}_t dt, \quad \hat{\pi}_T = 0.
\end{array} \right.$$

By applying Itô's formula to the ansatz we then obtain:

$$\begin{aligned}
dY_t &= \hat{\pi}_t (X_t - \bar{X}_t) dt + \pi_t d(X_t - \bar{X}_t) + \hat{\pi}_t \bar{X}_t dt + \hat{\pi}_t d\bar{X}_t + \psi_t dt + \phi_t dW \\
&= dt \left[\hat{\pi}_t (X_t - \bar{X}_t) + \psi_t - \bar{\psi}_t + \pi_t (\beta - \bar{\beta} + b_x (X_t - \bar{X}_t) + B(\alpha_t - \bar{\alpha}_t)) \right] \\
&\quad + dt \left[\hat{\pi}_t \bar{X}_t + \bar{\psi}_t + \hat{\pi}_t (\bar{\beta} + \hat{b}_x \bar{X}_t + \hat{B} \bar{\alpha}_t) \right] \\
&\quad + dW_t \left[\phi_t + \pi_t (\gamma + \sigma_x X_t + \bar{\sigma}_x \bar{X}_t + \Sigma \alpha_t + \bar{\Sigma} \bar{\alpha}_t) \right].
\end{aligned}$$

By comparing the two Itô's decompositions of Y , we get

$$\begin{cases} P_y(Y_t - \bar{Y}_t) + P_z(Z_t - \bar{Z}_t) \\ + P_\alpha(\alpha_t - \bar{\alpha}_t) + F - \bar{F} & = \hat{\pi}_t(X_t - \bar{X}_t) + \psi_t - \bar{\psi}_t \\ & \quad + \pi_t \left(\beta - \bar{\beta} + b_x(X_t - \bar{X}_t) + B(\alpha_t - \bar{\alpha}_t) \right) \\ \hat{P}_y \bar{Y}_t + \hat{P}_z \bar{Z}_t + \hat{P}_\alpha \bar{\alpha}_t + \bar{F} & = \left[\hat{\pi}_t \bar{X}_t + \bar{\psi}_t + \hat{\pi}_t \left(\bar{\beta} + \hat{b}_x \bar{X}_t + \hat{B} \bar{\alpha}_t \right) \right] \\ Z_t & = \left[\phi_t + \pi_t \left(\gamma + \sigma_x X_t + \tilde{\sigma}_x \bar{X}_t + \Sigma \alpha_t + \tilde{\Sigma} \bar{\alpha}_t \right) \right]. \end{cases} \quad (13)$$

We now substitute the Y by its ansatz in the best response equation (10), and obtain the system:

$$\begin{cases} (Id - S_z \pi_t \Sigma)(\alpha_t^* - \bar{\alpha}_t^*) & = (S_x + S_y \pi_t + S_z \pi_t \sigma_x)(X_t - \bar{X}_t) \\ & \quad + (H - \bar{H} + S_y(\eta_t - \bar{\eta}_t) + S_z(\phi_t - \bar{\phi}_t + \pi_t(\gamma - \bar{\gamma}))) \\ (Id - \hat{S}_z \pi_t \hat{\Sigma}) \bar{\alpha}_t^* & = (\hat{S}_x + \hat{S}_y \hat{\pi}_t + \hat{S}_z \pi_t \hat{\sigma}_x) \bar{X}_t + (\hat{H} + \hat{S}_y \bar{\eta}_t + \hat{S}_z(\bar{\phi}_t + \pi_t \bar{\gamma})). \end{cases} \quad (14)$$

To make the next computations slightly less painful we rewrite (14) as

$$\begin{cases} \alpha_t^* - \bar{\alpha}_t^* & = A_x(X_t - \bar{X}_t) + R_t - \bar{R}_t \\ \bar{\alpha}_t^* & = \hat{A}_x \bar{X}_t + \hat{R}_t \\ & \text{where} \\ A_x & := (Id - S_z \pi_t \Sigma)^{-1} (S_x + S_y \pi_t + S_z \pi_t \sigma_x) \\ \hat{A}_x & := (Id - \hat{S}_z \pi_t \hat{\Sigma})^{-1} (\hat{S}_x + \hat{S}_y \pi_t + \hat{S}_z \pi_t \hat{\sigma}_x) \\ R_t & := (Id - S_z \pi_t \Sigma)^{-1} (H + S_y \eta_t + S_z(\phi_t + \pi_t \gamma)) \\ \hat{R}_t & := (Id - \hat{S}_z \pi_t \hat{\Sigma})^{-1} (\hat{H} + \hat{S}_y \eta_t + \hat{S}_z(\phi_t + \pi_t \gamma)). \end{cases} \quad (15)$$

By injecting (14) into (13) we have:

$$\begin{cases} 0 & = [\hat{\pi}_t + \pi_t b_x - P_y \pi_t - P_z \pi_t (\sigma_x + \Sigma A_x) - (P_\alpha - \pi_t B) A_x] (X_t - \bar{X}_t) \\ & \quad + \psi_t - \bar{\psi}_t + \pi_t (\beta - \bar{\beta}) - (P_\alpha - \pi_t B)(R - \bar{R}) - (F - \bar{F}) \\ & \quad - P_z(\phi_t - \bar{\phi}_t + \pi_t(\gamma - \bar{\gamma} + \Sigma(R - \bar{R}))) - P_y(\eta_t - \bar{\eta}_t) \\ 0 & = [\hat{\pi}_t + \hat{\pi}_t \hat{b}_x - \hat{P}_y \hat{\pi}_t - \hat{P}_z \pi_t (\hat{\sigma}_x + \hat{\Sigma} \hat{A}_x) - (\hat{P}_\alpha - \hat{\pi}_t \hat{B}) \hat{A}_x] \bar{X}_t \\ & \quad + \bar{\psi}_t + \hat{\pi}_t \bar{\beta} - (\hat{P}_\alpha - \hat{\pi}_t \hat{B}) \hat{R} - \bar{F} - \hat{P}_z(\bar{\phi}_t + \pi_t(\bar{\gamma} + \hat{\Sigma} \hat{R})) - \hat{P}_y \bar{\eta}_t. \end{cases}$$

Thus we constrain the coefficients $(\pi, \hat{\pi}, \psi, \phi)$ of the ansatz of Y to satisfy:

$$\left\{ \begin{array}{l}
\dot{\pi}_t = -\pi_t b_x + P_y \pi_t + P_z \pi_t \sigma_x + (P_\alpha + P_z \Sigma)(Id - S_z \pi_t \Sigma)^{-1}(S_x + S_y \pi_t + S_z \pi_t \sigma_x) \\
\quad - \pi_t B(Id - S_z \pi_t \Sigma)^{-1}(S_x + S_y \pi_t + S_z \pi_t \sigma_x) \\
\pi_T = 0 \\
\dot{\hat{\pi}}_t = -\hat{\pi}_t \hat{b}_x + \hat{P}_y \hat{\pi}_t + \hat{P}_z \pi_t \hat{\sigma}_x + (\hat{P}_\alpha + \hat{P}_z \hat{\Sigma})(Id - \hat{S}_z \pi_t \hat{\Sigma})^{-1}(\hat{S}_x + \hat{S}_y \hat{\pi}_t + \hat{S}_z \pi_t \hat{\sigma}_x) \\
\quad - \hat{\pi}_t \hat{B}(Id - \hat{S}_z \pi_t \hat{\Sigma})^{-1}(\hat{S}_x + \hat{S}_y \hat{\pi}_t + \hat{S}_z \pi_t \hat{\sigma}_x) \\
\hat{\pi}_T = 0 \\
d\eta_t = \psi_t dt + \phi_t dW \\
\eta_T = r \\
\text{where:} \\
\psi_t - \bar{\psi}_t = -\pi_t(\beta - \bar{\beta}) + (P_\alpha - \pi_t B)(R - \bar{R}) + (F - \bar{F}) \\
\quad + P_z(\phi_t - \bar{\phi}_t + \pi_t(\gamma - \bar{\gamma} + \Sigma(R - \bar{R}))) + P_y(\eta_t - \bar{\eta}_t) \\
\bar{\psi}_t = -\hat{\pi}_t \bar{\beta} + (\hat{P}_\alpha - \hat{\pi}_t \hat{B})\bar{R} + \bar{F} + \hat{P}_z(\bar{\phi}_t + \pi_t(\bar{\gamma} + \hat{\Sigma}\bar{R})) + \hat{P}_y \bar{\eta}_t \\
R_t := (Id - S_z \pi_t \Sigma)^{-1}(H + S_y \eta + S_z(\phi + \pi \gamma)) \\
\hat{R}_t := (Id - \hat{S}_z \pi_t \hat{\Sigma})^{-1}(\hat{H} + \hat{S}_y \eta + \hat{S}_z(\phi + \pi \gamma)).
\end{array} \right. \tag{16}$$

We now have a feedback form for $(Y, Z) = ((Y^1, Z^1), \dots, (Y^n, Z^n))$. We can inject it in the best response functions α^* in order to obtain the optimal controls in feedback form. We then inject these latter in the state equation in order to obtain an explicit expression of $t \mapsto X_t^*$.

2.2.6 Step 6: check the validity

Let us now check the existence and uniqueness of $t \mapsto (K_t^i, \Lambda_t^i, (Y_t^i, Z_t^i), R_t^i, \pi_t, \hat{\pi}_t, (\eta_t, \phi_t))$ where $K^i \in L^\infty([0, T], \mathbb{S}_+^d)$, $\Lambda^i \in L^\infty([0, T], \mathbb{S}_+^d)$, $(Y_t^i, Z_t^i) \in \mathbb{S}_{\mathbb{F}}^2(\Omega \times [0, T], \mathbb{R}^d) \times L_{\mathbb{F}}^2(\Omega \times [0, T], \mathbb{R}^d)$, $R^i \in L^\infty([0, T], \mathbb{R})$, $\pi, \hat{\pi} \in L^\infty([0, T], \mathbb{R}^{nd \times d})$ and $(\eta, \phi) \in \mathbb{S}_{\mathbb{F}}^2(\Omega \times [0, T], \mathbb{R}^{nd}) \times L_{\mathbb{F}}^2(\Omega \times [0, T], \mathbb{R}^{nd})$, under the assumptions **(H1)**-**(H2)**. We recall that $t \mapsto (K_t^i, \Lambda_t^i, (Y_t^i, Z_t^i), R_t^i)$ and $t \mapsto (\pi_t, \hat{\pi}_t, (\eta_t, \phi_t))$ are solutions respectively to **(8)** and **(16)**. Fix $i \in \llbracket 1, n \rrbracket$:

(i) We first consider the coefficients K^i which follow Riccati equations:

$$\left\{ \begin{array}{l}
\dot{K}_t^i + Q_t^i + \sigma_{x,t}^\top K_t^i \sigma_{x,t} + K_t^i b_{x,t} + b_{x,t}^\top K_t^i - \rho K_t^i \\
-(I_{k,t}^i + \sigma_{k,t}^\top K_t^i \sigma_{k,t} + b_{k,t}^\top K_t^i)(N_{k,t}^i + \sigma_{k,t}^\top K_t^i \sigma_{k,t})^{-1}(I_{k,t}^i + \sigma_{k,t}^\top K_t^i \sigma_{k,t} + b_{k,t}^\top K_t^i) = 0 \\
K_T^i = P^i.
\end{array} \right.$$

By standard result in control theory (see [16], Ch. 6, Thm 7.2) under **(H1)** and **(H2)** there exists a unique solution $K^i \in L^\infty([0, T], \mathbb{S}_+^d)$.

(ii) Given K^i let us now consider the Λ^i 's. They also follow Riccati equations:

$$\left\{ \begin{array}{l}
\dot{\Lambda}_t^i + \hat{Q}_t^{iK} + \Lambda_t^i \hat{b}_{x,t} + \hat{b}_{x,t}^\top \Lambda_t^i - \rho \Lambda_t^i - (\hat{I}_t^{iK} + \hat{b}_{i,t}^\top \Lambda_t^i)(\hat{N}_t^{iK})^{-1}(\hat{I}_t^{iK} + \hat{b}_{i,t}^\top \Lambda_t^i) = 0 \\
\Lambda_T^i = \hat{P}^i
\end{array} \right.$$

where we define:

$$\begin{cases} \hat{Q}_t^{iK} & := \hat{Q}_t^i + \hat{\sigma}_{x,t}^\top K_t^i \hat{\sigma}_{x,t} \\ \hat{I}_t^{iK} & := \hat{I}_t^i + \hat{\sigma}_{i,t}^\top K_t^i \hat{\sigma}_{x,t} \\ \hat{N}_t^{iK} & := \hat{N}_t^i + \hat{\sigma}_{i,t}^\top K_t^i \hat{\sigma}_{i,t}. \end{cases}$$

We need the same arguments as for K^i . The only missing argument to conclude the existence and uniqueness of Λ^i is: $\hat{Q}^{iK} - (\hat{I}^{iK})^\top (\hat{N}^{iK})^{-1} \hat{I}^{iK} \geq 0$. As in [2] we can prove with some algebraic calculations that it is implied by the hypothesis $\hat{Q}^i - (\hat{I}^i)^\top (\hat{N}^i)^{-1} \hat{I}^i \geq 0$ that we made in **(H2)**.

- (iii) Given (K^i, Λ^i) we now consider the equation for (Y^i, Z^i) which is a linear mean-field BSDE of the form:

$$\begin{cases} dY_t^i & = \mathcal{V}_t^i + \mathcal{G}_t^i(Y_t^i - \mathbb{E}[Y_t^i]) + \hat{\mathcal{G}}_t^i \mathbb{E}[Y_t^i] + \mathcal{J}_t^i(Z_t^i - \mathbb{E}[Z_t^i]) + \hat{\mathcal{J}}_t^i \mathbb{E}[Z_t^i] + Z_t^i dW_t \\ Y_T^i & = r^i, \end{cases} \tag{17}$$

where the deterministic coefficients $\mathcal{G}_t^i, \hat{\mathcal{G}}_t^i, \mathcal{J}_t^i, \hat{\mathcal{J}}_t^i \in L^\infty([0, T], \mathbb{R}^{d \times d})$ and the stochastic process $\mathcal{V}^i \in L^2([0, T], \mathbb{R}^d)$ are defined as:

$$\begin{cases} \mathcal{V}_t^i & := -L_{x,t}^i - \Lambda_t^i \bar{\beta}_t - \sigma_{x,t}^\top K_t^i \gamma_t - \bar{\sigma}_{x,t}^\top K_t^i \bar{\gamma}_t - K_t^i (\beta_t - \bar{\beta}_t) - \\ & \quad \sum_{k \neq i} \left(U_{k,t}^i (\alpha_{k,t} - \bar{\alpha}_{k,t}) + V_{k,t}^i \bar{\alpha}_{k,t} \right) + U_{i,t}^{i\top} S_t^{i-1} (L_{k,t}^i - \mathbb{E}[L_{k,t}^i]) \\ & \quad + \sigma_{i,t}^\top K_t^i (\gamma_t - \mathbb{E}[\gamma_t]) + \frac{1}{2} \sum_{k \neq i} (J_{i,k,t}^i + J_{k,i,t}^{i\top}) (\alpha_{k,t} - \mathbb{E}[\alpha_{k,t}]) \\ & \quad + V_{i,t}^{i\top} \hat{S}_t^{i-1} (\mathbb{E}[L_{k,t}^i] + \hat{\sigma}_{i,t}^\top K_t^i - \mathbb{E}[\gamma_t]) + \frac{1}{2} \sum_{k \neq i} (\hat{J}_{i,k,t}^i + \hat{J}_{k,i,t}^{i\top}) \mathbb{E}[\alpha_{k,t}] \\ \mathcal{G}_t^i & := \rho I_d - b_{x,t}^\top + U_{i,t}^{i\top} S_t^{i-1} b_{i,t}^\top \\ \hat{\mathcal{G}}_t^i & = \rho I_d - \hat{b}_{x,t}^\top + V_{i,t}^{i\top} \hat{S}_t^{i-1} \hat{b}_{i,t}^\top \\ \mathcal{J}_t^i & := -\sigma_{x,t}^\top + U_{i,t}^{i\top} S_t^{i-1} \sigma_{k,t}^\top \\ \hat{\mathcal{J}}_t^i & := -\hat{\sigma}_{x,t}^\top + V_{i,t}^{i\top} \hat{S}_t^{i-1} \sigma_{k,t}^\top. \end{cases} \tag{18}$$

By standard results (see Thm. 2.1 in [12]) we obtain that there exists a unique solution $(Y^i, Z^i) \in \mathcal{S}_{\mathbb{P}}^2(\Omega \times [0, T], \mathbb{R}^{d \times d}) \times L_{\mathbb{P}}^2(\Omega \times [0, T], \mathbb{R}^{d \times d})$ to (17).

- (iv) Given $(K^i, \Lambda^i, Y^i, Z^i)$ we consider the equation of R^i which is a linear ODE whose solution is given by:

$$R_t^i = \int_t^T e^{-\rho(s-t)} h_s^i ds,$$

where h^i is a deterministic function defined by:

$$\begin{aligned} h_t^i & = -\mathbb{E}[\gamma_t^\top K_t^i \gamma_t + 2\beta_t^\top Y_t^i + 2\gamma_t^\top Z_t^i + \\ & \quad \sum_{k \neq i} (\alpha_{k,t} - \bar{\alpha}_{k,t})^\top S_{k,t}^i (\alpha_{k,t} - \bar{\alpha}_{k,t}) + \bar{\alpha}_{k,t}^\top \hat{S}_{k,t}^i \bar{\alpha}_{k,t} \\ & \quad + 2[O_{k,t}^i + \xi_{k,t}^i - \bar{\xi}_{k,t}^i]^\top \alpha_{k,t}] \\ & \quad + \mathbb{E} \left[(\xi_{i,t}^i - \bar{\xi}_{i,t}^i)^\top S_t^{i-1} (\xi_{i,t}^i - \bar{\xi}_{i,t}^i) - O_{i,t}^{i\top} \hat{S}_t^{i-1} O_{i,t}^i \right], \end{aligned} \tag{19}$$

and the expressions of the coefficients are recalled in (6).

- (v) The final step is to verify that the procedure to find a fixed point is valid. More precisely we need to ensure that $t \mapsto (\pi_t, \hat{\pi}_t, \eta_t)$ is well defined. It is difficult to ensure the well posedness of $t \mapsto (\pi_t, \hat{\pi}_t)$ for two reasons: first because π and $\hat{\pi}$ follow Riccati equations but are not squared matrices; and second because π appears in the equation followed by $\hat{\pi}$. We are not aware of any work addressing this kind of equations in a general setting.

If we suppose $t \mapsto (\pi_t, \hat{\pi}_t)$ well defined, then $t \mapsto (\eta_t, \phi_t)$ follows a linear mean-field BSDE of the type:

$$\begin{cases} d\eta_t &= \left(\mathcal{V}_t + \mathcal{G}_t(\eta_t - \mathbb{E}[\eta_t]) + \hat{\mathcal{G}}_t \mathbb{E}[\eta_t] + \mathcal{J}_t(\phi_t - \mathbb{E}[\phi_t]) + \hat{\mathcal{J}}_t \mathbb{E}[\phi_t] \right) dt + \phi_t dW_t \\ \eta_T &= r, \end{cases} \tag{20}$$

where the deterministic coefficients $\mathcal{G}_t^i, \hat{\mathcal{G}}_t^i, \mathcal{J}_t^i, \hat{\mathcal{J}}_t^i \in L^\infty([0, T], \mathbb{R}^{nd \times nd})$ and the stochastic process $\mathcal{V}^i \in L^2([0, T], \mathbb{R}^{nd})$ are defined as:

$$\begin{cases} \mathcal{V}_t &:= -\pi(\beta - \bar{\beta}) + F - \bar{F} + P_z \pi(\gamma - \bar{\gamma}) \\ &\quad + (P_\alpha - \pi B + P_z \pi \Sigma)(I_d - S_z \pi \Sigma)^{-1}(H - \bar{H} + S_z \pi(\gamma - \bar{\gamma})) \\ &\quad - \hat{\pi} \bar{\beta} + \bar{F} + \hat{P}_z \pi \bar{\gamma} + (\hat{P}_\alpha - \hat{\pi} \hat{B} + \hat{P}_z \pi \hat{\Sigma})(I_d - \hat{S}_z \pi \hat{\Sigma})^{-1}(\hat{H} + \hat{S}_z \hat{\pi} \bar{\gamma}) \\ \mathcal{G}_t &:= [P_y + (P_\alpha - \pi B + P_z \pi \Sigma)((I_d - S_z \pi \Sigma)^{-1} S_y)] \\ \hat{\mathcal{G}}_t &= \hat{P}_y \\ \mathcal{J}_t &:= [P_z + (P_\alpha - \pi B + P_z \pi \Sigma)((I_d - S_z \pi \Sigma)^{-1} S_z)] \\ \hat{\mathcal{J}}_t &:= -\hat{\sigma}_{x,t}^\top + V_{i,t}^{i\top} \hat{S}_{i,t}^{i-1} \sigma_{k,t}^\top. \end{cases}$$

Again, by standard results (see Thm. 2.1 in [12]) we obtain that there exists a unique solution $(\eta, \phi) \in \mathcal{S}^2(\Omega \times [0, T], \mathbb{R}^{d \times d}) \times L^2_{\mathbb{F}}(\Omega \times [0, T], \mathbb{R}^{d \times d})$ to (20).

To sum up the arguments previously presented, our main result provides the following characterization of the Nash equilibrium:

Theorem 2. *Suppose assumptions (H1) and (H2). Suppose also that the system associated with the fixed point search (16) is well defined. Then $\alpha^* = (\alpha_1, \dots, \alpha_n)$ defined by*

$$\begin{cases} \alpha_t^* - \bar{\alpha}_t^* &= S_{x,t}(X_t^* - \bar{X}_t^*) + S_{y,t}(Y_t - \bar{Y}_t) + S_{z,t}(Z_t - \bar{Z}_t) + H_t - \bar{H}_t \\ \bar{\alpha}_t^* &= \hat{S}_{x,t} \bar{X}_t^* + \hat{S}_{y,t} \bar{Y}_t + \hat{S}_{z,t} \bar{Z}_t + \bar{H}_t \end{cases}$$

where $S_x, S_y, S_z, H, \hat{S}_x, \hat{S}_y, \hat{S}_z, \hat{H}$ are defined in (11), (Y, Z) are in $\mathcal{S}^2_{\mathbb{F}}(\Omega \times [0, T], \mathbb{R}^{nd \times d}) \times L^2_{\mathbb{F}}(\Omega \times [0, T], \mathbb{R}^{nd \times d})$ and satisfy (12) and (16), is a Nash equilibrium.

3 Some extensions

3.1 The case of infinite horizon

Let us now tackle the infinite horizon case. The method is similar to the finite-horizon case but some adaptations are needed when dealing with the well posedness of $(K^i, \Lambda^i, Y^i, R^i)$ and the admissibility of the controls.

We redefine the set of admissible controls for for each player $i \in \llbracket 1, n \rrbracket$ as:

$$\mathcal{A}^i = \left\{ \alpha : \Omega \times [0, T] \rightarrow \mathbb{R}^{d_i} \text{ s.t. } \alpha \text{ is } \mathbb{F}\text{-adapted and } \int_0^\infty e^{-\rho u} \mathbb{E}[|\alpha_u|^2] du < \infty \right\}$$

while the controlled state defined on \mathbb{R}_+ now follows a dynamics of the form

$$\begin{cases} dX_t &= b(t, X_t^\alpha, \mathbb{E}[X_t^\alpha], \alpha_t, \mathbb{E}[\alpha_t])dt + \sigma(t, X_t^\alpha, \mathbb{E}[X_t^\alpha], \alpha_t, \mathbb{E}[\alpha_t])dW_t \\ X_0^\alpha &= X_0 \end{cases} \quad (21)$$

where for each $t \in [0, T], x, \bar{x} \in \mathbb{R}^d, a_i, \bar{a}_i \in \mathbb{R}^{d_i}$:

$$\begin{cases} b(t, x, \bar{x}, a, \bar{a}) &= \beta_t + b_x x + \tilde{b}_x \bar{x} + \sum_{i=1}^n b_i a_i + \tilde{b}_i \bar{a}_i \\ &= \beta_t + b_x x + \tilde{b}_x \bar{x} + B a_t + \tilde{B} \bar{a}_t \\ \sigma(t, x, \bar{x}, a, \bar{a}) &= \gamma_t + \sigma_x x + \tilde{\sigma}_x \bar{x} + \sum_{i=1}^n \sigma_i a_i + \tilde{\sigma}_i \bar{a}_i \\ &= \gamma_t + \sigma_x x + \tilde{\sigma}_x \bar{x} + \Sigma a + \tilde{\Sigma} \bar{a}. \end{cases} \quad (22)$$

Notice that now only the coefficients β and γ are allowed to be stochastic processes. The other linear coefficients are constant matrices.

The goal of each player $i \in \llbracket 1, n \rrbracket$ during the game is still to minimize her cost functional with respect to control α_i over \mathcal{A}^i , and given control α^{-i} of the other players:

$$J^i(\alpha_i, \alpha^{-i}) = \mathbb{E} \left[\int_0^\infty e^{-\rho t} f^i(t, X_t^\alpha, \mathbb{E}[X_t^\alpha], \alpha_t, \mathbb{E}[\alpha_t]) dt \right] \quad (23)$$

where for each $t \in [0, T], x, \bar{x} \in \mathbb{R}^d, a_i, \bar{a}_i \in \mathbb{R}^{d_i}$ we have set the running cost for each player as:

$$\begin{cases} f^i(t, x, \bar{x}, a, \bar{a}) &= (x - \bar{x})^\top Q^i(x - \bar{x}) + \bar{x}^\top [Q^i + \tilde{Q}^i] \bar{x} \\ &+ \sum_{k=1}^n a_k^\top I_k^i(x - \bar{x}) + \bar{a}_k^\top (I_k^i + \tilde{I}_k^i) \bar{x} \\ &+ \sum_{k=1}^n (a_k - \bar{a}_k)^\top N_k^i(a_k - \bar{a}_k) + \bar{a}_k (N_k^i + \tilde{N}_k^i) \bar{a}_k \\ &+ \sum_{0 \leq k \neq l \leq n} (a_k - \bar{a}_k)^\top G_{k,l}^i(a_l - \bar{a}_l) + a_k^\top (G_{k,l}^i + \tilde{G}_{k,l}^i) a_l \\ &+ 2[L_{x,t}^{i\top} x + \sum_{k=1}^n L_{k,t}^{i\top} a_k]. \end{cases}$$

Note that the only coefficients that we allow to be time dependent are L_x^i and L_k^i for $k \in \llbracket 1, n \rrbracket$ which may be stochastic processes.

3.1.1 Assumptions

We detail below the new assumptions:

(H1') The coefficients in (22) satisfy:

- a) $\beta, \gamma \in L^2_{\mathbb{F}}(\Omega \times [0, T], \mathbb{R}^d)$,
- b) $b_x, \tilde{b}_x, \sigma_x, \tilde{\sigma}_x \in \mathbb{R}^{d \times d}; b_i, \tilde{b}_i, \sigma_i, \tilde{\sigma}_i \in \mathbb{R}^{d \times d_i}$.

(H2') The coefficients of the cost functional (3) satisfy:

- a) $Q^i, \tilde{Q}^i \in \mathbb{S}^d_+; N^i_k, \tilde{N}^i_k \in \mathbb{S}^{d_k}_+; I^i_k, \tilde{I}^i_k \in \mathbb{R}^{d_k \times d}$,
- b) $L^i_x \in L^2_{\mathbb{F}}(\Omega \times \mathbb{R}^*_+, \mathbb{R}^d), L^i_k \in L^2_{\mathbb{F}}(\Omega \times \mathbb{R}^*_+, \mathbb{R}^{d_k})$,
- c) $N^i_k > 0 \quad Q^i - I^{i\top}_k (N^i_k)^{-1} I^i_k \geq 0$,
- d) $\tilde{N}^i_{k,t} > 0 \quad \hat{Q}^i - \hat{I}^{i\top}_t (\hat{N}^i_t)^{-1} \hat{I}^i_t \geq 0$.

(H3') $\rho > 2 (|b_x| + |\tilde{b}_x| + 8(|\sigma_x|^2 + |\tilde{\sigma}_x|^2))$.

As shown below, the new hypothesis **(H3')** ensure the well posedness of our problem. Notice first that by **(H1')** and classical results, there exists a unique strong solution X^α to the SDE (21). Furthermore by **(H1')** and **(H3')** we obtain by similar arguments as in [2] the following estimate:

$$\int_0^\infty e^{-\rho u} \mathbb{E}[|X_u^\alpha|^2] du \leq C_\alpha (1 + \mathbb{E}[|X_0|^2]) < \infty, \tag{24}$$

in which C_α is a constant depending on $\alpha = (\alpha_1, \dots, \alpha_n)$ only through $\int_0^\infty e^{-\rho} \mathbb{E}[|\alpha_u|^2] du$. Finally by **(H2')** and (24) the minimizing problem (23) is well defined for each player.

3.1.2 A weak submartingale optimality principle on infinite horizon

We now give an easy adaptation of the weak submartingale optimality principle in the case of infinite horizon.

Lemma 2 (Weak submartingale optimality principle). *Suppose there exists a couple*

$(\alpha^*, (\mathcal{W}^{\cdot,i})_{i \in \llbracket 1, n \rrbracket})$, *where* $\alpha^* \in \mathcal{A}$ *and* $\mathcal{W}^{\cdot,i} = \{\mathcal{W}_t^{\alpha^*,i}, t \in \mathbb{R}^*_+, \alpha \in \mathcal{A}\}$ *is a family of adapted processes indexed by* \mathcal{A} *for each* $i \in \llbracket 1, n \rrbracket$, *such that:*

- (i) *For every* $\alpha \in \mathcal{A}$, $\mathbb{E}[\mathcal{W}_0^{\alpha,i}]$ *is independent of the control* $\alpha_i \in \mathcal{A}^i$;
- (ii) *For every* $\alpha \in \mathcal{A}$, $\lim_{t \rightarrow \infty} e^{-\rho t} \mathbb{E}[\mathcal{W}_t^{\alpha,i}] = 0$;
- (iii) *For every* $\alpha \in \mathcal{A}$, *the map* $t \in \mathbb{R}^*_+ \mapsto \mathbb{E}[\mathcal{S}_t^{\alpha,i}]$, *with*

$$\mathcal{S}_t^{\alpha,i} = e^{-\rho t} \mathcal{W}_t^{\alpha,i} + \int_0^t e^{-\rho u} f^i(u, X_u^\alpha, \mathbb{P}_{X_u^\alpha}, \alpha_u, \mathbb{P}_{\alpha_u}) du$$
is well defined and non-decreasing;
- (iv) *The map* $t \mapsto \mathbb{E}[\mathcal{S}_t^{\alpha^*,i}]$ *is constant for every* $t \in \mathbb{R}^*_+$;

Then α^* is a Nash equilibrium and $J^i(\alpha^*) = \mathbb{E}[\mathcal{W}_0^{\alpha^*,i}]$. Moreover, any other Nash-equilibrium $\tilde{\alpha}$ such that $\mathbb{E}[\mathcal{W}_0^{\tilde{\alpha},i}] = \mathbb{E}[\mathcal{W}_0^{\alpha^*,i}]$ and $J^i(\tilde{\alpha}) = J^i(\alpha^*)$ for any $i \in \llbracket 1, n \rrbracket$ satisfies the condition (iv).

Proof. The proof is exactly the same as in Lemma 1.

Let us now describe the steps to follow in order to apply Lemma 2. Since they are similar to the ones in the finite-horizon case, we only report the main changes.

Steps 1-3

For each player $i \in \llbracket 1, n \rrbracket$ we still search for a random field $\{w_t^i(x, \bar{x}), t \in [0, T], x, \bar{x} \in \mathbb{R}^d\}$ of the form $w_t^i(x, \bar{x}) = K_t^i \cdot (x - \bar{x})^{\otimes 2} + \Lambda_t^i \cdot \bar{x}^{\otimes 2} + 2Y_t^{i\top} x + R_t^i$ for which the optimality principle in Lemma 2 now leads to the system:

$$\begin{cases} dK_t^i = -\Phi_t^{i0} dt \\ d\Lambda_t^i = -\Psi_t^{i0} dt \\ dY_t^i = -\Delta_t^{i0} dt + Z_t^i dW_t \\ dR_t^i = -\Gamma_t^{i0} dt, \quad t \geq 0. \end{cases} \tag{25}$$

Notice that there are no terminal conditions anymore since we are in the infinite horizon case. The coefficients $\Phi^{i0}, \Psi^{i0}, \Delta_t^{i0}, \Gamma_t^{i0}$ are defined in (7). The fourth step is exactly the same as in the finite horizon case.

Step 5

We now search for a fixed point of the best response functions. Let us define $Y = (Y^1, \dots, Y^n)$ and propose an ansatz in a feedback form $Y: Y_t = \pi(X_t - \mathbb{E}[X_t]) + \hat{\pi}\mathbb{E}[X_t] + \eta_t$ where $\pi, \hat{\pi} \in L^\infty(\mathbb{R}_+, \mathbb{R}^{nd \times d})$ and $\eta \in L_{\mathbb{F}}^2(\Omega \times \mathbb{R}_+, \mathbb{R}^{nd})$ satisfy

$$\begin{cases} 0 &= -\pi b_x + P_y \pi + P_z \pi \sigma_x + (P_\alpha + P_z \Sigma)(Id - S_z \pi \Sigma)^{-1}(S_x + S_y \pi + S_z \pi \sigma_x) \\ &\quad - \pi B (Id - S_z \pi \Sigma)^{-1}(S_x + S_y \pi + S_z \pi \sigma_x) \\ 0 &= -\hat{\pi} \hat{b}_x + \hat{P}_y \hat{\pi} + \hat{P}_z \pi \hat{\sigma}_x + (\hat{P}_\alpha + \hat{P}_z \hat{\Sigma})(Id - \hat{S}_z \pi \hat{\Sigma})^{-1}(\hat{S}_x + \hat{S}_y \hat{\pi} + \hat{S}_z \pi \hat{\sigma}_x) \\ &\quad - \hat{\pi} \hat{B} (Id - \hat{S}_z \pi \hat{\Sigma})^{-1}(\hat{S}_x + \hat{S}_y \hat{\pi} + \hat{S}_z \pi \hat{\sigma}_x) \\ d\eta_t &= \psi_t dt + \phi_t dW_t \\ &\text{with:} \\ \psi_t - \bar{\psi}_t &= -\pi_t(\beta - \bar{\beta}) + (P_\alpha - \pi_t B)(R - \bar{R})(F - \bar{F}) + \\ &\quad + P_z(\phi_t - \bar{\phi}_t + \pi_t(\gamma - \bar{\gamma} + \Sigma(R_t - \bar{R}_t))) + P_y(\eta_t - \bar{\eta}_t) \\ \bar{\psi}_t &= -\hat{\pi}_t \bar{\beta} + (\hat{P}_\alpha - \hat{\pi}_t \hat{B}) \bar{R}_t + \bar{F} + \hat{P}_z(\bar{\phi} + \pi_t(\bar{\gamma} + \hat{\Sigma} \bar{R}_t)) + \hat{P}_y \bar{\eta}_t \\ R_t &:= (Id - S_z \pi_t \Sigma)^{-1}(H + S_y \eta_t + S_z(\phi_t + \pi_t \gamma)) \\ \hat{R}_t &:= (Id - \hat{S}_z \pi_t \hat{\Sigma})^{-1}(\hat{H} + \hat{S}_y \eta_t + \hat{S}_z(\phi_t + \pi_t \gamma)) \end{cases} \tag{26}$$

where the coefficients are defined in (11).

Step 6

We finally tackle the well-posedness of (25) and (26).

- (i) We first consider the ODE for K^i . Since the map $(t, k) \mapsto \phi_t^{i0}(k)$ does not depend on time (all the coefficient being constant) we search for a constant non-negative

matrix $K^i \in \mathbb{S}^d$ satisfying $\Phi^{i0}(K^i) = 0$, more precisely solution to:

$$\begin{aligned} Q^i + \sigma_x^T K^i \sigma_x + K^i b_x + b_x^T K^i - \rho K^i \\ - (I_k^i + \sigma_k^T K^i \sigma_x + b_k^T K^i)(N_k^i + \sigma_k^T K^i \sigma_k)^{-1}(I_k^i + \sigma_k^T K^i \sigma_x + b_k^T K^i) = 0. \end{aligned} \tag{27}$$

As in [2] we can show using a limit argument that there exists $K^i \in \mathbb{S}_+^d$ solution to (27). The argument for Λ^i is the same as for K^i .

- (ii) Given (K^i, Λ^i) the equation for (Y^i, Z^i) is a linear mean-field BSDE on infinite horizon:

$$dY_t^i = \mathcal{V}_t^i + \mathcal{G}^i(Y_t^i - \mathbb{E}[Y_t^i]) + \hat{\mathcal{G}}^i \mathbb{E}[Y_t^i] + \mathcal{J}^i(Z_t^i - \mathbb{E}[Z_t^i]) + \hat{\mathcal{J}}^i \mathbb{E}[Z_t^i] + Z_t^i dW_t,$$

where the coefficient are defined in (18). Notice that now $\mathcal{G}^i, \hat{\mathcal{G}}^i, \mathcal{J}^i, \hat{\mathcal{J}}^i$ are all constant matrices. To the best of our knowledge, there are no general results ensuring the existence for such equation. We then add the following assumption:

(H4') There exists a solution $(Y^i, Z^i) \in \times \mathcal{S}_{\mathbb{F}}^2(\Omega \times \mathbb{R}_+, \mathbb{R}^d) \times L_{\mathbb{F}}^2(\Omega \times \mathbb{R}_+, \mathbb{R}^d)$

- (iii) Given $(K^i, \Lambda^i, Y^i, Z^i)$ the equation for R^i is a linear ODE whose solution is:

$$R_t^i = \int_t^\infty e^{-\rho(s-t)} h_s^i ds,$$

where h^i is a deterministic function defined in (19).

- (iv) We now study the well posedness of the fixed point procedure. More precisely we need to ensure that the process $t \rightarrow (\pi, \hat{\pi}, \eta_t)$ defined as a solution of the system (26), recalled below, is well defined. Note that in the infinite horizon framework we search for constant π and $\hat{\pi}$.

$$\begin{cases} 0 &= -\pi b_x + P_y \pi + P_z \pi \sigma_x + (P_\alpha + P_z \Sigma)(Id - S_z \pi \Sigma)^{-1}(S_x + S_y \pi + S_z \pi \sigma_x) \\ &\quad - \pi B (Id - S_z \pi \Sigma)^{-1}(S_x + S_y \pi + S_z \pi \sigma_x) \\ 0 &= -\hat{\pi} \hat{b}_x + \hat{P}_y \hat{\pi} + \hat{P}_z \pi \hat{\sigma}_x + (\hat{P}_\alpha + \hat{P}_z \hat{\Sigma})(Id - \hat{S}_z \pi \hat{\Sigma})^{-1}(\hat{S}_x + \hat{S}_y \hat{\pi} + \hat{S}_z \pi \hat{\sigma}_x) \\ &\quad - \hat{\pi} \hat{B} (Id - \hat{S}_z \pi \hat{\Sigma})^{-1}(\hat{S}_x + \hat{S}_y \hat{\pi} + \hat{S}_z \pi \hat{\sigma}_x) \\ d\eta_t &= \psi_t dt + \phi_t dW_t, \quad t \geq 0. \end{cases} \tag{28}$$

Existence of $(\pi, \hat{\pi})$ in whole generality is a difficult problem. Let us first rewrite the system (28) as:

$$F((\pi, \hat{\pi}), \mathcal{C}) = 0$$

Where $\mathcal{C} = (b_x, \sigma_x, B, \Sigma, \hat{b}_x, \hat{\sigma}_x, \hat{B}, \hat{\Sigma}, S_x, S_y, S_z, \hat{S}_x, \hat{S}_y, \hat{S}_z, P_y, P_z, P_\alpha, \hat{P}_y, \hat{P}_z, \hat{P}_\alpha)$. Note that F is continuously differentiable on its domain of definition. Thus, if $\left[\frac{\partial F}{\partial \pi}, \frac{\partial F}{\partial \hat{\pi}} \right] (\pi, \hat{\pi}, \mathcal{C})$ is invertible for, then, by the implicit function theorem, there exists an open set U containing \mathcal{C} and a continuously differentiable function $g : U \mapsto (\mathbb{R}^{nd \times d})^2$ such that for all admissible coefficients $\mathcal{C} \in U : F(g(\mathcal{C}), \mathcal{C}) = 0$ and the solutions $(\pi, \hat{\pi}) = g(\mathcal{C})$ are unique. It means that if we find a solution

to (28) while the condition $\left[\frac{\partial F}{\partial \pi}, \frac{\partial F}{\partial \hat{\pi}}\right]$ is invertible, then for small perturbations on the coefficients we still have solutions for $(\pi, \hat{\pi})$.

Let us now give sufficient conditions to ensure the existence of $(\pi, \hat{\pi})$ in a simplified setting where the state $t \rightarrow X_t$ belongs to \mathbb{R} and all the players are symmetric in the sense that all the coefficients associated with each player are equals ($b_1 = \dots = b_2, Q^1 = \dots = Q^n, etc.$). We suppose also that the volatility is not controlled i.e. $\sigma_i = 0$ for all $i \in \llbracket 1, n \rrbracket$. In such a case $\pi, \hat{\pi} \in \mathbb{R}^{n \times 1}, \pi_1 = \dots = \pi_n, \hat{\pi}_n = \dots = \hat{\pi}_1$ and the systems (28) of coupled equations now reduces to two coupled second order equations:

$$\begin{cases} \pi_1^2 [nb_1 S_{y,1}] + \pi_1 [-b_x + P_{y,1} + P_{z,1} \sigma_x + \sum_{j \neq 1} P_{\alpha,1,j} S_{y,1} - nb_1 S_{x,1}] \\ + \sum_{j \neq 1} P_{\alpha,1,j} S_{x,j} = 0 \\ \hat{\pi}_1^2 [n\hat{b}_1 \hat{S}_{y,1}] + \hat{\pi}_1 [-\hat{b}_x + \hat{P}_{y,1} + \sum_{j \neq 1} \hat{P}_{\alpha,1,j} \hat{S}_{y,1} - n\hat{b}_1 \hat{S}_{x,1}] \\ + \sum_{j \neq 1} \hat{P}_{\alpha,1,j} \hat{S}_{x,j} + \pi_1 \hat{P}_{z,1} \hat{\sigma}_x = 0. \end{cases}$$

If we note:

$$\begin{cases} a & := nb_1 S_{y,1} \\ b & := -b_x + P_{y,1} + P_{z,1} \sigma_x + \sum_{j \neq 1} P_{\alpha,1,j} S_{y,1} - nb_1 S_{x,1} \\ c & := \sum_{j \neq 1} P_{\alpha,1,j} S_{x,j}, \end{cases}$$

then a sufficient condition for π_1 to exists is simply: $b^2 - 4ac \geq 0$. Since $a \leq 0$ and $c \geq 0$ we have two possibilities a priori for π_1 . We choose the positive one to ensure that $\alpha \in \mathcal{A}$. Then if we note:

$$\begin{cases} \hat{a} & := n\hat{b}_1 \hat{S}_{y,1} \\ \hat{b} & := -\hat{b}_x + \hat{P}_{y,1} + \sum_{j \neq 1} \hat{P}_{\alpha,1,j} \hat{S}_{y,1} - n\hat{b}_1 \hat{S}_{x,1} \\ \hat{c}(\pi_1) & := \sum_{j \neq 1} \hat{P}_{\alpha,1,j} \hat{S}_{x,j} + \pi_1 \hat{P}_{z,1} \hat{\sigma}_x, \end{cases}$$

a sufficient condition for $\hat{\pi}_1$ to exist is $\hat{b}^2 - 4\hat{a}\hat{c}(\pi_1) \geq 0$. To ensure that there is a positive solution we also need $\hat{c}(\pi_1) \geq 0$.

- (v) Let us finally verify that $\alpha^* \in \mathcal{A}$. Let us consider the candidate for the optimal control for each player:

$$\begin{aligned} \alpha^* - \bar{\alpha}^* &= A_x(X - \bar{X}) + R_t - \bar{R}_t \\ \bar{\alpha}^* &= \hat{A}_x \bar{X} + \bar{R}_t \end{aligned}$$

where the coefficients are defined in (15) and X^* is the state process optimally controlled. Since $R, \hat{R} \in L^2_{\mathbb{F}}(\Omega \times \mathbb{R}_+, \mathbb{R}^{\sum_i d_i})$ and given that the coefficient are constant in the infinite horizon case, we need to verify that:

$$\int_0^\infty e^{-\rho u} \mathbb{E}[|X_u^* - \mathbb{E}[X_u^*]|^2] du < \infty \quad \int_0^\infty e^{-\rho u} |\mathbb{E}[X_u^*]|^2 du < \infty.$$

As we will see below, we will have to choose ρ large enough to ensure these conditions. From the above expressions we see that X^* satisfies:

$$\begin{cases} dX_t^* &= b_t^* dt + \sigma_t^* dW_t \\ X_0 &= x_0 \end{cases}$$

with:

$$b_t^* = \beta_t^* + B^*(X_t^* - \mathbb{E}[X_t^*]) + \hat{B}^* \mathbb{E}[X_t^*] \quad \sigma_t^* = \gamma_t^* + \Sigma^*(X_t^* - \mathbb{E}[X_t^*]) + \hat{\Sigma}^* \mathbb{E}[X_t^*]$$

where we define

$$\begin{aligned} B^* &= b_x + BA_x, & \hat{B}^* &= \hat{b}_x + \hat{B}\hat{A}_x, & \Sigma^* &= \sigma_x + \Sigma A_x, \\ \hat{\Sigma}^* &= \hat{\sigma}_x + \hat{\Sigma}\hat{A}_x, \\ \beta_t^* &= \beta_t + B(R_t - \mathbb{E}[R_t]) + \hat{B}\bar{R}_t, \\ \gamma_t^* &= \sigma_t + \Sigma(R_t - \mathbb{E}[R_t]) + \hat{\Sigma}\bar{R}_t. \end{aligned}$$

By Itô's formula we have:

$$\begin{aligned} \frac{d}{dt} e^{-\rho t} |\bar{X}_t^*|^2 &\leq e^{-\rho t} \left(-\rho |\bar{X}_t^*|^2 + 2(\bar{b}_t^*)^\top \bar{X}_t^* \right) \\ &\leq e^{-\rho t} \left(|\bar{X}_t^*|^2 (-\rho + 2\hat{B}^*) + 2|\bar{X}_t^*| |\bar{\beta}_t^*| \right) \\ &\leq e^{-\rho t} \left(|\bar{X}_t^*|^2 (-\rho + 2\hat{B}^* + \varepsilon) + \frac{1}{\varepsilon} |\bar{\beta}_t^*|^2 \right). \end{aligned}$$

If we now set:

$$K = |\mathbb{E}[X_0^*]|^2 + \frac{1}{\varepsilon} \int_0^\infty e^{-\rho u} |\bar{\beta}_t^*|^2 du, \quad C = -\rho + 2\hat{B}^* + \varepsilon,$$

then, by Grownall inequality we obtain:

$$e^{-\rho t} |\mathbb{E}[X_t^*]|^2 \leq Ke^{Ct}.$$

Therefore, in order to have $\int_0^\infty e^{-\rho u} |\mathbb{E}[X_u^*]|^2 du < \infty$, we shall impose that $\rho > 2\hat{B}^*$. Finally, by Itô's formula we also have:

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[e^{-\rho t} |X_t^* - \bar{X}_t^*|^2] &\leq e^{-\rho t} \mathbb{E} \left[-\rho |X_t^* - \bar{X}_t^*|^2 + 2(b_t^* - \bar{b}_t^*)^\top (X_t^* - \bar{X}_t^*) + |\sigma_t^*|^2 \right] \\ &\leq e^{-\rho t} \mathbb{E} \left[|X_t^* - \bar{X}_t^*|^2 (-\rho + 2B^* + \varepsilon) + \frac{1}{\varepsilon} |\beta_t^* - \bar{\beta}_t^*|^2 \right. \\ &\quad \left. + 4(|\gamma_t^*|^2 + |\Sigma^*|^2 |X_t^* - \mathbb{E}[X_t^*]|^2 + |\hat{\Sigma}^*|^2 |\mathbb{E}[X_t^*]|^2) \right] \\ &\leq e^{-\rho t} \mathbb{E} \left[|X_t^* - \bar{X}_t^*|^2 (-\rho + 2B^* + \varepsilon + 4|\Sigma^*|^2) \right. \\ &\quad \left. + \frac{1}{\varepsilon} |\beta_t^* - \bar{\beta}_t^*|^2 + 4(|\gamma_t^*|^2 + |\hat{\Sigma}^*|^2 |\mathbb{E}[X_t^*]|^2) \right]. \end{aligned}$$

If we now set:

$$\begin{aligned} K &= |X_0^* - \bar{X}_0^*|^2 + \int_0^\infty e^{-\rho u} \mathbb{E} \left[\frac{1}{\varepsilon} |\beta_u^* - \bar{\beta}_u^*|^2 + 4(|\gamma_u^*|^2 + |\hat{\Sigma}^*|^2 |\mathbb{E}[X_u^*]|^2) \right] du \\ C &= -\rho + 2B^* + 4|\Sigma^*|^2 + \varepsilon, \end{aligned}$$

then, by Gronwall inequality we obtain:

$$e^{-\rho t} \mathbb{E} [|X_t^* - \bar{X}_t^*|^2] \leq Ke^{-Ct}.$$

This time in order to ensure the convergence $\int_0^\infty e^{-\rho u} \mathbb{E} [|X_u^* - \bar{X}_u^*|^2] du < \infty$, we will add the constraint $\rho > 2B^* + 4|\Sigma^*|^2$. To conclude, in order to ensure that $\alpha^* \in \mathcal{A}$ we make the following assumption:

(H5*) $\rho > \max [2\hat{B}^*, 2B^* + 4|\Sigma^*|^2].$

3.2 The case of common noise

Let W and W^0 be two independent Brownian motions defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ where $\mathbb{F} = \{\mathcal{F}_t\}_{t \in [0, T]}$ is the filtration generated by the pair (W, W^0) . Let $\mathbb{F}^0 = \{\mathcal{F}_t^0\}_{t \in [0, T]}$ be the filtration generated by W^0 . For any X_0 and $\alpha \in \mathcal{A}$ as in Section 1, the controlled process X^α is defined by:

$$\begin{cases} dX_t &= b(t, X_t^\alpha, \mathbb{E}[X_t^\alpha | W^0], \alpha_t, \mathbb{E}[\alpha_t | W^0])dt \\ &\quad + \sigma(t, X_t^\alpha, \mathbb{E}[X_t^\alpha | W^0], \alpha_t, \mathbb{E}[\alpha_t | W^0])dW_t \\ &\quad + \sigma^0(t, X_t^\alpha, \mathbb{E}[X_t^\alpha | W^0], \alpha_t, \mathbb{E}[\alpha_t | W^0])dW_t^0 \\ X_0^\alpha &= X_0 \end{cases}$$

where for each $t \in [0, T], x, \bar{x} \in \mathbb{R}^d, a_i, \bar{a}_i \in \mathbb{R}^{d_i}$:

$$\begin{cases} b(t, x, \bar{x}, \alpha, \bar{\alpha}) &= \beta_t + b_{x,t}x + \tilde{b}_{x,t}\bar{x} + \sum_{i=1}^n b_{i,t}\alpha_i + \tilde{b}_{i,t}\bar{\alpha}_i \\ &= \beta_t + b_{x,t}x + \tilde{b}_{x,t}\bar{x} + B_t\alpha + \tilde{B}_t\bar{\alpha} \\ \sigma(t, x, \bar{x}, \alpha, \bar{\alpha}) &= \gamma_t + \sigma_{x,t}x + \tilde{\sigma}_{x,t}\bar{x} + \sum_{i=1}^n \sigma_{i,t}\alpha_i + \tilde{\sigma}_{i,t}\bar{\alpha}_i \\ &= \gamma_t + \sigma_{x,t}x + \tilde{\sigma}_{x,t}\bar{x} + \Sigma_t\alpha + \tilde{\Sigma}_t\bar{\alpha} \\ \sigma^0(t, x, \bar{x}, \alpha, \bar{\alpha}) &= \gamma_t^0 + \sigma_{x,t}^0x + \tilde{\sigma}_{x,t}^0\bar{x} + \sum_{i=1}^n \sigma_{i,t}^0\alpha_i + \tilde{\sigma}_{i,t}^0\bar{\alpha}_i \\ &= \gamma_t^0 + \sigma_{x,t}^0x + \tilde{\sigma}_{x,t}^0\bar{x} + \Sigma_t^0\alpha + \tilde{\Sigma}_t^0\bar{\alpha}. \end{cases}$$

Since we will condition on W^0 , we assume that the coefficients $b_x, \tilde{b}_x, b_i, \tilde{b}_i, \sigma_x, \tilde{\sigma}_x, \sigma_i, \tilde{\sigma}_i, \sigma_x^0, \tilde{\sigma}_x^0, \sigma_i^0, \tilde{\sigma}_i^0$ are essentially bounded and \mathbb{F}^0 -adapted processes, whereas β, γ, γ^0 are square integrable \mathbb{F}^0 -adapted processes. The problem of each player i is to minimize over $\alpha_i \in \mathcal{A}^i$, and given control α^{-i} of the other players, a cost functional of the form

with f^i, g^i as in (3). We now suppose that $Q^i, \tilde{Q}^i, I^i, \tilde{I}^i, N^i, \tilde{N}^i$ are essentially bounded and \mathbb{F}^0 -adapted, L_x^i, L_k^i are square-integrable \mathbf{F} -adapted processes, P^i, \tilde{P}^i are essentially bounded \mathcal{F}_T^0 -measurable random variable and r^i are square-integrable \mathcal{F}_T -measurable random variable. Hypothesis c) and d) of (H2) still holds. As in step 1 we guess a random field of the type $\mathcal{W}_t^{\alpha, i} = w_t^i(X_t^\alpha, \mathbb{E}[X_t^\alpha | W_t^0])$ where w^i is of the form $w_t^i(x, \bar{x}) = K_t^i \cdot (x - \bar{x})^{\otimes 2} + \Lambda_t^i \cdot \bar{x}^{\otimes 2} + 2Y_t^{i\top}x + R_t^i$ with suitable coefficients K^i, Λ^i, Y^i, R^i . Given that the quadratic coefficient in f^i, g^i are \mathbb{F}^0 -adapted we guess that K^i, Λ^i are also \mathbb{F}^0 -adapted. Since the linear coefficients in f^i, g^i and the affine coefficients in b, σ, σ^0 are \mathbb{F} -adapted, we guess that Y^i is \mathbb{F} -adapted as well. Thus for each player we look for processes $(K^i, \Lambda^i, Y^i, R^i)$ valued in $\mathbb{S}_+^d \times \mathbb{S}_+^d \times \mathbb{R}^d \times \mathbb{R}$ and of the form:

$$\begin{cases} dK_t^i = \dot{K}_t^i dt + Z_t^{K^i} dW_t^0 & K_T^i = P^i \\ d\Lambda_t^i = \dot{\Lambda}_t^i dt + Z_t^{\Lambda^i} dW_t^0 & \Lambda_T^i = P^i + \tilde{P}^i \\ dY_t^i = \dot{Y}_t^i dt + Z_t^i dW_t + Z_t^{0, Y^i} dW_t^0 & Y_T^i = r^i \\ dR_t^i = \dot{R}_t^i dt & R_T^i = 0 \end{cases}$$

where $\dot{K}^i, \dot{\Lambda}^i, Z^{K^i}, Z^{\Lambda^i}$ are \mathbb{F}^0 -adapted processes valued in \mathbb{S}^d ; $\dot{Y}^i, Z^{Y^i}, Z^{0, Y^i}$ are \mathbb{F} -adapted processes valued in \mathbb{R}^d and R^i are continuous functions valued in \mathbb{R} . In step 2 we now consider, for each player $i \in \llbracket 1, n \rrbracket$, a family of processes of the form:

$$\mathcal{S}_t^{\alpha, i} = e^{-\rho t} w_t^i(X_t^\alpha, \mathbb{E}[X_t^\alpha | W_t^0]) + \int_0^t e^{-\rho u} f^i(u, X_u^\alpha, \mathbb{E}[X_u^\alpha | W_u^0]), \alpha_u, \mathbb{E}[\alpha_u | W_u^0]) du.$$

By Itô’s formula we then obtain for (5) and (6):

$$\left\{ \begin{aligned} \Phi_t^i &= Q_t^i + \sigma_{x,t}^\top K_t^i \sigma_{x,t} + \sigma_t^{0\top} K_t^i \sigma_t^0 + K_t^i b_{x,t} + b_{x,t}^\top K_t^i + Z_t^{K^i} \sigma_{x,t}^0 + \sigma_{x,t}^{0\top} Z_t^{K^i} - \rho K_t^i \\ \Psi_t^i &= \hat{Q}_t^i + \hat{\sigma}_{x,t}^\top K_t^i \hat{\sigma}_{x,t} + \hat{\sigma}_{x,t}^{0\top} \Lambda_t^i \hat{\sigma}_{x,t}^0 + \Lambda_t^i \hat{b}_{x,t} + \hat{b}_{x,t}^\top \Lambda_t^i + Z_t^{\Lambda^i} \hat{\sigma}_{x,t}^0 + \hat{\sigma}_{x,t}^{0\top} Z_t^{\Lambda^i} - \rho \Lambda_t^i \\ \Delta_t^i &= L_{x,t}^i + b_{x,t}^\top Y_t^i + \tilde{b}_{x,t}^\top \bar{Y}_t^i + \sigma_{x,t}^\top Z_t^i + \tilde{\sigma}_{x,t}^\top \bar{Z}_t^i + \sigma_{x,t}^{0\top} Z_t^{0,Y^i} + \tilde{\sigma}_{x,t}^{0\top} \bar{Z}_t^{0,Y^i} + \Lambda_t^i \bar{\beta}_t \\ &\quad + \sigma_{x,t}^\top K_t^i \gamma_t + \tilde{\sigma}_{x,t}^\top K_t^i \bar{\gamma}_t + K_t^i (\beta_t - \bar{\beta}_t) + Z_t^{K^i} (\gamma_t^0 - \bar{\gamma}_t^0) + Z_t^{\Lambda^i} \bar{\gamma}_t^0 - \rho Y_t^i \\ &\quad + \sum_{k \neq i} U_{k,t}^{i\top} (\alpha_{k,t} - \bar{\alpha}_{k,t}) + V_{k,t}^{i\top} \bar{\alpha}_{k,t} \\ \Gamma_t^i &= \gamma_t^\top K_t^i \gamma_t + \bar{\gamma}_t^{0\top} \Lambda_t^i \bar{\gamma}_t^0 + (\gamma_t^0 - \bar{\gamma}_t^0)^\top K_t^i (\gamma_t^0 - \bar{\gamma}_t^0) + 2\beta_t^\top Y_t^i + 2\gamma_t^\top Z_t^i + 2\gamma_t^{0\top} Z_t^{0,Y^i} \\ &\quad + \sum_{k \neq i} (\alpha_{k,t} - \bar{\alpha}_{k,t})^\top S_{k,t}^i (\alpha_{k,t} - \bar{\alpha}_{k,t}) \\ &\quad + \bar{\alpha}_{k,t}^\top \hat{S}_{k,t}^i \bar{\alpha}_{k,t} + 2[O_{k,t}^i + \xi_{k,t}^i - \bar{\xi}_{k,t}^i]^\top \alpha_{k,t} - \rho R_t^i, \end{aligned} \right.$$

with

$$\left\{ \begin{aligned} S_{k,t}^i &= N_{k,t}^i + \sigma_{k,t}^\top K_t^i \sigma_{k,t} + (\sigma_{k,t}^0)^\top K_t^i \sigma_{k,t}^0 \\ \hat{S}_{k,t}^i &= \hat{N}_{k,t}^i + \hat{\sigma}_{k,t}^\top K_t^i \hat{\sigma}_{k,t} + (\hat{\sigma}_{k,t}^0)^\top \Lambda_t^i \hat{\sigma}_{k,t}^0 \\ U_{k,t}^i &= \hat{I}_{k,t}^i + \sigma_{k,t}^\top K_t^i \sigma_{x,t} + (\sigma_{k,t}^0)^\top K_t^i \sigma_{x,t}^0 + (\sigma_{k,t}^0)^\top Z_t^{K^i} + b_{k,t}^\top K_t^i \\ V_{k,t}^i &= \tilde{I}_{k,t}^i + \hat{\sigma}_{k,t}^\top K_t^i \hat{\sigma}_{x,t} + (\hat{\sigma}_{k,t}^0)^\top \Lambda_t^i \hat{\sigma}_{x,t}^0 + (\hat{\sigma}_{k,t}^0)^\top Z_t^{\Lambda^i} + \hat{b}_{k,t}^\top \Lambda_t^i \\ O_{k,t}^i &= \bar{L}_{k,t}^i + \hat{b}_{k,t}^\top \bar{Y}_t^i + \hat{\sigma}_{k,t}^\top \bar{Z}_t^i + \hat{\sigma}_{k,t}^\top K_t^i \bar{\gamma}_t + (\hat{\sigma}_{k,t}^0)^\top \Lambda_t^i \bar{\gamma}_t^0 + (\hat{\sigma}_{k,t}^0)^\top Z_t^{0,Y^i} \\ &\quad + \frac{1}{2} \sum_{k \neq i} (\hat{J}_{i,k,t}^i + \hat{J}_{k,i,t}^{i\top}) \bar{\alpha}_{k,t} \\ J_{k,l,t}^i &= G_{k,l,t}^i + \sigma_{k,t}^\top K_t^i \sigma_{l,t} + (\sigma_{k,t}^0)^\top K_t^i \sigma_{l,t}^0 \\ \hat{J}_{k,l,t}^i &= \hat{G}_{k,l,t}^i + \hat{\sigma}_{k,t}^\top K_t^i \hat{\sigma}_{l,t} + (\hat{\sigma}_{k,t}^0)^\top \Lambda_t^i \hat{\sigma}_{l,t}^0 \\ \xi_{k,t}^i &= L_{k,t}^i + b_{k,t}^\top Y_t^i + \sigma_{k,t}^\top Z_t^i + (\sigma_{k,t}^0)^\top Z_t^{0,Y^i} + \sigma_{k,t}^\top K_t^i \gamma_t + (\sigma_{k,t}^0)^\top K_t^i \gamma_t^0 \\ &\quad + \frac{1}{2} \sum_{k \neq i} (J_{i,k,t}^i + J_{k,i,t}^{i\top}) \alpha_{k,t}. \end{aligned} \right.$$

Note that we now denote by \bar{U} the conditional expectation with respect to W_t^0 , i.e. $\bar{U} = \mathbb{E}[U|W_t^0]$. Then, at **step 3**, we constraint the coefficients $(K^i, \Lambda^i, Y^i, R^i)$ to satisfy the following problem:

$$\left\{ \begin{aligned} dK_t^i &= -\Phi_t^{i0} dt + Z_t^{K^i} dW_t^0 & K_T^i &= P^i \\ d\Lambda_t^i &= -\Psi_t^{i0} dt + Z_t^{\Lambda^i} dW_t^0 & \Lambda_T^i &= P^i + \tilde{P}^i \\ dY_t^i &= -\Delta_t^{i0} dt + Z_t^i dW_t + Z_t^{0,Y^i} dW_t^0 & Y_T^i &= r^i \\ dR_t^i &= -\Gamma_t^{i0} dt & R_T^i &= 0 \end{aligned} \right. \tag{29}$$

where $\Phi^{i0}, \Psi^{i0}, \Delta^{i0}, \Gamma^{i0}$ are defined in (8). Thus we obtain the best response functions of the players:

$$\alpha_{i,t} = -S_{i,t}^{i-1} U_{i,t}^i (X_t - \mathbb{E}[X_t|W_t^0]) - S_{i,t}^{i-1} (\xi_{i,t}^i - \bar{\xi}_{i,t}^i) - \hat{S}_{i,t}^{i-1} (V_{i,t}^i \mathbb{E}[X_t|W_t^0] + O_{i,t}^i).$$

We then proceed to **step 5** and to the search of a fixed point in the space of controls. The only difference at that point is in the ansatz for $t \mapsto Y_t$. Since we consider the case of common noise, we now search for an ansatz of the form $Y_t = \pi_t(X_t - \bar{X}_t) + \hat{\pi}_t \bar{X}_t + \eta_t$, where $(\pi, \hat{\pi}, \eta) \in L^\infty([0, T], \mathbb{R}^{nd \times d}) \times L^\infty([0, T], \mathbb{R}^{nd \times d}) \times S_{\mathbb{F}}^2(\Omega \times [0, T], \mathbb{R}^{nd})$ satisfy:

$$\begin{cases} d\eta_t &= \psi_t dt + \phi_t dW_t + \phi_t^0 dW_t^0 \\ \eta_T &= r = (r^i)_{i \in \llbracket 1, n \rrbracket} \\ d\pi_t &= \dot{\pi}_t dt + Z_t^{0, \pi^i} dW_t^0 \\ \pi_T &= 0 \\ d\hat{\pi}_t &= \hat{\pi}_t dt + Z_t^{0, \hat{\pi}^i} dW_t^0 \\ \hat{\pi}_T &= 0. \end{cases} \tag{30}$$

The method to determine the coefficients is then similar. Existence and uniqueness of a solution (K^i, Λ^i) to the backward stochastic Riccati equation in (29) is discussed in [13], section 3.2. The existence of a solution (Y^i, Z^i, Z^{0, Y^i}) to the linear mean-field BSDE in (29) is obtained as in **step 6** thanks to Thm. 2.1 in [12]. As in the previous section the existence of a solution $(\pi, \hat{\pi}) \in L_{\mathbb{F}^0}^\infty(\Omega \times \mathbb{R}_+, \mathbb{R}^{nd \times d})$ (essentially bounded \mathbb{F}^0 -adapted functions) of (30) in the general case is a conjecture and needs to be verified in each example. We are not aware of any work tackling the existence of solutions in such situation. Given $(\pi, \hat{\pi})$ the existence of (η, ϕ, ϕ^0) solution to (30) is ensured as in the previous section by Thm. 2.1 in [12].

3.3 The case of multiple Brownian motions

We quickly sketch an extension to the case where there are multiple Brownian motions driving the state equation. The assumptions on the coefficients are the same as in the previous part. Only the length of the calculus changes. Let us now consider the state dynamic:

$$dX_t = b(t, X_t^\alpha, \mathbb{E}[X_t^\alpha], \alpha_t, \mathbb{E}[\alpha_t])dt + \sum_{\ell=1}^\kappa \sigma^\ell(t, X_t^\alpha, \mathbb{E}[X_t^\alpha], \alpha_t, \mathbb{E}[\alpha_t])dW_t^\ell$$

where $\Phi^{i0}, \Psi^{i0}, \Delta^{i0}, \Gamma_t^{i0}$ are defined in (8).

$$\begin{cases} \alpha_t &= (\alpha_{1,t}, \dots, \alpha_{n,t}) \\ b(t, x, \bar{x}, a, \bar{a}) &= \beta_t + b_{x,t}x + \tilde{b}_{x,t}\bar{x} + \sum_{i=1}^n b_{i,t}a_{i,t} + \tilde{b}_{i,t}\bar{a}_{i,t} \\ \sigma^\ell(t, x, \bar{x}, a, \bar{a}) &= \gamma_t^\ell + \sigma_{x,t}^\ell x + \tilde{\sigma}_{x,t}^\ell \bar{x} + \sum_{i=1}^n \sigma_{i,t}^\ell a_{i,t} + \tilde{\sigma}_{i,t}^\ell \bar{a}_{i,t}. \end{cases} \tag{31}$$

We require the coefficients in (31) to satisfy an adaptation of **(H1)** where $\gamma, \sigma_x, \tilde{\sigma}_x, (\sigma_i, \tilde{\sigma}_i)_{i \in \llbracket 1, n \rrbracket}$ are replaced by $\gamma^\ell, \sigma_x^\ell, \tilde{\sigma}_x^\ell, (\sigma_i^\ell, \tilde{\sigma}_i^\ell)_{i \in \llbracket 1, n \rrbracket}$ for $\ell \in \{1, \dots, \kappa\}$.

To take into account the multiple Brownian motions in **step 1**, we now search for random fields of the form $w_t^i(x, \bar{x}) = K_t^i \cdot (x - \bar{x})^{\otimes 2} + \Lambda_t^i \cdot \bar{x}^{\otimes 2} + 2Y_t^{i\top} x + R_t^i$ with the processes $(K^i, \Lambda^i, Y^i, (Z^{\ell, i})_{\ell \in \llbracket 1, \kappa \rrbracket}, R^i)$ in $(L^\infty([0, T], \mathbb{S}_+^d))^2 \times \mathcal{S}_{\mathbb{F}^0}^2(\Omega \times [0, T], \mathbb{R}^d) \times (L_{\mathbb{F}^0}^2(\Omega \times [0, T], \mathbb{R}^d))^\kappa \times L^\infty([0, T], \mathbb{R})$ and solution to:

$$\begin{cases} dK_t^i = \dot{K}_t^i dt & K_T^i = P^i \\ d\Lambda_t^i = \dot{\Lambda}_t^i dt & \Lambda_T^i = P^i + \tilde{P}^i \\ dY_t^i = \dot{Y}_t^i dt + \sum_{\ell} Z_t^{i,\ell} dW_t^\ell & Y_T^i = r^i \\ dR_t^i = \dot{R}_t^i dt & R_T^i = 0 \end{cases}$$

where $(\dot{K}^i, \dot{\Lambda}^i, \dot{R}^i)$ are deterministic processes valued in $\mathbb{S}_+^d \times \mathbb{S}_+^d \times \mathbb{R}$ and (\dot{Y}^i, Z^i) are adapted processes valued in \mathbb{R}^d .

The method then follows the same steps with generalized coefficients and at **step 2** we obtain generalized coefficient for (5) and (6):

$$\begin{cases} \Phi_t^i = Q_t^i + K_t^i b_{x,t} + b_{x,t}^\top K_t^i + \sum_r \sigma_{x,t}^{r\top} K_t^i \sigma_{x,t}^r - \rho K_t^i \\ \Psi_t^i = \hat{Q}_t^i + \Lambda_t^i \hat{b}_{x,t} + \hat{b}_{x,t}^\top \Lambda_t^i + \sum_r \hat{\sigma}_{x,t}^{r\top} \Lambda_t^i \hat{\sigma}_{x,t}^r - \rho \Lambda_t^i \\ \Delta_t^i = L_{x,t}^i + b_{x,t}^\top Y_t^i + \bar{b}_{x,t}^\top \bar{Y}_t^i + \Lambda_t^i \bar{\beta}_t + K_t^i (\beta_t - \bar{\beta}_t) \\ \quad + \sum_{\ell} \sigma_{x,t}^{\ell\top} K_t^i \gamma_t^\ell + \tilde{\sigma}_{x,t}^{\ell\top} K_t^i \bar{\gamma}_t^\ell + \sigma_{x,t}^{\ell\top} Z_t^{i,\ell} + \tilde{\sigma}_{x,t}^{\ell\top} \bar{Z}_t^{i,\ell} - \rho Y_t^i \\ \quad + \sum_{k \neq i} U_{k,t}^i (\alpha_{k,t} - \bar{\alpha}_{k,t}) + V_{k,t}^i \bar{\alpha}_{k,t} \\ \Gamma_t^i = 2\beta_t^\top Y_t^i + \sum_{\ell} \gamma_t^{\ell\top} K_t^i \gamma_t^\ell + 2\gamma_t^{\ell\top} Z_t^{i,\ell} \\ \quad + \sum_{k \neq i} (\alpha_{k,t} - \bar{\alpha}_{k,t})^\top S_{k,t}^i (\alpha_{k,t} - \bar{\alpha}_{k,t}) + \bar{\alpha}_{k,t}^\top \hat{S}_{k,t}^i \bar{\alpha}_{k,t} \\ \quad + 2[O_{k,t}^i + \xi_{k,t}^i - \bar{\xi}_{k,t}^i]^\top \alpha_{k,t} \end{cases} \quad (32)$$

$$\begin{cases} S_{k,t}^i = N_{k,t}^i + \sum_{\ell} \sigma_{k,t}^{\ell\top} K_t^i \sigma_{k,t}^\ell \\ \hat{S}_{k,t}^i = \hat{N}_{k,t}^i + \sum_{\ell} \hat{\sigma}_{k,t}^{\ell\top} K_t^i \hat{\sigma}_{k,t}^\ell \\ U_{k,t}^i = I_{k,t}^i + b_{k,t}^\top K_t^i + \sum_{\ell} \sigma_{k,t}^{\ell\top} K_t^i \sigma_{x,t}^\ell \\ V_{k,t}^i = \hat{I}_{k,t}^i + \hat{b}_{k,t}^\top \Lambda_t^i + \sum_{\ell} \hat{\sigma}_{k,t}^{\ell\top} K_t^i \hat{\sigma}_{x,t}^\ell \\ O_{k,t}^i = \bar{L}_{k,t}^i + \bar{b}_{k,t}^\top \bar{Y}_t^i + \sum_{\ell} \bar{\sigma}_{k,t}^{\ell\top} \bar{Z}_t^{i,\ell} + \hat{\sigma}_{k,t}^{\ell\top} K_t^i \bar{\gamma}_t^\ell \\ J_{k,l,t}^i = G_{k,l,t}^i + \sum_{\ell} \sigma_{k,t}^{\ell\top} K_t^i \sigma_{l,t}^\ell \\ \hat{J}_{k,l,t}^i = \hat{G}_{k,l,t}^i + \sum_{\ell} \hat{\sigma}_{k,t}^{\ell\top} K_t^i \hat{\sigma}_{l,t}^\ell \\ \xi_{k,t}^i = L_{k,t}^i + b_{k,t}^\top Y_t^i + \sum_{\ell} \sigma_{k,t}^{\ell\top} Z_t^{i,\ell} + \sigma_{k,t}^{\ell\top} K_t^i \gamma_t^\ell. \end{cases} \quad (33)$$

From these extended formulas we can then constrain the coefficients as in **step 3** and obtain (8) with now the generalized coefficients defined in (32) and (33). The **step 4** is then straightforward and we obtain the best response functions:

$$\begin{aligned} \alpha_{i,t} &= a_t^{i,0}(X_t, \mathbb{E}[X_t]) + a_t^{i,1}(\mathbb{E}[X_t]) \\ &= -(S_{i,t}^i)^{-1} U_{i,t}^i (X_t - \mathbb{E}[X_t]) - (S_{i,t}^i)^{-1} (\xi_{i,t}^i - \bar{\xi}_{i,t}^i) - (\hat{S}_{i,t}^i)^{-1} (V_{i,t}^i \mathbb{E}[X_t] + O_{i,t}^i). \end{aligned}$$

From **step 4** we can then continue to **step 5**, i.e. the fixed point search. The only difference at that point is in the ansatz for $t \mapsto Y_t$. Since we consider the case with multiple Brownian motions we now search for an ansatz of the form $Y_t = \pi_t(X_t - \bar{X}_t) + \hat{\pi}_t \bar{X}_t + \eta_t$ where $(\pi, \hat{\pi}, \eta) \in L^\infty([0, T], \mathbb{R}^{nd \times d}) \times L^\infty([0, T], \mathbb{R}^{nd \times d}) \times \mathbb{S}_{\mathbb{F}}^2(\Omega \times [0, T], \mathbb{R}^{nd})$ satisfy:

$$\begin{cases} d\eta_t &= \psi_t dt + \sum_{\ell} \phi_t^{\ell} dW_t^{\ell} \\ \eta_T &= r = (r^i)_{i \in \llbracket 1, n \rrbracket} \\ d\pi_t &= \dot{\pi}_t dt \\ \pi_T &= 0 \\ d\hat{\pi}_t &= \dot{\hat{\pi}}_t dt \\ \hat{\pi}_T &= 0. \end{cases}$$

The method to determine the coefficients $\pi, \hat{\pi}, \eta$ is then similar. The validity of the computations i.e. **Step 6** can be done exactly as in the case of a single brownian motion.

4 Example

We now focus on a toy example to illustrate the previous results. Let us consider a two player game where the state dynamics is simply a Brownian motion that two players can control. The goal of each player is to get the state near its own target $t \mapsto T_t^i$, where $t \mapsto T_t^i, i = 1, 2$, is a stochastic process. In order to add mean-field terms we suppose that each player try also to minimize the variance of the state and the variance of their controls.

$$\begin{cases} dX_t &= (b_1 \alpha_{1,t} + b_2 \alpha_{2,t}) dt + \sigma dW_t \\ J^i(\alpha_1, \alpha_2) &= \mathbb{E} \left[\int_0^{\infty} e^{-\rho u} \left(\lambda^i \text{Var}(X_u) + \delta^i (X_u - T_u^i)^2 + \theta^i \text{Var}(\alpha_{i,u}) + \xi^i \alpha_{i,u}^2 \right) du \right]. \end{cases}$$

where $(\lambda^i, \delta^i, \theta^i, \xi^i) \in \mathbb{R}_+^4$. In order to fit to the context described in the first section we rewrite the cost function as follows:

$$\begin{aligned} J^i(\alpha_1, \alpha_2) &= \mathbb{E} \left[\int_0^{\infty} e^{-\rho u} \left((\lambda^i + \delta^i) (X_u - \bar{X}_u)^2 + \delta^i \bar{X}_u^2 + (\theta^i + \xi)^2 (\alpha_{i,u} - \bar{\alpha}_{i,u}) \right. \right. \\ &\quad \left. \left. + \xi^i \bar{\alpha}_{i,u}^2 + 2X_u [-2\delta^i T^i] + \delta^i (T_u^i)^2 \right) du \right]. \end{aligned}$$

Since the terms $\delta^i (T^i)^2$ do not influence the optimal control of the players, we work with the slightly simplified cost function:

$$\begin{aligned} \tilde{J}^i(\alpha_1, \alpha_2) &= \mathbb{E} \left[\int_0^{\infty} e^{-\rho u} \left((\lambda^i + \delta^i) (X_u - \bar{X}_u)^2 + \delta^i \bar{X}_u^2 + (\theta^i + \xi)^2 (\alpha_{i,u} - \bar{\alpha}_{i,u}) \right. \right. \\ &\quad \left. \left. + \xi^i \bar{\alpha}_{i,u}^2 + 2X_u [-2\delta^i T^i] \right) du \right]. \end{aligned}$$

Following the method explained in the previous section, we use Theorem 2 in order to find a Nash equilibrium. We obtain the feedback form of the open loop controls and the dynamics of the state:

$$\begin{cases} \alpha^i - \bar{\alpha}^i &= -\frac{P^i}{b_i} ((K^i + \pi^i)(X_t - \bar{X}_t) + \eta_t^i - \bar{\eta}_t^i) \\ \bar{\alpha}^i &= -\frac{\tilde{P}^i}{b_i} ((\Lambda^i + \tilde{\pi}^i)\bar{X}_t + \bar{\eta}_t^i) \\ \bar{X}_t &= \bar{X}_0 e^{-\tilde{a}t} + \int_0^t e^{-\tilde{a}(t-u)} \bar{\gamma}_u du \\ X_t - \bar{X}_t &= (X_0 - \bar{X}_0) e^{-at} + \int_0^t e^{-a(t-u)} [(\gamma_u - \bar{\gamma}_u) du + \sigma dW_u] \end{cases} \tag{34}$$

where $K^i \in \mathbb{R}_+, \Lambda^i \in \mathbb{R}_+, a \in \mathbb{R}, \tilde{a} \in \mathbb{R}, \pi \in \mathbb{R}^2, \tilde{\pi} \in \mathbb{R}^2, \eta \in L^2_{\mathbb{F}}((0, \infty), \mathbb{R}^2), \gamma \in L^2_{\mathbb{F}}((0, \infty), \mathbb{R}^2)$ satisfy:

$$\begin{cases} K^i &= \frac{-\rho + \sqrt{\rho^2 + 4P^i(\lambda^i + \delta^i)}}{2P^i} \\ \Lambda^i &= \frac{-\rho + \sqrt{\rho^2 + 4\tilde{P}^i\delta^i}}{2\tilde{P}^i} \\ a &= \sum_{i=1}^2 P^i (K^i + \pi^i) \\ \tilde{a} &= \sum_{i=1}^2 \tilde{P}^i (\Lambda^i + \tilde{\pi}^i) \\ 0 &= P_y \pi - (\pi B - P_\alpha)(S_x + S_y \pi) \\ 0 &= \tilde{P}_y \tilde{\pi} - (\tilde{\pi} B - \tilde{P}_\alpha)(\tilde{S}_x + \tilde{S}_y \tilde{\pi}) \\ \eta_t - \bar{\eta}_t &= -\int_t^\infty e^{[P_y - (\pi B - P_\alpha)S_y](t-u)} \\ &\quad \times \mathbb{E} [H_u - \bar{H}_u | \mathcal{F}_t] du \\ \bar{\eta}_t &= -\int_t^\infty e^{[\tilde{P}_y - (\tilde{\pi} B - \tilde{P}_\alpha)\tilde{S}_y](t-u)} \bar{H}_u du \\ \gamma_t - \bar{\gamma}_t &= -\sum_{i=1}^2 P^i (\eta_{i,t} - \bar{\eta}_{i,t}) \\ \bar{\gamma}_t &= -\sum_{i=1}^2 \tilde{P}^i \bar{\eta}_{i,t} \end{cases} \begin{cases} S_x &= -(P^1 K^1 / b_1, P^2 K^2 / b_2) \\ \tilde{S}_x &= -(\tilde{P}^1 \Lambda^1 / b_1, \tilde{P}^2 \Lambda^2 / b_2) \\ S_y &= -(\mathbf{1}_{i=j} P^i / b_i)_{i,j \in \{1,2\}} \\ \tilde{S}_y &= -(\mathbf{1}_{i=j} \tilde{P}^i / b_i)_{i,j \in \{1,2\}} \\ P_\alpha &= -(\mathbf{1}_{i \neq j} K^i b_j)_{i,j \in \{1,2\}} \\ \tilde{P}_\alpha &= -(\mathbf{1}_{i \neq j} \Lambda^i b_j)_{i,j \in \{1,2\}} \\ P_y &= (\mathbf{1}_{i=j} (P^i K^i + \rho))_{i,j \in \{1,2\}} \\ \tilde{P}_y &= (\mathbf{1}_{i=j} (\tilde{P}^i \Lambda^i + \rho))_{i,j \in \{1,2\}} \\ H &= (\delta^1 T^1, \delta^2 T^2) \\ B &= (b_1, b_2) \\ P^i &= \frac{b_i^2}{\theta_i + \xi^i} \\ \tilde{P}^i &= \frac{b_i^2}{\xi^i}. \end{cases} \tag{35}$$

From (34) and (35) we can study and simulate the influence of the different parameters of the cost-function of the first player. We notice that (λ^1, θ^1) only influence $X - \mathbb{E}[X]$ and the feedback form of $\alpha_1 - \bar{\alpha}_1$ (zero-mean terms).

- If $\lambda^1 \rightarrow \infty$ then $\pi_1 \sim \lambda^1$ and $K^1 \rightarrow \infty$ which implies that $X_t - \mathbb{E}[X]_t \rightarrow 0$ for all $t \geq 0$. This is expected since the term λ^1 penalizes the variance of the state $\text{Var}(X_t)$ in the cost function of the first player. See Figure 1.
- If $\delta^1 \rightarrow \infty$ then $(\pi, \tilde{\pi}) \rightarrow \infty$ and $K^1 \sim \delta^1$ and $\Lambda^1 \sim \delta^1$ which imply that $X_t \rightarrow T_t^1$ for all $t \geq 0$. This is also expected since the term δ^1 penalizes the quadratic gap between the state X and the target T^1 . See Figure 2.
- If $\theta^1 \rightarrow \infty$ then $P^1 \rightarrow 0, P^1 K^1 \rightarrow 0, P^1 \pi^1 \rightarrow 0$ and $P^1 \eta_t^1 \rightarrow 0$ for every $t \geq 0$. We then have $\alpha_t - \bar{\alpha}_t \rightarrow 0$ for all $t \geq 0$ and all the terms relative to the first player in $X - \mathbb{E}[X]$ disappear. Given that θ^1 penalizes the variance of the control of the first player, this convergence is also intuitive.
- If $\xi^1 \rightarrow \infty$ then $(P^1, \tilde{P}^1) \rightarrow 0$ which imply that $(\alpha_{1,t}, \bar{\alpha}_{1,t}) \rightarrow 0$ for all $t \geq 0$ and all the terms relative to the first player in $X_t - \mathbb{E}[X_t]$ and $\mathbb{E}[X_t]$ disappear for all $t \geq 0$. This means that the first player becomes powerless.

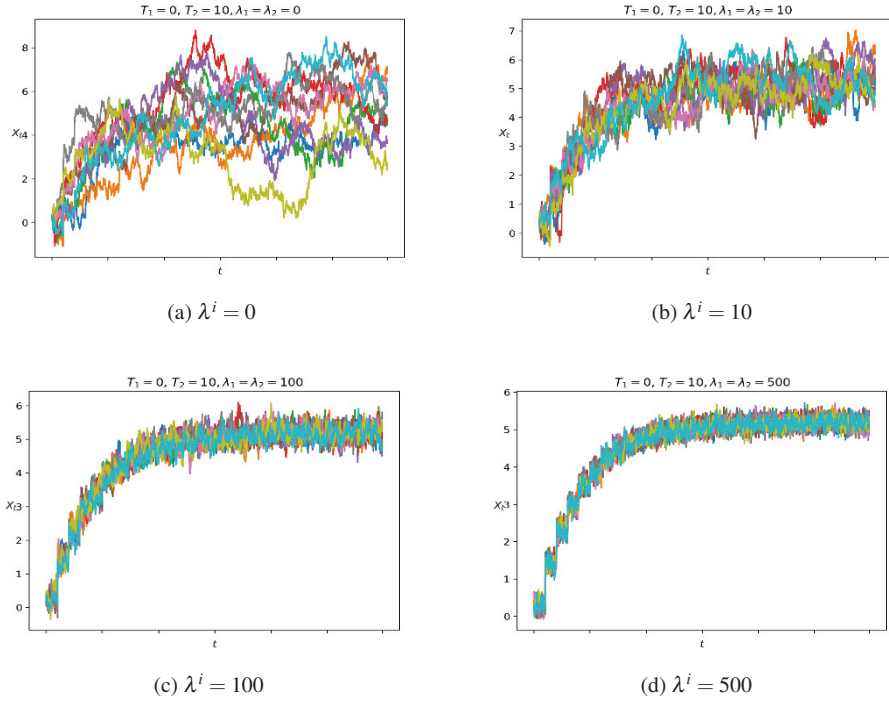


Fig. 1: Nash equilibrium with:
 $b_i = \sigma = \delta^i = \theta^i = \xi^i = 1, \rho = 3, T^1 = 0, T^2 = 10$

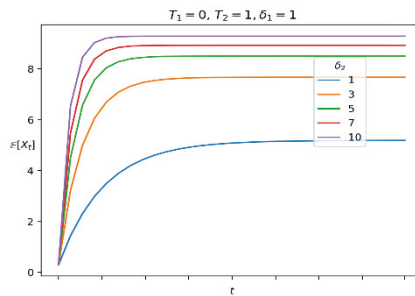


Fig. 2: $t \mapsto \mathbb{E}[X_i]$
 $b_i = \sigma = \delta^i = \theta^i = \xi^i = 1, \rho = 3, T^1 = 0, T^2 = 10$

References

1. A. Aurell and B. Djehiche. Mean-field type modeling of nonlocal crowd aversion in mean-field modeling of nonlocal crowd aversion in pedestrian crowd dynamics. *SIAM Journal on*

- Control and Optimization*, 56(1):434–455, 2018.
2. M. Basei and H. Pham. A weak martingale approach to linear-quadratic mckean-vlasov stochastic control problems. *Journal of Optimization Theory and Applications*, to appear, 2018.
 3. A. Bensoussan, J. Frehse, and P. Yam. *Mean field games and mean field type control theory*. Springer Briefs in Mathematics, 2013.
 4. A. Bensoussan, T. Huang, and M. Laurière. Mean field control and mean field game models with several populations. arXiv: 1810.00783.
 5. R. Carmona and F. Delarue. Forward-backward stochastic differential equations and controlled mckean-vlasov dynamics. *Annals of Probability*, 43(5):2647–2700, 2015.
 6. R. Carmona and F. Delarue. *Probabilistic Theory of Mean Field Games with Applications vol I and II*. Springer, 2018.
 7. A. Cosso and H. Pham. Zero-sum stochastic differential games of generalized mckean-vlasov type. *Journal de Mathématiques Pures et Appliquées*, to appear, 2018.
 8. B. Djehiche, J. Barreiro-Gomez, and H. Tembine. Electricity price dynamics in the smart grid: A mean-field-type game perspective. In *23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS2018)*, pages 631–636, 2018.
 9. T. Duncan and H. Tembine. Linear-quadratic mean-field-type games: a direct method. *Games*, 9(7), 2018.
 10. P.J. Graber. Linear-quadratic mean-field type control and mean-field games with common noise, with application to production of an exhaustible resource. *Applied Mathematics and Optimization*, 74(3):459–486, 2016.
 11. J. Huang, X. Li, and J. Yong. Linear-quadratic optimal control problem for mean-field stochastic differential equations in infinite horizon. *Mathematical Control and Related Fields*, 5(1):97–139, 2015.
 12. X. Li, J. Sun, and J. Xiong. Linear quadratic optimal control problems for mean-field backward stochastic differential equations. *Applied Mathematics & Optimization*, to appear, 2017.
 13. H. Pham. Linear quadratic optimal control of conditional mckean-vlasov equation with random coefficients and applications. *Probability, Uncertainty and Quantitative Risk*, 1:7, 2016.
 14. H. Pham and X. Wei. Dynamic programming for optimal control of stochastic McKean-Vlasov dynamics. *SIAM Journal on Control and Optimization*, 55(2):1069–1101, 2017.
 15. J. Yong. Linear-quadratic optimal control problem for mean-field stochastic differential equations in infinite horizon. *SIAM Journal on Control and Optimization*, 51(4):2809–2838, 2013.
 16. J. Yong and X.Y. Zhou. *Stochastic controls: Hamiltonian Systems and HJB Equations*. SMAP. Springer, 1999.



Stochastic Multigroup Epidemic Models: Duration and Final Size

Aadrita Nandi and Linda J.S. Allen

Abstract The epidemic duration, the final epidemic size, and the probability of an outbreak are studied in stochastic multigroup epidemic models. Two models are considered, where the transmission rate for each group either depends on the infectious individuals or on the susceptible individuals, referred to as Model 1 and Model 2, respectively. Such models are applicable to emerging and re-emerging infectious diseases. Applying a multitype branching process approximation, it is shown for Model 1 that an outbreak is dependent primarily on group reproduction numbers, whereas for Model 2, this dependence is due to group recovery rates. The probability distributions for epidemic duration and for final size are a mixture of two distributions, that depend on whether an outbreak occurs. Given there is an outbreak, it is shown that the mean final size of the stochastic multigroup model agrees well with the final size obtained from the underlying deterministic model. These methods can be extended to more general stochastic multigroup models and to other stochastic epidemic models with multiple stages, patches, hosts, or pathogens.

1 Introduction

The Susceptible-Infectious-Recovered (SIR) epidemic model was introduced in 1927 by Kermack and McKendrick [21]. It has been applied to many infectious diseases and has served as a framework for development of more realistic epidemic models. The basic reproduction number and final size of an epidemic are well-known concepts that have been studied extensively in the SIR model and have been extended to more complex settings, such as epidemic models with multiple stages,

A. Nandi

Texas Tech University, Lubbock, TX 79409-1042, e-mail: aadrita.nandi@ttu.edu

L. Allen

Texas Tech University, Lubbock, TX 79409-1042, e-mail: linda.j.allen@ttu.edu

patches, hosts, or pathogens, e.g., [3, 9, 12, 19, 22, 23, 31, 32, 36] and references therein.

For the stochastic, continuous-time Markov chain (CTMC) SIR epidemic model, the threshold parameters for disease outbreak, final epidemic size, and epidemic duration have also been studied (e.g., [5, 7, 35, 38]). Whittle in 1955 was the first to introduce the concept of a minor or a major epidemic in the stochastic SIR model [38]. The term “minor” implies that only a few individuals become infected, whereas “major” implies a relatively large number of individuals are infected (an outbreak) [38]. Unlike the deterministic model, the initial number of infectious individuals is important. The probability of a minor epidemic is given by

$$\mathbb{P}_{minor} = \begin{cases} \left(\frac{1}{\mathcal{R}_0}\right)^i, & \mathcal{R}_0 > 1, \\ 1, & \mathcal{R}_0 < 1, \end{cases}$$

where $\mathbb{P}_{major} = 1 - \mathbb{P}_{minor}$ and i is the initial number of infected individuals. Analytic formulas for the duration and final size distributions have also been computed for the stochastic SIR epidemic model [5, 7, 11, 35]. In more complex stochastic epidemic settings, some computational and analytical methods have been applied to obtain threshold parameters and approximations for the distributions for epidemic duration and final size, e.g., [2, 6, 8, 11, 20, 25].

In this investigation, we consider two stochastic multigroup epidemic models, where the group transmission rates are determined by either (1) the infectious individuals within the group, such as superspreaders or nonsuperspreaders or (2) the susceptible individuals within the group that are defined by vaccination coverage or immunity levels. We refer to these models as Model 1 and Model 2, respectively. In [27], we investigated these two models for two groups but focused our study on the probability of a minor versus a major epidemic. Edholm et al. [14] applied a more complex two-group model to the study of superspreaders in outbreaks of Ebola and Middle East Respiratory Syndrome (MERS). We extend this investigation to n groups and as in [27], we compare the probability of a minor to a major epidemic when the epidemic is initiated by an infected individual in a particular group. It is verified for Model 1 that the group with the largest basic reproduction number, the superspreaders, has the greatest probability of initiating an epidemic while in Model 2, it is the group with the lowest recovery rate, i.e., lowest immunity. In addition, we investigate the epidemic duration and the final epidemic size. The epidemic duration and final size can be separated into bimodal probability distributions, a mixed distribution, where each mode corresponds to a minor or a major epidemic. For $\mathcal{R}_0 > 1$, the duration and final size of minor and major epidemics are compared to several well-known distributions, e.g. normal, gamma, lognormal, and Weibull. For comparison purposes, the two models have the same recovery rates, the same group sizes, and the same group reproduction numbers.

In the next section, the multigroup CTMC SIR models are defined. A multitype branching process approximation is used to compute the probabilities of a minor and a major epidemic in Section 3. In Sections 4 and 5, numerical examples with three

groups illustrate the probability distributions for duration and final size and the best fit of these distributions to some well-known distributions. In the concluding Section 6, the mathematical results and their biological implications are summarized.

2 Stochastic Multigroup SIR Model

The stochastic formulation is based on the well-known deterministic ordinary differential equation (ODE) multigroup model,

$$\begin{aligned} \frac{dS_k}{dt} &= -\frac{S_k}{N} \sum_{j=1}^n \beta_{kj} I_j, \\ \frac{dI_k}{dt} &= \frac{S_k}{N} \sum_{j=1}^n \beta_{kj} I_j - \gamma_k I_k, \\ \frac{dR_k}{dt} &= \gamma_k I_k, \end{aligned} \tag{1}$$

where S_k , I_k , and R_k denote susceptible, infectious, and recovered individuals from group k , respectively. All recovered individuals can be put into one group R , since there is no re-infection within this group. The transmission rate from an infectious individual from group j to a susceptible individual in group k is β_{kj} . The recovery rate of an infectious individual in group k is γ_k . Group k has a constant population size N_k and therefore, the total population size is constant, $N = \sum_{k=1}^n N_k$.

We consider two models, where either the transmission rate β_{kj} depends only on the infectious group j or the susceptible group k . These two cases are referred to as Model 1 and Model 2, respectively,

$$\begin{aligned} \text{Model 1 : } \beta_{kj} &= \beta_j^I \\ \text{Model 2 : } \beta_{kj} &= \beta_k^S. \end{aligned} \tag{2}$$

In either case, the basic reproduction number is

$$\mathcal{R}_0 = \sum_{i=1}^n \frac{\beta_i^C}{\gamma_i} \frac{N_i}{N} = \sum_{i=1}^n \mathcal{R}_{0i} \frac{N_i}{N}, \tag{3}$$

where $C = I$ or $C = S$ identifies the dependence on either the infectious or susceptible groups. The parameter \mathcal{R}_{0i} is the i th group reproduction number. The definition of \mathcal{R}_0 follows directly from the next generation matrix approach [37].

The assumptions regarding the transmission rates have applications to recent emerging and re-emerging infectious diseases. For example, the assumption that the transmission rate depends on the infectious individuals, β_j^I in Model 1, implies that after infection, infectious individuals in group j determine transmission, not the susceptible individuals in that group. Two examples of emerging diseases where

this assumption is meaningful are Severe Acute Respiratory Syndrome (SARS) and MERS [40]. In 2003 and 2015, these two emerging diseases were caused by super-spreaders, infectious individuals that infect a large proportion of individuals before being identified [40]. The behavior of these infectious individuals and the particular environment (e.g., hospital settings) contributed to their spread. The assumption in Model 2 is that the transmission rate depends on the susceptible individuals in group k , β_k^I , and not the infectious individuals in that group. Re-emerging diseases such as measles, mumps, and tuberculosis have a long history. They have been controlled via vaccination or treatment programs. However, due to vaccine waning, lack of a second vaccine dose, introduction of new strains with little cross protection, or development of antibiotic resistance, re-emergence of these diseases has occurred [28]. Susceptible individuals in Model 2 are divided into groups according to their susceptibility to infection or re-infection via the transmission rate β_k^S . Re-infection rates and recovery rates depend on each group, but after infection, their transmission rate is the same as other groups. The duration is longer for individuals with low levels of immunity. Therefore, the multigroup models differentiate between susceptible and infectious individuals through behavior after infection or by immunity levels which affect whether individuals become infected within a group.

For the CTMC models, the infinitesimal transition rates corresponding to Model 1 and Model 2 are summarized in Table 1. Each of the discrete random variables satisfy $S_j, I_j, R_j \in \{0, 1, \dots, N_j\}$ with the restriction $S_j + I_j + R_j = N_j$ for $j = 1, \dots, n$. The interevent time is exponentially distributed with parameter λ_C , $C = I, S$:

$$\lambda_I = \sum_{j=1}^n \left[\gamma_j i_j + \frac{s_j}{N} \sum_{k=1}^n \beta_k^I i_k \right],$$

$$\lambda_S = \sum_{j=1}^n \left[\gamma_j i_j + \beta_j^S \frac{s_j}{N} \sum_{k=1}^n i_k \right],$$

where s_j and i_j are the values of the random variables S_j and I_j , respectively.

Table 1 Infinitesimal transition probabilities for Models 1 and 2

Model 1		
Event	Change	Probability
$S_j \rightarrow I_j$	$(\Delta S_j, \Delta I_j) = (-1, 1)$	$\frac{s_j}{N} \sum_{k=1}^n \beta_k^I i_k \Delta t + o(\Delta t)$
$I_j \rightarrow R$	$(\Delta I_j, \Delta R_j) = (-1, 1)$	$\gamma_j i_j \Delta t + o(\Delta t)$
Model 2		
Event	Change	Probability
$S_j \rightarrow I_j$	$(\Delta S_j, \Delta I_j) = (-1, 1)$	$\beta_j^S \frac{s_j}{N} \sum_{k=1}^n i_k \Delta t + o(\Delta t)$
$I_j \rightarrow R$	$(\Delta I_j, \Delta R) = (-1, 1)$	$\gamma_j i_j \Delta t + o(\Delta t)$

3 Probability of Minor and Major Epidemics

Multitype branching process (MBP) theory is used to approximate the dynamics of the CTMC models near the DFE, $S_j = N_j, i = 1, \dots, n$. In the MBP only the n random variables, $I_j, j = 1, \dots, n$, are considered. For this linear approximation near the DFE, the random variables I_j eventually equal zero or approach infinity. The group size is accounted for in the terms N_j/N . The approximation is valid for small initial values and sufficiently large population sizes, as we also make the assumption that the random variables I_j are independent.

Differential equations for the probability of disease extinction can be derived in terms of probability generating functions (pgfs) for each infectious group by applying the backward Kolmogorov differential equations and from the assumption of independence of the random variables [1, 2, 4, 13, 18, 27, 29]. For each group j , assume $I_j(0) = 1$ and $I_k(0) = 0$ for $k \neq j$, i.e., $I(0) = (I_1(0), \dots, I_n(0)) = e_j$, where e_j is the j th unit vector. The differential equations for disease extinction are

$$\frac{dp_j}{dt} = \omega_j[f_j(p_1, \dots, p_n) - p_j], \quad j = 1, \dots, n, \tag{4}$$

where

$$p_j(t) = p_{(e_j, \mathbf{0})}(t) = \mathbb{P}(I(t) = \mathbf{0} | I(0) = e_j),$$

the notation $\mathbf{0}$ means the zero vector, and $f_j(p_1, \dots, p_n)$ is the pgf for group j . The parameter ω_j is the waiting time parameter in the j th group. The asymptotic probability of extinction is found by computing the stationary solution of the differential equations, that is, the minimal fixed point of the pgfs, $f_j(q) = q_j$ and $q = (q_1, \dots, q_n) \in (0, 1]^n$ [2, 4, 13, 18, 29]. In general, the independent assumption of the variables I_j implies for small initial values $I(0) = (i_1, \dots, i_n)$, the asymptotic probability of extinction is approximately

$$\mathbb{P}_{minor} = \prod_{j=1}^n q_j^{i_j} \tag{5}$$

and $\mathbb{P}_{major} = 1 - \mathbb{P}_{minor}$. When $\mathcal{R}_0 \leq 1$, the only fixed point in $(0, 1]^n$ is $q = (1, \dots, 1)$, so that the $\mathbb{P}_{minor} = 1$. The following properties of the pgfs ensure that \mathcal{R}_0 determines the asymptotic probability of disease extinction,

$$\lim_{t \rightarrow \infty} p_{(e_j, \mathbf{0})}(t) = q_j,$$

- (i) each pgf $f_j(u_1, \dots, u_n)$ is nonlinear,
- (ii) $f_j(0, \dots, 0) > 0$ and
- (iii) the Jacobian matrix of the system (4) when evaluated at the fixed point $(1, \dots, 1)$ is irreducible.

The Jacobian matrix in (iii) is often referred to as the expectation matrix M . See e.g., [2, 4, 13, 18, 29].

For Model 1, the waiting time parameter is $\omega_j = \gamma_j + \beta_j^I$ and the pgfs f_j are given by

$$f_j(u_1, u_2, \dots, u_n) = \frac{\gamma_j + \beta_j^I u_j \sum_{k=1}^n \frac{N_k}{N} u_k}{\gamma_j + \beta_j^I}, \tag{6}$$

for $u_i \in [0, 1]$, $i, j = 1, \dots, n$ [27]. The pgfs in (6) satisfy properties (i)-(iii). The following result for Model 1 shows that the relation between the probabilities of extinction for different groups depend on their group reproduction numbers $\mathcal{R}_{0j} = \beta_j^I / \gamma_j$. If group i has a smaller reproduction number than group j , then group i has a larger probability of extinction. The following theorem simplifies the proof in [27] for two groups and extends it to n groups.

Theorem 1. *Assume in the MBP approximation for Model 1 that the group basic reproduction numbers satisfy $\beta_i^I / \gamma_i < \beta_j^I / \gamma_j$ for some i, j . Then the probability of extinction for groups i and j satisfy one of the following:*

- (a) *If $\mathcal{R}_0 > 1$, then $0 < q_j < q_i < 1$ and all other extinction probabilities are less than one.*
- (b) *If $\mathcal{R}_0 \leq 1$, then $q_i = 1$ for all $i = 1, \dots, n$.*

In the special case, $\beta_i^I / \gamma_i = \beta / \gamma$ for all $i = 1, \dots, n$, and if $\mathcal{R}_0 = \beta / \gamma > 1$, then $q_i = 1 / \mathcal{R}_0$ and if $\mathcal{R}_0 \leq 1$, then $q_i = 1$ for all $i = 1, \dots, n$.

Proof. Assume $\beta_i^I / \gamma_i < \beta_j^I / \gamma_j$. Rearranging the expression $f_j(u_1, \dots, u_n) = u_j$ leads to

$$\sum_{k=1}^n \frac{N_k}{N} u_k - 1 = \frac{\gamma_j}{\beta_j^I} \left(1 - \frac{1}{u_j} \right). \tag{7}$$

A similar identity holds for $f_i(u_1, \dots, u_n) = u_i$. Hence, we can equate the right sides for i and j . That is,

$$\frac{\gamma_i}{\beta_i^I} \left(1 - \frac{1}{u_i} \right) = \frac{\gamma_j}{\beta_j^I} \left(1 - \frac{1}{u_j} \right).$$

Thus, the solutions u_i and u_j must be less than one for $\mathcal{R}_0 > 1$ which means the solutions q_i and q_j satisfy $0 < q_j < q_i < 1$ and if $\mathcal{R}_0 \leq 1$ there is only one fixed point and $q_j = q_i = 1$. It also follows from (7), that all other $q_k = 1$ since

$$\sum_{k=1}^n \frac{N_k}{N} q_k = 1.$$

In the special case $\beta_i^I / \gamma_i = \beta / \gamma$ for all $i = 1, \dots, n$, it follows that $\mathcal{R}_0 = \beta / \gamma$ and $u_i = u_j = u$ for all i, j . In particular, the fixed point of the pgfs simplifies to a single equation:

$$f(u) = \frac{\gamma + \beta u^2}{\gamma + \beta} = u,$$

with two solutions for u , $q = \gamma / \beta$ and 1. If $\mathcal{R}_0 > 1$, then the minimal fixed point is $q = 1 / \mathcal{R}_0$ and if $\mathcal{R}_0 > 1$, the minimal fixed point is $q = 1$.

For Model 2, the waiting time parameter for group j is $\omega_j = \gamma_j + \sum_{k=1}^n \beta_k^S \frac{N_k}{N}$ and the pgfs f_j have the following form

$$f_j(u_1, u_2, \dots, u_n) = \frac{\gamma_j + u_j \sum_{k=1}^n \beta_k^S \frac{N_k}{N} u_k}{\gamma_j + \sum_{k=1}^n \beta_k^S \frac{N_k}{N}} \tag{8}$$

for $u_i \in [0, 1], i, j = 1, \dots, n$. These pgfs also satisfy the three properties (i)-(iii) and a similar result holds for the probability of extinction when the recovery rates between two groups differ. When transmission depends on the susceptibility of the group j , then differences between probability of extinction of the groups only depends on the average duration of infection $1/\gamma_j$. This is reasonable as there are no differences between the groups once they become infected, except for the duration of infection. The proof of the following theorem is similar to that of Theorem 1.

Theorem 2. *Assume in the MBP approximation for Model 2 that the recovery rates satisfy $\gamma_i > \gamma_j$ for some i and j . Then the probability of extinction for group i and j satisfy one of the following:*

- (a) *If $\mathcal{R}_0 > 1$, then $0 < q_j < q_i < 1$ and all other extinction probabilities are less than one.*
- (b) *If $\mathcal{R}_0 \leq 1$, then $q_i = 1$ for all $i=1, \dots, n$.*

In the special case, $\gamma_i = \gamma$ for all $i = 1, \dots, n$, if $\mathcal{R}_0 = \sum_k \beta_k^S \frac{N_k}{N} / \gamma > 1$, then $q_i = 1/\mathcal{R}_0$ and if $\mathcal{R}_0 \leq 1$, then $q_i = 1$ for all $i = 1, \dots, n$.

Proof. Assuming $\gamma_i > \gamma_j$ and rearranging the expression $f_j(u_1, \dots, u_n) = u_j$ leads to

$$\sum_{k=1}^n \beta_k^S \frac{N_k}{N} (u_k - 1) = \gamma_j \left(1 - \frac{1}{u_j} \right).$$

Equating the right sides for i and j leads to

$$\gamma_i \left(\frac{1}{u_i} - 1 \right) = \gamma_j \left(\frac{1}{u_j} - 1 \right).$$

The remaining steps of the proof for (a) and (b) and the special case with $\gamma_i = \gamma$ follow in a manner similar to the proof of Theorem 1 by noting that $\mathcal{R}_0 = \sum_k \beta_k^S \frac{N_k}{N} / \gamma$.

Theorems 1 and 2 can be extended to more general models with latent or exposed stages when there are no disease-related deaths, e.g., SEIR-type models. The same pgfs apply to stochastic SEIR multigroup models [27].

4 Duration of an Epidemic

In general, the pdf for the duration of an epidemic is a mixture of two distributions corresponding to the minor and major epidemics,

$$D_{I(0)}(t) = \mathbb{P}_{minor}d_1(t) + \mathbb{P}_{major}d_2(t), \quad t \in [0, \infty). \quad (9)$$

where \mathbb{P}_{minor} and \mathbb{P}_{major} depend on the initial number of infected individuals $I(0)$ as in equation (5).

Various computational methods have been used to estimate the duration of an epidemic, e.g., [5, 7, 11, 35]. For example, the duration of a minor epidemic can be estimated from the system of differential equations in (4) with zero initial conditions. These probabilities give an estimate for the cumulative distribution function (cdf) for duration of a minor epidemic, as the zero state is an absorbing state. The system of differential equations is solved until a time $T \gg 0$ such that $p_{(e_j, \mathbf{0})}(T) \approx q_j$. The numerical solution $p_{(e_j, \mathbf{0})}(t)$ on $[0, T]$ is scaled by q_j , to yield a cdf for initial conditions $I_j(0) = 1$ and $I_k(0) = 0$ for $k \neq j$. For other initial conditions, $I_j(0) = i_j$, the probabilities are raised to the power i_j . Taking the derivative of the cdf with respect to time, yields the pdf for the duration of infection.

4.1 Best Fitting Distributions

As no analytical formulas for the duration of infection for the n -group model are known, some well-known continuous pdfs including normal gamma, lognormal, and Weibull, are used for comparison purposes. Separate fits are applied for the minor and for the major epidemic, scaling the pdfs by either \mathbb{P}_{minor} or \mathbb{P}_{major} . The goal is not to find the best distribution among all possible candidates, but to make some comparisons of the epidemic duration in the CTMC model with some well-known distributions. In particular, software available in **R** is applied and the fit for each cumulative distribution compared using the skewness-kurtosis graph generated from numerical solution of the CTMC model [10]. The Cullen and Frey graph [10] in the “fitdistrplus” package in the statistical computing package **R** is used estimate the model parameters, compare the various models, and to test for the goodness of fit. The Kolmogorov-Smirnov goodness of fit test statistic is applied and the quantile-quantile plots (Q-Q plots) are used to visualize this fit [24, 26]. This package handles both discrete and continuous data and applies maximum likelihood estimation to give the best fit parameters, the Akaike’s Information Criteria (AIC) among competing models and Q-Q plots for the best fitting cumulative distribution. The model with the smallest AIC value from the maximum-likelihood estimation is the best fitting model, using the value $-2\loglikelihood + 2n_{par}$ for comparison, where $n_{par} = 2$. The mean and standard deviation are reported for each distribution, e.g., if the best fit is a gamma function, then we write $\text{Gamma}(\mu, \sigma)$, where the mean is μ and standard deviation is σ .

4.2 Numerical Examples

Numerical examples for Models 1 and 2 with three groups are presented to illustrate the epidemic duration. The parameter values for the CTMC model, transmission and recovery rates and the reproduction numbers are defined in equations (2) and (3) and are summarized in Table 2. The parameters are hypothetical but reasonable based on a time scale of one day, e.g., the average duration of the infection for group 1 is one day but for group 3 it is five days. Group 1 has the largest size and the smallest reproduction number, and group 3 has the largest reproduction number. The two models have the same parameter values and the same threshold values for \mathcal{R}_0 and \mathcal{R}_{0i} , $i = 1, 2, 3$. Since $\mathcal{R}_{01} < \mathcal{R}_{02} < \mathcal{R}_{03}$ and $\gamma_1 > \gamma_2 > \gamma_3$, the fixed points for these models satisfy $q_1 > q_2 > q_3$ (Theorems 1 and 2). The Gillespie algorithm [16] is used to simulate the CTMC model until absorption into a disease-free state, where the total infectious population equals zero. For each of the figures, probability histograms are generated from 10^5 sample paths.

$$\begin{aligned} \text{Model 1 : } (q_1, q_2, q_3) &= (0.859, 0.550, 0.234) \\ \text{Model 2 : } (q_1, q_2, q_3) &= (0.650, 0.527, 0.271). \end{aligned} \tag{10}$$

Table 2 Parameter values for Models 1 and 2 with three groups, $N = 2000$ and $\mathcal{R}_0 = 2.8$

Parameters	Group 1	Group 2	Group 3
N_i	1200	400	400
β_i^I or β_i^S	0.5	1.5	2
γ_i	1.0	0.6	0.2
\mathcal{R}_{0i}	0.5	2.5	10

Only three numerical examples are presented for each model, with three sets of initial conditions $I(0) = (2, 0, 0)$, $I(0) = (0, 2, 0)$, and $I(0) = (0, 0, 2)$. To fit the pdfs for minor and major epidemics the data are divided into two sets (values for $t < 20$ and $t \geq 20$) and scaled by \mathbb{P}_{minor} and \mathbb{P}_{major} (Figures 1 and 2). The best fitting distributions, calculated separately for minor and major epidemics, among normal, gamma, lognormal, and Weibull (smallest AIC value, Table 5 in Appendix 7) to the data generated from 10^5 sample paths are

$$\begin{aligned} D_{(2,0,0)}(t) &= 0.738\text{Gamma}(1.633, 1.326)(t) + 0.262\text{Lognormal}(51.092, 7.837)(t) \\ D_{(0,2,0)}(t) &= 0.302\text{Gamma}(2.327, 1.957)(t) + 0.698\text{Lognormal}(50.242, 7.783)(t) \\ D_{(0,0,2)}(t) &= 0.055\text{Weibull}(3.115, 2.368)(t) + 0.945\text{Lognormal}(49.061, 7.665)(t). \end{aligned} \tag{11}$$

For Model 1, the duration of minor epidemic is also calculated from the differential equations in (4). Mean and standard deviation of this pdf are reported in Table 3.

Table 3 Mean and standard deviation for duration of a minor epidemic in Model 1, calculated from the differential equations in (4).

Initial value	\mathbb{P}_{minor}	Mean	Standard Deviation
$I(0) = (2, 0, 0)$	0.738	1.631	1.455
$I(0) = (0, 2, 0)$	0.302	2.305	1.986
$I(0) = (0, 0, 2)$	0.055	3.042	2.346

The curves for duration fitted from the statistical package **R** in Table 5 in Appendix 7 are close to those calculated from the MBP in Table 3 with the closest agreement for $I(0) = (0, 0, 2)$. But applying the Kolmogorov-Smirnov goodness of fit test, the null hypothesis that the specific distribution is a good fit is rejected at the significance level $\alpha = 0.05$ for cases $I(0) = (2, 0, 0)$ and $I(0) = (0, 2, 0)$ [33]. But for the case $I(0) = (0, 0, 2)$, the null hypothesis that Weibull is the correct distribution cannot be rejected at significance level $\alpha = 0.05$; the critical value is 0.0183 and Kolmogorov-Smirnov test statistic is 0.0122 (Table 5 in Appendix 7). The Q-Q plots in Appendix 8 show that the extreme values of the duration do not agree with the cdf of the normal distribution; the CTMC duration for a major epidemic is right-skewed [39]. The graphs of the best fitting pdfs for epidemic duration are overlaid on the probability histograms generated from the 10^5 sample paths of the CTMC in Figure 1.

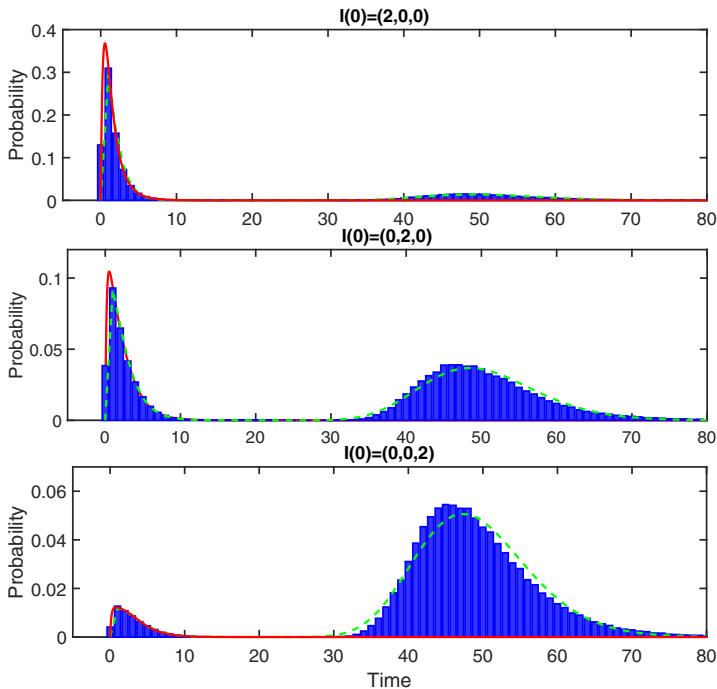


Fig. 1 Duration of an epidemic in Model 1 with 3 groups. Approximations for minor and major epidemic by the best fitting distribution yield either a gamma or Weibull and a lognormal distribution, respectively. The overlaid dashed curves represent the best fitting pdfs and the overlaid solid curves for the minor epidemic are the approximation from the MBP equations (4).

Similar analyses and fit of the pdfs were performed for Model 2 with three groups, applying parameter values in Table 2. The best fitting distributions for epidemic duration, calculated separately for minor and major epidemics, for the three sets of initial conditions are

$$\begin{aligned}
 D_{(2,0,0)}(t) &= 0.422\text{Lognormal}(1.422, 1.751)(t) + 0.578\text{Lognormal}(48.39, 7.234)(t) \\
 D_{(0,2,0)}(t) &= 0.277\text{Gamma}(1.775, 1.470)(t) + 0.723\text{Lognormal}(48.25, 7.353)(t) \\
 D_{(0,0,2)}(t) &= 0.073\text{Gamma}(2.626, 2.046)(t) + 0.927\text{Lognormal}(47.68, 7.280)(t).
 \end{aligned}
 \tag{12}$$

The mean and standard deviation of the minor epidemic in Model 1, calculated from the numerical solution of the pdf in (4) are summarized in Table 4.

Table 4 Mean and standard deviation for Duration of a minor epidemic in Model 2, calculated from the differential equations in (4).

Initial value	\mathbb{P}_{minor}	Mean	Standard Deviation
$I(0) = (2, 0, 0)$	0.422	1.356	1.349
$I(0) = (0, 2, 0)$	0.277	1.765	1.583
$I(0) = (0, 0, 2)$	0.073	2.570	2.045

Table 6 in Appendix [30] summarizes the smallest AIC values for all distributions tested. At significance level $\alpha = 0.05$, the only distribution that cannot be rejected is the minor epidemic with $I(0) = (0, 0, 2)$. The critical value in this case is 0.0185, and the Kolmogorov-Smirnov test statistic is 0.0142 for a gamma distribution [33]. The Q-Q plots in Appendix 8 (Figure 5) show that these fitted distributions differ from the CTMC multigroup model in the tails of the distribution; the CTMC duration for a major epidemic is right-skewed [39]. The graphs of the best fitting distributions for Model 1 are overlaid on the probability histograms in Figure 2.

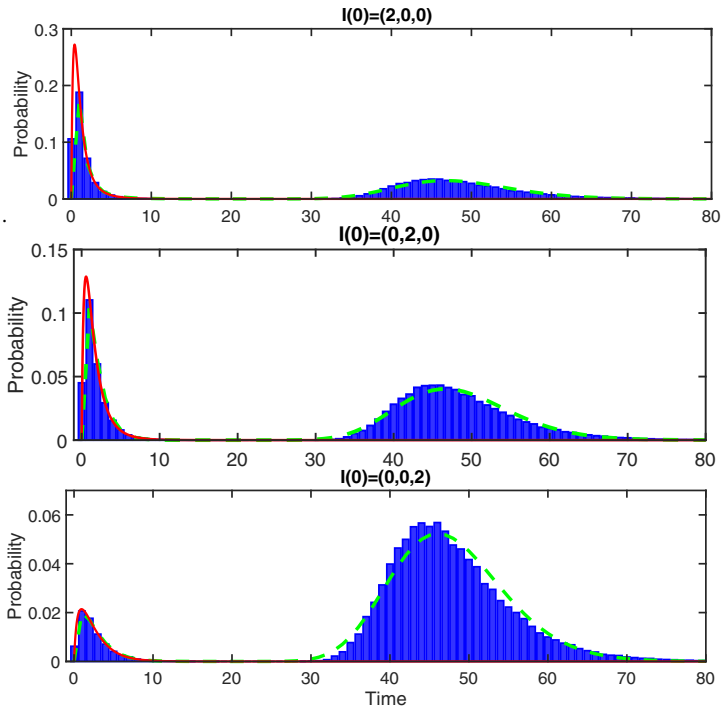


Fig. 2 Duration of an epidemic for Model 2 with 3 groups. Approximations for minor and major epidemic by the best fitting distribution yield either a lognormal or a gamma distribution. The overlaid dashed curves represent the best fitting pdfs and the overlaid solid curves for the minor epidemic are the approximation from the MBP equations (4).

5 Final Size of an Epidemic

The final size of an epidemic is a discrete distribution with values from $\{0, \dots, N\}$. As in the case for the duration of an epidemic, the final size of an epidemic is a mixture of two distributions, corresponding to minor and major epidemics,

$$H_{I(0)}(x) = \mathbb{P}_{minor}h_1(x) + \mathbb{P}_{major}h_2(x), \quad x \in \{0, 1, 2, \dots, N\}.$$

For the large population size $N = 2000$ and group sizes $N_i \geq 400$, the discrete distributions h_i are approximated with continuous pdfs. The discrete distributions Poisson, binomial, negative binomial, and geometric fit poorly, especially for the minor epidemic.

5.1 Best Fitting Distributions

Analytical formulas for final size in the stochastic multigroup model are not known. The continuous pdfs normal, gamma, lognormal, and Weibull are used for comparison purposes with the CTMC final size. Separate fits are applied for the final size in the case of either a minor or a major epidemic, separating the data into two sets (size < 1000 and ≥ 1000) and scaling the pdfs by either \mathbb{P}_{minor} or \mathbb{P}_{major} . As in the fits for duration, the goal in fitting final size is not to find the best distribution among the possible candidates, but to make comparisons of the epidemic final size in the CTMC model with these well-known distributions. Also, a MBP approximation is used to estimate final size in a minor epidemic and the mean final size for a major epidemic is compared with a formula for epidemic final size from the ODE multigroup model [22, 23].

5.2 Minor Epidemic

The final size of a minor epidemic for the SIR model was approximated by Bailey in 1975 for $I(0) = 1$ and generalized to $I(0) = s > 1$ [5, 15, 17]:

$$\mathbb{P}(n) = \frac{s(2n - s)! \mathcal{R}_0^{n-s}}{(2n - s)n!(n - s)!(\mathcal{R}_0 + 1)^{2n-s}}, \quad n = s, s + 1, \dots, \tag{13}$$

where $\mathbb{P}(n)$ is the probability the final size equals n . The value of n includes the initial number s . This estimate comes from a MBP approximation with no bound on the final size.

To derive an approximate final size of a minor epidemic, similar to Bailey, we also apply methods used in the MBP approximation of the multigroup model. Here, we only discuss the computations for the case of i_j initial infectious individuals from group j , and either no new infections or at most one new infection. The transition probabilities in Table 1 are applied with the assumption that $s_j = N_j/N = \alpha_j$.

Suppose $I_j(0) = i_j$ and $I_k(0) = i_k = 0$ for $k \neq j$ and there are no new infections, then all i_j individuals must recover. In Model 1, the recovery of the first individual has approximate probability

$$\frac{\gamma_j i_j}{\sum_{\ell=1}^n [\alpha_\ell \sum_{k=1}^n \beta_k^\ell i_k] + \sum_{k=1}^n \gamma_k i_k} = \frac{\gamma_j}{\beta_j^\ell + \gamma_j}.$$

Thus, for Model 1, the probability that all i_j individuals recover is simply

$$\left[\frac{\gamma_j}{\beta_j^I + \gamma_j} \right]^{i_j} \tag{14}$$

This formula is equivalent to the formula $\mathbb{P}(s)$ in (13) for $i_j = s$. For Model 2, the probability that all i_j individuals recover is

$$\left[\frac{\gamma_j}{\sum_{j=1}^n \beta_j^S \alpha_j + \gamma_j} \right]^{i_j} \tag{15}$$

If there is only one new transmission prior to recovery, then the transmission to other groups and the order of the transmission and the recovery events must be considered. This leads to a combinatorial problem. For example, if the first event is a new transmission, there are n potential new transmissions, one to each of the n groups. If the transmission is to the same group j , then there must be $i_j + 1$ recoveries (order is not important here). But for a transmission to group $k \neq j$, the i_j individuals in group j must recover and also the one individual in group k (order of these $i_j + 1$ recovery events is important).

The full combinatorial problem is not considered here, instead the estimates from (14) and (15) are applied in the numerical examples for three groups. Note that the relation between the two formulas in (14) and (15) depends on the relation between β_j^I and the weighted sum $\sum_j \beta_j^S \alpha_j$.

5.3 Major epidemic

Recently, Magal et al. [22, 23] computed the final epidemic size in a ODE multi-group model. The implicit formula for final size is computed for each group i by solving the following system of equations for $S_i(\infty)$,

$$S_i(\infty) = S_i(0) \exp \left(\sum_{j=1}^n \frac{\beta_{ij}}{N\gamma_j} [S_j(\infty) - S_j(0) - I_j(0)] \right), \quad i = 1, \dots, n, \tag{16}$$

with transmission rates β_{ij} for Models 1 and 2 given in (2). The final size of the epidemic for group i is $N_i - S_i(\infty)$ and the final epidemic size for the entire population is $N - \sum_{i=1}^n S_i(\infty)$.

5.4 Numerical Examples

Similar to the numerical examples in Section 4.2, only three numerical examples are presented for each model, with initial conditions $I(0) = (2, 0, 0)$, $I(0) = (0, 2, 0)$, and $I(0) = (0, 0, 2)$. All of the fitted distributions for final size do not include the two initial infectious individuals. The values q_i , $i = 1, 2, 3$ from the MBP are given in equations (10) and the parameter values in Table 2. The best fitting distributions for final size, calculated separately for minor and major epidemics, with smallest AIC value are lognormal for minor epidemic and normal for major epidemic,

$$\begin{aligned}
 H_{(2,0,0)}(t) &= 0.738\text{Lognormal}(2.598, 2.341)(t) + 0.262\text{Normal}(1846, 24.86)(t) \\
 H_{(0,2,0)}(t) &= 0.302\text{Lognormal}(4.82, 5.465)(t) + 0.698\text{Normal}(1847, 24.77)(t) \\
 H_{(0,0,2)}(t) &= 0.055\text{Lognormal}(7.50, 9.354)(t) + 0.945\text{Normal}(1846, 24.80)(t).
 \end{aligned} \tag{17}$$

The Kolmogorov-Smirnov goodness of fit values are given in Appendix 7 in Tables 7 and 8 for Models 1 and 2, respectively. For the large sample size 10^5 all the fitted distributions are rejected at the $\alpha = 0.05$ level. The normal Q-Q plots for the major epidemic show that the final size in the CTMC model differ in tails of the distribution. The CTMC final size distribution is left-skewed [39]. The final size of the ODE multigroup Model 1, calculated from the formulas in (16) equals 1848 (not counting the two initial infectious individuals). This estimate agrees well with the mean final size for a major epidemic for the fitted distributions, ≈ 1846 -1847. Probability histograms of the final size for each set of initial conditions and the best fitting distributions for Model 1 are graphed in Figure 3.

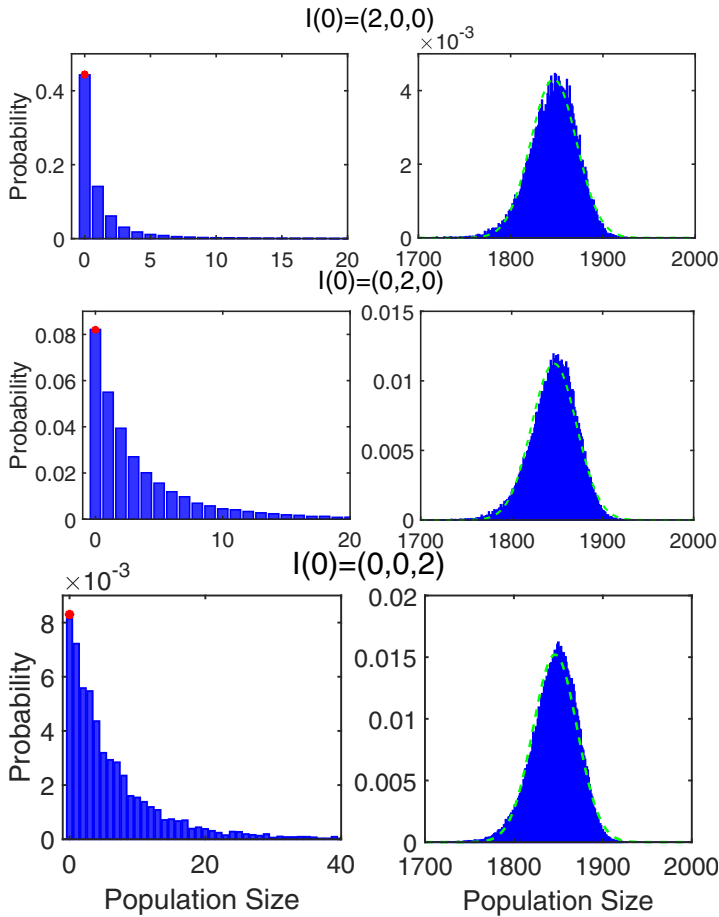


Fig. 3 Final size of an epidemic in Model 1 with 3 groups. The dots denote the probability there is no additional infectious individuals, as calculated from (14) and the dashed curves are the best fitting distributions for a major epidemic.

The corresponding best fitting distributions for final size, calculated separately for minor and major epidemics, in Model 2 are

$$\begin{aligned}
 H_{(2,0,0)}(t) &= 0.422\text{Lognormal}(2.388, 1.927)(t) + 0.578\text{Normal}(1402.7, 35.787)(t) \\
 H_{(0,2,0)}(t) &= 0.277\text{Lognormal}(2.703, 2.333)(t) + 0.723\text{Normal}(1401.6, 35.889)(t) \\
 H_{(0,0,2)}(t) &= 0.073\text{Lognormal}(3.309, 3.112)(t) + 0.927\text{Normal}(1401.6, 35.827)(t).
 \end{aligned}
 \tag{18}$$

The final size in the ODE multigroup Model 2, calculated from equations (16), is 1403-1404 (not counting the two initial infectious individuals). This is in close agreement with the mean final size in the best fitting distributions, ≈ 1402 -1403. Graphs of the best fitting distributions are overlaid on the probability histograms generated from 10^5 sample paths of the CTMC model.

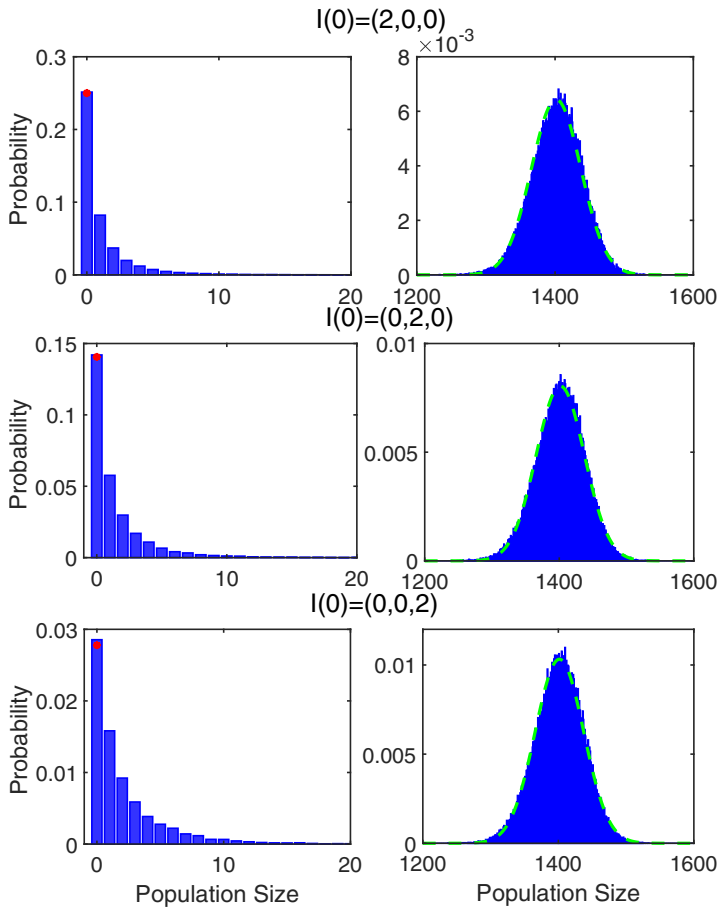


Fig. 4 Final size of an epidemic in Model 2 with 3 groups. The dots denote the probability there is no additional infectious individuals, as calculated from formula (15) and the dashed curves are the best fitting distributions for the major epidemic.

6 Discussion

The duration and final epidemic size of the CTMC multigroup model with transmission dependent on either the infectious group or the susceptible group were fit to some well-known distributions. The best fitting distributions show that there are distinct differences between the CTMC duration and final size and these well-known distributions. The differences are greatest in the tails of the distribution, where the CTMC duration is right-skewed and the CTMC final size is left-skewed (Figures 5 and 6). It was also shown that the final size estimate from the underlying ODE multigroup model is a good prediction of the mean final size for a major epidemic in the stochastic multigroup models. It is notable that the final size in Model 1 (Fig-

ure 3 and in equation (17)) is much greater than in Model 2 (Figure 4 and equation (18)) but the two models do not differ significantly in the epidemic duration. Model 1 is applicable to emerging diseases and Model 2 to re-emerging diseases. Therefore, a larger final size is of significant public health concern, as superspreaders are responsible for recent emerging diseases (SARS, MERS, and Ebola).

The value of \mathcal{R}_0 plays an important role in duration and final size. Generally, if $\mathcal{R}_0 \gg 1$ the final sizes are larger but of shorter duration but for $\mathcal{R}_0 \approx 1$, the final sizes are smaller but of longer duration (e.g., [3, 9, 22, 23, 35]). Our results depend on \mathcal{R}_0 and the population sizes. In the numerical examples, $\mathcal{R}_0 = 2.8$ and N_i ranges from 400 to 1200. For \mathcal{R}_0 sufficiently large, the minor and major epidemics are well separated. For large population sizes, application of the MBP approximation is possible. Simulations of other numerical examples for Models 1 and 2 with two and three groups, group population sizes > 50 , initial conditions with 1 or 2 infectious individuals, and $\mathcal{R}_0 > 1.5$ showed similar qualitative results for probability of a minor or a major epidemic and final size.

The assumptions regarding the transmission rates for Models 1 and 2 are restrictive. Generally individual heterogeneity is much more complex. Behavior, genetics, the environment, and physiological conditions impact disease transmission [34]. This investigation has shed some light on duration and final size in these two specific multigroup models. But further investigation is needed to understand the impact of group size, transmission and recovery rates on duration and final size in more general multigroup models. The methods applied in this investigation can be extended to more general stochastic multigroup models and to other stochastic epidemic models with multiple stages, patches, hosts, or pathogens.

Appendix

7 AIC and Goodness of Fit

The following four tables summarize the AIC values and the Kolmogorov-Smirnov goodness of fit values for duration and final size for Models 1 and 2.

Table 5 AIC values for duration of a minor and a major epidemic in Model 1. †KS=Kolmogorov-Smirnov goodness of fit values, reported for the model with the smallest AIC value.

I(0)=(2,0,0)				
pdf	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	213169	0.0286	182027	-
Lognormal	214606	-	181319	0.0309
Normal	265379	-	184122	-
Weibull	215054	-	189547	-
I(0)=(0,2,0)				
pdf	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	110271	0.0319	482261	-
Lognormal	112586	-	480310	0.0315
Normal	128675	-	487997	-
Weibull	110555	-	502391	-
I(0)=(0,0,2)				
KS†	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	22461	-	653311	-
Lognormal	23205	-	650403	0.0344
Normal	24714	-	661706	-
Weibull	22447	0.0122	681952	-

Table 6 AIC values for duration of a minor and a major epidemic in Model 2. †KS=Kolmogorov-Smirnov goodness of fit values, reported for the model with the smallest AIC value.

I(0)=(2,0,0)				
pdf	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	108353	-	392755	-
Lognormal	107487	0.0286	391218	0.0322
Normal	145811	-	397243	-
Weibull	109334	-	408980	-
I(0)=(0,2,0)				
pdf	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	85563	0.0267	492022	-
Lognormal	86381	-	489985	0.0331
Normal	105433	-	497925	-
Weibull	86159	-	513133	-
I(0)=(0,0,2)				
pdf	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	28152	0.0142	629820	-
Lognormal	28686	-	627161	0.0327
Normal	32054	-	637516	-
Weibull	28273	-	657153	-

Table 7 AIC values for final size of a minor and a major epidemic in Model 1 †KS=Kolmogorov-Smirnov goodness of fit values, reported for the model with the smallest AIC value.

I(0)=(2,0,0)				
pdf	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	117445	-	238963	-
Lognormal	107041	0.2796	239016	-
Normal	157697	-	238864	0.0361
Weibull	119389	-	239473	-
I(0)=(0,2,0)				
pdf	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	114808	-	643904	-
Lognormal	110675	0.147	644036	-
Normal	139172	-	643653	0.0349
Weibull	115434	-	645599	-
I(0)=(0,0,2)				
pdf	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	28704	-	873844	-
Lognormal	28242	0.0929	874022	-
Normal	33468	-	873506	0.033
Weibull	28795	-	876315	-

Table 8 AIC values for final size of a minor and a major epidemic in Model 2. †KS=Kolmogorov-Smirnov goodness of fit values, reported for the model with the smallest AIC value.

I(0)=(2,0,0)				
pdf	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	63361	-	575268	-
Lognormal	58745	0.2870	575385	-
Normal	80800	-	5750743	0.0188
Weibull	65049	-	580039	-
I(0)=(0,2,0)				
pdf	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	54486	-	720316	-
Lognormal	51130	0.2547	720441	-
Normal	68414	-	720115	0.0149
Weibull	55592	-	727130	-
I(0)=(0,0,2)				
pdf	Minor		Major	
	AIC	KS†	AIC	KS†
Gamma	19597	-	927418	-
Lognormal	18753	0.2083	927594	-
Normal	23750	-	927127	0.0174
Weibull	19864	-	935818	-

8 Q-Q Plots

The lognormal distribution is the best fit for the duration of a major epidemic in (11) and (12). The Q-Q plot of the logarithm of this distribution for initial condition $I(0) = (0,0,2)$ gives a normal Q-Q plot, graphed in Figure 5. The normal Q-Q plot shows that the true distribution differs from this lognormal distribution in the extreme values; the CTMC duration is right-skewed [39]. Due to the large number of points in the Q-Q plots, we only show the fit based on a major epidemic for 10^4 sample paths which is similar to 10^5 sample paths.

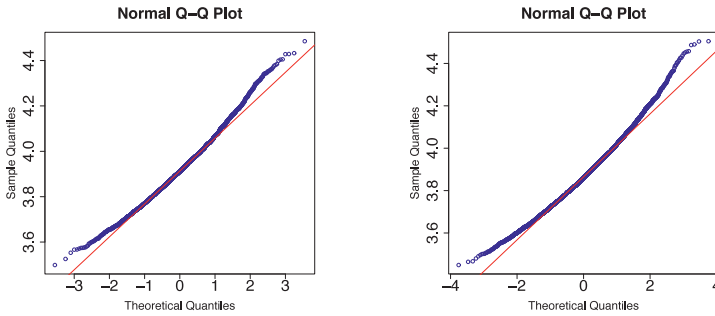


Fig. 5 Normal Q-Q plot of duration of major epidemic for Model 1 (left) and Model 2 (right) with 3 groups for initial value $(0, 0, 2)$. Sample size is $(1 - q_3^2)10^4$.

The normal distribution is the best fit for final size of a major epidemic in (17) and (18). The normal Q-Q plot for the final size with initial condition $I(0) = (0, 0, 2)$ is graphed in Figure 6 which shows differences from the normal in the extreme values; the CTMC final size is left-skewed [39].

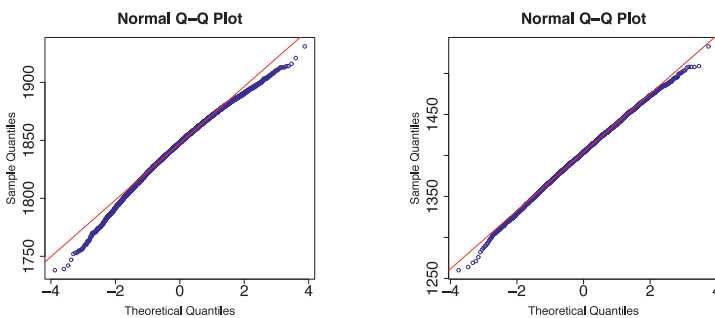


Fig. 6 Normal Q-Q Plot of final size of major epidemic for Model 1 (left) and Model 2 (right) with 3 groups for initial condition $(0, 0, 2)$. Sample size is $(1 - q_3^2)10^4$.

References

1. L. J. S. Allen. A primer on stochastic epidemic models: formulation, numerical simulation, and analysis. *Infectious Disease Modelling*, 2(1):3–10, 2017.
2. L. J. S. Allen and P. van den Driessche. Relations between deterministic and stochastic thresholds for disease extinction in continuous- and discrete-time infectious disease models. *Mathematical Biosciences*, 243:99–108, 2013.
3. J. Arino, F. Brauer, P. van den Driessche, J. Watmough, and J. Wu. A final size relation for epidemic models. *Mathematical Biosciences and Engineering*, 4(2):159–175, 2007.
4. K. B. Athreya and P. E. Ney. *Branching Processes*. Springer-Verlag, New York, 1972.
5. N. T. J. Bailey. *The Mathematical Theory of Infectious Diseases and Its Applications*. Charles Griffin & Company, Ltd., London, 2nd edition, 1975.
6. F. Ball and D. Clancy. The final size and severity of a generalised stochastic multitype epidemic model. *Advances in Applied Probability*, 25(4):721–736, 1993.
7. A. D. Barbour. The duration of the closed stochastic epidemic. *Biometrika*, 62:477–482, 1975.
8. A. J. Black and J. V. Ross. Computation of epidemic final size distribution. *Journal of Theoretical Biology*, 367:159–165, 2015.
9. F. Brauer. Epidemic models with heterogeneous mixing and treatment. *Bulletin of Mathematical Biology*, 70(7):1869–1885, 2008.
10. A. C. Cullen and H. C. Frey. *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. Plenum Press, New York, 2nd edition, 1999.
11. D. J. Daley and J. Gani. *Epidemic Modelling: An Introduction*. Cambridge University Press, Cambridge, 1999.
12. O. Diekmann, J.A.P. Heesterbeek, and J.A.J. Metz. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28(4):365–382, 1990.
13. K. S. Dormann, J. S. Sinsheimer, and K. Lange. In the garden of branching processes. *SIAM Review*, 46(2):202–229, 2004.
14. C. J. Edholm, B. O. Emerenini, A. L. Murillo, O. Saucedo, N. Shakiba, X. Wang, L. J. S. Allen, and A. Peace. Searching for superspreaders: Identifying epidemic patterns associated with superspreading events in stochastic models. In *Understanding Complex Biological Systems with Mathematics*, pages 1–29. Springer, 2018.
15. C. P. Farrington, M. N. Kanaan, and N. J. Gay. Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, 4:279–295, 2003.
16. D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.
17. N. C. Grassly and C. Fraser. Seasonal infectious disease epidemiology. *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1600):2541–2550, 2006.
18. T. E. Harris. *The Theory of Branching Processes*. Springer-Verlag, Berlin, 1963.
19. J. A. P. Heesterbeek and M. G. Roberts. The type-reproduction number T in models for infectious disease control. *Mathematical Biosciences*, 206(1):3–10, 2007.
20. T. House, J. V. Ross, and D. Sirl. How big is an outbreak likely to be? Methods for epidemic final-size calculation. *Proceedings of the Royal Society A*, 469(2150):22, 2013.
21. W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London*, 115:700–721, 1927.
22. P. Magal, O. Seydi, and G. Webb. Final size of an epidemic for a two-group SIR model. *SIAM Journal on Applied Mathematics*, 76:2042–2059, 2016.
23. P. Magal, O. Seydi, and G. Webb. Final size of a multi-group SIR epidemic model: Irreducible and non-irreducible modes of transmission. *Mathematical Biosciences*, 301:59–67, 2018.
24. F. J. Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.
25. J. C. Miller. A note on the derivation of epidemic final sizes. *Bulletin of Mathematical Biology*, 74:2125–2141, 2012.

26. A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics, 3rd Ed.* McGraw-Hill series in probability and statistics, 1973.
27. A. Nandi and L. J. S. Allen. Stochastic two-group models with transmission dependent on host infectivity or susceptibility. *Journal of Biological Dynamics*, 2018 (In press).
28. National Institutes of Health (US). Biological Sciences Curriculum Study. NIH Curriculum Supplement Series [Internet]. Bethesda (MD): National Institutes of Health (US); 2007. Understanding Emerging and Re-emerging Infectious Diseases. <https://www.ncbi.nlm.nih.gov/books/NBK20370/>. Accessed: 2018-11-24.
29. S. Pénişson. *Conditional limit theorems for multitype branching processes and illustration in epidemiological risk analysis*. PhD thesis, Universität Potsdam; Université Paris Sud-Paris XI, 2010.
30. RDocumentation. Akaike's an information criterion. <https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/AIC>. Accessed: 2018-11-01.
31. M. G. Roberts and J. A. P. Heesterbeek. A new method for estimating the effort required to control an infectious disease. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1522):1359–1364, 2003.
32. Z. Shuai, J.A.P. Heesterbeek, and P. van den Driessche. Extending the type reproduction number to infectious disease control targeting contacts between types. *Journal of Mathematical Biology*, 67:1067–1082, 2013.
33. N. Smirnov. Table for Estimating the Goodness of Fit of Empirical Distributions. *Annals of Mathematical Statistics*, 19:279–281, 1948.
34. R. A. Stein. Super-spreaders in infectious diseases. *International Journal of Infectious Diseases*, 15:e510–e513, 2011.
35. W. Tritch and L. J. S. Allen. Duration of minor epidemic. *Infectious Disease Modelling*, 3:60–73, 2018.
36. P. van den Driessche and J. Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180:29–48, 2002.
37. P. van den Driessche and J. Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180:29–48, 2002.
38. P. Whittle. The outcome of a stochastic epidemic: A note on Bailey's paper. *Biometrika*, 42:116–122, 1955.
39. M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, 1968.
40. G. Wong, W. Liu, Y. Liu, B. Zhou, Y. Bi, and G.F. Gao. MERS, SARS, and Ebola: the role of super-spreaders in infectious disease. *Cell Host & Microbe*, 18(4):398–401, 2015.



\mathcal{H}_2 Dynamic Output Feedback Control for Hidden Markov Jump Linear Systems

A. M. de Oliveira, O. L. V. Costa, J. Daafouz

Abstract In this note, we discuss the design of \mathcal{H}_2 dynamic output feedback controllers for a class of jump systems whose switching is induced by a Markov chain. The observation model is based on hidden Markov chains, in which only a random variable conditioned on the jump process of the plant is available to the controller. In this context, we consider a type of sub-optimal *ad hoc separation procedure* in which a state-feedback controller is given in order to obtain the remaining controller matrices by means of the linear matrix inequality formulation. In the case of perfect observation of the Markov chain, the conditions also become necessary allowing us to calculate optimal \mathcal{H}_2 controllers also provided by the classical results of the literature. Clusterized and mode-independent controllers can also be synthesized via our formulation. Two illustrative examples are presented.

1 Introduction

The study of systems subject to abrupt changes has attracted a great deal of effort in the last decades, in part due to the ease of modelling complex dynamics in the same application. This phenomenon may arise in systems subject to faults or possessing different operation points, such as in the case of nonlinear systems. Particularly, the very nature of failures, that are usually random, asks for a class of systems that presents not only a set of different dynamics, but also a type of *switching* that could reach out for this kind of unpredictable behaviour. In this context, the so-called Markov jump linear systems, for short MJLS, has appeared as an important

A. M. de Oliveira and O. L. V. Costa

Departamento de Engenharia de Telecomunicações e Controle, Escola Politécnica da Universidade de São Paulo, CEP: 05508-010 - São Paulo, Brazil. e-mail: marcorin@usp.br, osvaldo@lac.usp.br

J. Daafouz

Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France. e-mail: Jamal.Daafouz@univ-lorraine.fr

tool for modelling abrupt dynamic changes and consequently acting in the system to ensure stability and performance properties. There is by now a great number of works concerning MJLS in the literature. We refer the interested reader to [4, 9, 10, 19, 36, 53], and the references therein.

Specifically, in the realm of control systems theory, the influence of the state space methods and the interest in optimal control due to the works of Rudolf Kalman in the 60's can still be felt nowadays and paved the way to the development of the \mathcal{H}_∞ theory in the 80's, see for instance, [16, 18], and the references therein. Among the contributions of that time, we mention the revival interest in the Lyapunov theory in [31] and [32], the introduction of the concepts of Controllability and Observability and the use of Calculus of Variations for solving the Linear Quadratic Regulator (LQR) control in [29], and the Kalman filter and the "geometric feel" in which it was solved in [30]. The synthesis of all the mentioned contributions is the elegant *Linear Quadratic Gaussian* (LQG) control theory and the so-called *separation principle*: the optimal quadratic control for linear systems is the joint use of the Kalman filter for estimating the states and the state-feedback control coming from the LQR theory. Earlier works on LQG are, for instance, [25] and [27], that led to a great deal of discussion, such as the ones in [3, 15, 16, 17, 18, 22, 28, 48, 54], and the references therein. An even more ever-lasting product of that time is the use of Riccati equations in control theory that curiously, as pointed out in [50], was used for the critics of the optimal control framework for solving the \mathcal{H}_∞ control in [18] or applied to robust control techniques in [17].

The extension of the LQG theory to MJLS is considered in [6, 26], and the more general \mathcal{H}_2 theory, in [13]. Especially concerning the \mathcal{H}_2 control presented in [13], we notice that the results echo the ones derived in the case without jumps, leading to the \mathcal{H}_2 *separation principle* for MJLS, in which the controller is composed by the optimal \mathcal{H}_2 state-feedback controller and the optimal \mathcal{H}_2 observer, both structures obtained through coupled algebraic Riccati equations (CARE). It is interesting to note that the use of the Kalman filter could be prohibitive in terms of memory usage for the *off-line* calculation of the gains, since it would depend on all possible trajectories of the Markov chain up to the current instant. Alternatively, by means of the linear matrix inequality (LMI) formulation, see, for instance, [5], design conditions for optimal \mathcal{H}_2 and \mathcal{H}_∞ dynamic output feedback controllers were presented in [23]. However, a fundamental characteristic shared by all those works is the hypothesis that the Markov chain (or the mode) can be perfectly measured, that is the same to assume that the controller would have access to some parts of the state. Such a controller is called *mode-dependent* in the literature. This assumption greatly simplifies the problem, but also raises very important questions, such as how to effectively measure the underlying jump process in order to use it in real processes.

Arguably, there are two most relevant formulations concerning *partial observations* of the Markov chain in the literature: the *cluster* and *mode-independent* cases. In the former case, introduced in [52], the modes of the Markov chain are grouped in disjoint sets called *clusters*, and then the controller would have access to which cluster the Markov chain currently is. As for the mode-independent formulation, it is considered that the controller would not have any access to the jump process of

the plant and then it would remain fixed throughout the time. Since all the states of the Markov chain can be grouped in a unique set, or conversely, can constitute a bijection with respect to the disjoint sets, it follows that the cluster case encompasses the mode-independent and -dependent formulations. A graph for a Markov chain of three states grouped in two clusters is shown in Fig. 1. Nowadays, the few

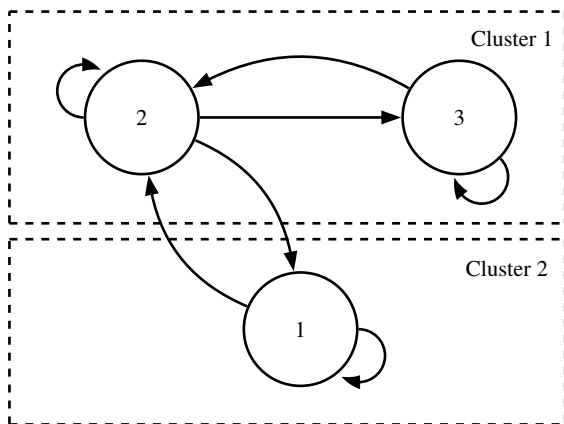


Fig. 1 Example of a Markov chain and clusterization

works studying the design of dynamic output feedback controllers for MJLS that consider the *cluster case* and the *mode-independent* formulations depend on very simplifying assumptions and/or present sub-optimal results due to the high degree of non-linearity involved. In [21], LMI design conditions were given for model-based controllers, and due to this structural choice, the system matrices must be equal inside a given cluster. By writing the original design problem as one of calculating static output feedback controllers, the works [38] and [39] presented sub-optimal conditions following the two steps procedure of [37] and [46]. More recently the work [40] introduced sub-optimal conditions for obtaining cluster controllers by means of an algorithm that uses a mode-dependent controller as an input.

Another trend that can be found in the literature is the use of observation models that are linked to Active Fault-tolerant Control Systems (AFTCS), that aim to present suitable approximations for the behavior of Fault Detection and Isolation (FDI) devices. In this sense, the model called *detector approach* or *hidden MJLS*, that was introduced in [8] and considered more recently in [11], consists of a hidden Markov chain (HMC) in which the jump process of the plant cannot be measured, but instead the controller would only have access to an observed variable that is conditioned on the Markov chain. This modelling could represent a very simple model for FDI processes, see, for instance, [11], or could be viewed as an asynchronous phenomenon between the controller and the plant, as characterized in [55]. A recent work that considered such modelling concerning stabilizing dynamic output feedback control is [43], in which a type of *ad hoc separation procedure* is presented,

echoing the results given in [45] for the uncertain linear time-invariant (LTI) systems. It must be pointed out that the detector approach encompasses the aforementioned cases: the situation of perfect observation of the Markov chain, as well as the cluster and mode-independent formulations. Alternatively, some works such as [34] and [35] employed a different observed variable that consists of another Markov process conditioned on the Markov chain of the plant. In this context, the works [1] and [2] studied the design of dynamic output feedback controllers for continuous-time MJLS, where in the latter one, the design conditions are given in LMI, but the resulting controller would also depend on the Markov chain of the plant. A more practical approach (that is similar to the one considered in this note) is presented in [1], in which bilinear matrix inequalities (BMI) conditions are obtained for mixed $\mathcal{H}_2/\mathcal{H}_\infty$ controllers. A similar observation model for the discrete-time case is used in [33] for the \mathcal{H}_∞ control, but as in [2], the final controller seems to depend also on the Markov chain of the plant.

In this note, we are concerned with the design of \mathcal{H}_2 dynamic output feedback controllers for hidden MJLS, that is, the observation model employed in [11] and [43]. By considering the similar transformations as in [43], we derive BMI conditions that, if fulfilled, provide dynamic output feedback controllers that switch according to the observed variable, guarantee the stability of the closed-loop system in some stochastic sense, and impose an upper bound on its \mathcal{H}_2 norm. For solving the BMI, we use the strategy that has been called the *ad hoc separation procedure* in the literature, see, for instance, [43] and [45], in which a state-feedback controller is provided for calculating the remaining “filter-like” structure, and then only LMIs must be solved. The final controller is composed by both the state-feedback gains of the first stage and the remaining calculated matrices. An additional and desirable property is also guaranteed and shown in this note, namely, that for the case in which we can perfectly measure the Markov chain, our conditions also become necessary, allowing us to obtain optimal \mathcal{H}_2 dynamic controllers such as the ones given in [11] and [23] (with some suitable modifications). Due to some properties of the hidden MJLS formulation, we are also able to obtain clusterized and mode-independent structures, thus providing alternative and arguably simpler design conditions compared to the ones presented in [21, 39, 38, 40], and the references therein. Besides, it must be pointed out that, even though the problems tackled in [1, 2, 33] are somewhat similar to the one considered in this work, the observation models, as well as the results, are different, as previously explained.

The structure of this chapter is as follows. Sect. 2 introduces the notation, Sect. 3, the preliminary discussions such as the problem formulation, the definitions and results, and the main goal. Sect. 4 presents the main result, that is, sufficient design conditions for the \mathcal{H}_2 dynamic output feedback control considering the asynchronous phenomenon between the controller and the plant, that becomes also necessary for the mode-dependent case. In Sect. 5, we present two examples. The first one traces the parallel between our work and the *separation principle* of [13]. The second one is inserted in the context of AFTCS, through the design of dynamic controllers for an unmanned aerial vehicle subject to actuator failures. Our final remarks are presented in Sect. 6 and some auxiliary results, in the Appendix.

2 Notation

For \mathbb{Y} and \mathbb{X} complex Banach spaces, we set $\mathbb{B}(\mathbb{Y}, \mathbb{X})$ as the space of bounded linear operators of \mathbb{Y} into \mathbb{X} , and for simplicity, $\mathbb{B}(\mathbb{Y}) \triangleq \mathbb{B}(\mathbb{Y}, \mathbb{Y})$. The spectral radius of $T \in \mathbb{B}(\mathbb{Y})$ is represented by $r_\sigma(T)$. The real n -dimensional Euclidean space is denoted by \mathbb{R}^n , and the norm bounded linear space of all $m \times n$ real matrices is represented by $\mathbb{B}(\mathbb{R}^n, \mathbb{R}^m)$, with $\mathbb{B}(\mathbb{R}^n) \triangleq \mathbb{B}(\mathbb{R}^n, \mathbb{R}^n)$. The superscript $'$ indicates the transpose of a matrix, the identity operator is represented by I (or by I_n , for a $n \times n$ identity matrix), the null operator, by 0 (or equivalently $0_{n \times m}$ whenever the dimensions are needed), the trace operator by $\text{Tr}(\cdot)$, and the block diagonal matrix, by $\mathbf{diag}(\cdot)$. Considering a square matrix $S \in \mathbb{B}(\mathbb{R}^n)$, we define the operator $\text{Her}(S) = S + S'$, and for a symmetric matrix, the symbol \bullet represents a symmetric block. For N and M positive integers, the sets \mathbb{N} and \mathbb{M} are defined, respectively, by $\mathbb{N} \triangleq \{1, 2, 3, \dots, N\}$ and $\mathbb{M} \triangleq \{1, 2, 3, \dots, M\}$. Furthermore, the set $\mathbb{H}^{n,m}$ represents the linear space of all N -sequences of real matrices $V = (V_1, V_2, \dots, V_N)$, $V_i \in \mathbb{B}(\mathbb{R}^n, \mathbb{R}^m)$, $i \in \mathbb{N}$, and for simplicity, we set $\mathbb{H}^n \triangleq \mathbb{H}^{n,n}$ and $\mathbb{H}^{n+} \triangleq \{V \in \mathbb{H}^n; V_i \geq 0, i \in \mathbb{N}\}$. For $P, V \in \mathbb{H}^{n+}$, we write that $P \geq V$ ($P > V$) if $P_i - V_i \geq 0$ ($P_i - V_i > 0$) for all $i \in \mathbb{N}$. The Banach space $(\|\cdot\|_2, \mathbb{H}^{n,m})$ is a Hilbert space with the norm induced by the inner product

$$\langle V; S \rangle \triangleq \sum_{i \in \mathbb{N}} \text{Tr}(V_i' S_i)$$

for $V, S \in \mathbb{H}^{n,m}$. Let $(\Omega, \mathfrak{F}, \{\mathfrak{F}_k\}, \text{Prob})$ be a stochastic basis, with $\mathbf{E}(\cdot)$ representing the expected value operator and $\mathbf{E}(\cdot | \cdot)$, the conditional expectation operator. For $A \in \mathfrak{F}$, $\mathbf{1}_A$ represents the indicator function of the event A (if $\omega \in A$, then $\mathbf{1}_A(\omega) = 1$).

3 Preliminaries

On a probability space $(\Omega, \mathfrak{F}, \text{Prob})$ with filtration $\{\mathfrak{F}_k\}$ we consider the MJLS

$$\mathcal{G} : \begin{cases} x(k+1) = A_{\theta(k)}x(k) + B_{\theta(k)}u(k) & + J_{\theta(k)}w(k) \\ y(k) = L_{\theta(k)}x(k) & + H_{\theta(k)}w(k) \\ z(k) = C_{\theta(k)}x(k) + D_{\theta(k)}u(k), \end{cases} \quad (1)$$

where $x(k) \in \mathbb{R}^n$ is the state variable, $u(k) \in \mathbb{R}^m$ is the control input, $w(k) \in \mathbb{R}^r$ is the exogenous input, $y(k) \in \mathbb{R}^p$ is the measured output, and $z(k) \in \mathbb{R}^q$ is the controlled output. The variable $\theta(k)$ is a Markov chain with state space \mathbb{N} respecting

$$\text{Prob}(\theta(k+1) = j | \mathfrak{F}_k) = \text{Prob}(\theta(k+1) = j | \theta(k)) = p_{\theta(k)j},$$

for all $j \in \mathbb{N}$, and transition probability matrix given by $\mathbb{P} \triangleq [p_{ij}]$. Considering the random variable $\theta_0 \sim \mu$, we set $\theta(0) = \theta_0$. We also set $x(0) = x_0$, where $x_0 \in \mathbb{R}^n$ is a second order random vector, unless otherwise stated.

Assumption 1 *The Markov chain $\theta(k)$ is ergodic, see, for instance, [47], with limiting distribution $\nu_i \triangleq \lim_{k \rightarrow \infty} \nu_i(k)$, where $\nu_i(k) \triangleq \text{Prob}(\theta(k) = i)$.*

Assumption 2 *The transition probability matrix \mathbb{P} is nondegenerate, i.e., $\sum_{i \in \mathbb{N}} P_{ij} > 0$ for all $j \in \mathbb{N}$.*

The only available variables to the controller are $y(k)$ and $\hat{\theta}(k)$, where $\hat{\theta}(k)$ represents the output of some detector, or an asynchronous behavior with respect to the Markov chain $\theta(k)$. The dynamic output feedback controller has the following structure,

$$\mathcal{C} : \begin{cases} x_c(k+1) = A_{c\hat{\theta}(k)}x_c(k) + B_{c\hat{\theta}(k)}y(k) \\ u(k) = C_{c\hat{\theta}(k)}x_c(k) \\ x_c(0) = x_{c0}, \end{cases} \quad (2)$$

where $x_c(k) \in \mathbb{R}^n$ and $x_{c0} \in \mathbb{R}^n$ is a second order random vector. The observed jump variable $\hat{\theta}(k)$ takes its values in the set \mathbb{M} , and, by considering the σ -field $\hat{\mathfrak{F}}_k$ generated by

$$\{x(0), x_c(0), w(0), \theta(0), \hat{\theta}(0), \dots, x(k), x_c(k), w(k), \theta(k)\},$$

for $k > 0$, and

$$\{x(0), x_c(0), w(0), \theta(0)\},$$

for $k = 0$, we assume that

$$\text{Prob}(\hat{\theta}(k) = l \mid \hat{\mathfrak{F}}_k) = \text{Prob}(\hat{\theta}(k) = l \mid \theta(k)) = \alpha_{\theta(k)l},$$

for all $l \in \mathbb{M}$. The set $\mathbb{M}_{\theta(k)}$ defines all possible outcomes of $\hat{\theta}(k)$ for a given $\theta(k)$, that is, $\mathbb{M}_{\theta(k)} \triangleq \{l \in \mathbb{M} : \alpha_{\theta(k)l} > 0\}$, and $\Upsilon \triangleq [\alpha_{il}]$. Finally, unless otherwise stated, w is taken as a wide-sense white noise sequence (see, for instance, [15]), that is, $\mathbf{E}(w(k)) = 0$ and $\mathbf{E}(w(k)w(s)') = I_r \delta_{k-s}$, independent of $\theta(k)$, $\hat{\theta}(k)$, and x_0 .

Remark 1 *The joint process $(\theta(k), \hat{\theta}(k))$ is a hidden Markov chain, see, for instance, [47], or a hidden Markov model (HMM). The properties of this model reflects in the MJLS theory with the following properties taken from [11]:*

- *The mode-dependent case. If $N = M$ and $\alpha_{ii} = 1$ for all $i \in \mathbb{N}$, we would have that $\hat{\theta} = \theta$, that is, the case of perfect observation of θ .*
- *The cluster case of [52]. Considering that there is a set \mathbb{M} for $M \leq N$ such that \mathbb{N} can be grouped in disjoint sets \mathbb{N}^s satisfying $\mathbb{N} = \bigcup_{s \in \mathbb{M}} \mathbb{N}^s$, by defining $g : \mathbb{N} \rightarrow \mathbb{M}$ such that $g(i) = s$ for all $i \in \mathbb{N}^s$, then g would map the Markov states into their respective clusters. Equivalently through the HMM formulation, we would have that $\mathbb{M}_i = \{g(i)\}$ and $\alpha_{ig(i)} = 1$, and therefore $\hat{\theta}(k)$ would indicate the corresponding cluster of $\theta(k)$.*
- *If $M = 1$ and $\alpha_{i1} = 1$ for all $i \in \mathbb{N}$, then we would get the mode-independent case, that is, the detector cannot provide any useful information regarding the Markov chain.*

Alternatively, the cluster case can also be obtained by means of the hidden MJLS formulation through the next assumption, taken from [41]. These characteristics will be discussed in details in the examples of Sect. 5.

Assumption 3 ([41]) *The modes of the Markov chain can be partitioned into κ disjoint subsets \mathbb{N}^s , such that $\bigcup_{s=1}^{\kappa} \mathbb{N}^s = \mathbb{N}$, and for all $s \in \{1, \dots, \kappa\}$, $i \in \mathbb{N}^s$, we have that $\mathbb{M}_i = \mathbb{M}^s$ for disjoint sets \mathbb{M}^s , $\bigcup_{s=1}^{\kappa} \mathbb{M}^s = \mathbb{M}$, and $\alpha_{il} = \alpha_l^s$, for all $l \in \mathbb{M}^s$.*

Connecting (1) and (2) yields to the following closed-loop system

$$\mathcal{G}_c : \begin{cases} \tilde{x}(k+1) = A_{\theta(k)\hat{\theta}(k)}\tilde{x}(k) + J_{\theta(k)\hat{\theta}(k)}w(k) \\ z(k) = C_{\theta(k)\hat{\theta}(k)}\tilde{x}(k), \end{cases} \quad (3)$$

where $\tilde{x}(k)' \triangleq [x(k)' \ x_c(k)']$, $\tilde{x}(k) \in \mathbb{R}^{2n}$, and

$$\left[\begin{array}{c|c} A_{\theta(k)\hat{\theta}(k)} & J_{\theta(k)\hat{\theta}(k)} \\ \hline C_{\theta(k)\hat{\theta}(k)} & 0 \end{array} \right] \triangleq \left[\begin{array}{cc|c} A_{\theta(k)} & B_{\theta(k)}C_c\hat{\theta}(k) & J_{\theta(k)} \\ B_c\hat{\theta}(k)L_{\theta(k)} & A_c\hat{\theta}(k) & B_c\hat{\theta}(k)H_{\theta(k)} \\ \hline C_{\theta(k)} & D_{\theta(k)}C_c\hat{\theta}(k) & 0 \end{array} \right]. \quad (4)$$

We present next the basic concepts that underlies our discussion.

Definition 1. System (3) with $w = 0$ is said to be stochastically stable (SS) if, for every θ_0 and every x_0 with finite second moment,

$$\|\tilde{x}\|_2^2 \triangleq \sum_{k=0}^{\infty} \mathbf{E} (\|\tilde{x}(k)\|^2) < \infty.$$

■

The set of admissible controllers is defined as follows

$$\mathcal{C} \triangleq \{ \mathcal{C} \text{ as in (2) such that (3) is SS} \}. \quad (5)$$

Consider the following operators for $V \in \mathbb{H}^{n_x}$,

$$\mathcal{E}_i(V) \triangleq \sum_{j \in \mathbb{N}} p_{ij}V_j, \quad (6)$$

$$\mathcal{L}_i(V) \triangleq \sum_{l \in \mathbb{M}_i} \alpha_{il}\mathcal{A}'_{il}\mathcal{E}_i(V)\mathcal{A}_{il}, \quad (7)$$

$$\mathcal{T}_j(V) \triangleq \sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} \alpha_{il}p_{ij}\mathcal{A}_{il}V_i\mathcal{A}'_{il}, \quad (8)$$

for $\mathcal{E}, \mathcal{L}, \mathcal{T} \in \mathbb{H}^{n_x}$ and $\mathcal{A}_{il} \in \mathbb{B}(\mathbb{R}^{n_x})$ for all $i \in \mathbb{N}$, $l \in \mathbb{M}_i$. The following theorem, adapted from [11], states necessary and sufficient conditions for evaluating the stochastic stability of (3) for a given controller structure \mathcal{C} .

Theorem 1 ([11, 12]). *The following assertions are equivalent for $\mathcal{A}_{il} = A_{il}$ and $n_x = 2n$ in (7)-(8).*

- i. $\mathcal{C} \in \mathfrak{C}$.
- ii. $r_\sigma(\mathcal{L}) < 1$.
- iii. $r_\sigma(\mathcal{T}) < 1$.
- iv. There exists $P \in \mathbb{H}^{n_x}$, $P > 0$, such that

$$P - \mathcal{L}(P) > 0. \tag{9}$$

- v. There exists $Q \in \mathbb{H}^{n_x}$, $Q > 0$, such that

$$Q - \mathcal{T}(Q) > 0. \tag{10}$$

Moreover, for $\mathcal{C} \in \mathfrak{C}$ there exists a unique solution $P \in \mathbb{H}^{n_x}$, $P > 0$, of $P = \mathcal{V}(P) + S$ for $\mathcal{V} \in \{\mathcal{L}, \mathcal{T}\}$ and $S \in \mathbb{H}^{n_x}$, $S > 0$. For $P = \mathcal{V}(P) + S$ and $\bar{P} = \mathcal{V}(\bar{P}) + \bar{S}$, if $S \geq (>) \bar{S} \geq (>) 0$, then $P \geq (>) \bar{P} \geq (>) 0$. ■

Define $A \triangleq (A_1, \dots, A_N)$, $B \triangleq (B_1, \dots, B_N)$, and $K \triangleq (K_1, \dots, K_M)$, where $K_l \in \mathbb{B}(\mathbb{R}^n, \mathbb{R}^m)$, $l \in \mathbb{M}$. We are now interested in defining the concept of stochastic stabilizability for hidden MJLS.

Definition 2 (Stochastic stabilizability). The pair (A, B) is said to be stochastically stabilizable if there exists K such that (9) ((10)) holds for $\mathcal{A}_{il} = A_i + B_i K_l$, for all $i \in \mathbb{N}$, $l \in \mathbb{M}_i$. ■

The set of stabilizing state-feedback controllers is defined in the following

$$\mathfrak{K} \triangleq \{K \text{ such that (9) or (10) hold for } \mathcal{A}_{il} = A_i + B_i K_l\}. \tag{11}$$

We now investigate the performance index considered in this work.

Proposition 1. If $\mathcal{C} \in \mathfrak{C}$, then

$$\begin{aligned} \lim_{k \rightarrow \infty} \mathbf{E}(\|z(k)\|^2) &= \sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} \alpha_{il} p_{ij} \text{Tr}(C_{il} \bar{Q}_i C'_{il}) \\ &= \sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} v_i \alpha_{il} \text{Tr}(J'_{il} \mathcal{E}_i(\bar{P}) J_{il}), \end{aligned}$$

where $\bar{Q} \in \mathbb{H}^{2n+}$ is the unique solution of

$$\bar{Q} = \mathcal{T}(\bar{Q}) + \bar{\mathbf{J}}, \tag{12}$$

for $\bar{\mathbf{J}}_j \triangleq \sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} v_i \alpha_{il} p_{ij} J_{il} J'_{il}$, $\bar{\mathbf{J}} \in \mathbb{H}^{2n+}$, and $\bar{P} \in \mathbb{H}^{2n+}$ is the unique solution of

$$\bar{P} = \mathcal{L}(\bar{P}) + \mathbf{C}, \tag{13}$$

for $\mathbf{C}_i \triangleq \sum_{l \in \mathbb{M}_i} \alpha_{il} C'_{il} C_{il}$, $\mathbf{C} \in \mathbb{H}^{2n+}$. ■

Proof. Define $\bar{Q}_i(k) \triangleq \mathbf{E}(x(k)x(k)'\mathbf{1}_{\theta(k)=i})$. Then, by recalling that $w(k)$ is independent of $\theta(k)$, $\hat{\theta}(k)$, and x_0 (and then also of $x(k)$), we have that

$$\bar{Q}_j(k+1) = \sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} p_{ij} \alpha_{il} [A_{il} \bar{Q}_i(k) A'_{il} + v_i(k) J_{il} J'_{il}] ,$$

where we recall that $v_i(k) = \text{Prob}(\theta(k) = i)$. Note also that

$$\begin{aligned} \mathbf{E}(\|z(k)\|^2) &= \text{Tr} \left(\mathbf{E} \left(C_{\hat{\theta}(k)\theta(k)} x(k) x(k)' C'_{\hat{\theta}(k)\theta(k)} \right) \right) \\ &= \sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} p_{ij} \alpha_{il} \text{Tr}(C_{il} \bar{Q}_i(k) C'_{il}) . \end{aligned}$$

Considering Assumption 1 and that $\mathcal{C} \in \mathfrak{C}$, we have that $v_i(k) \rightarrow v_i$ and $\bar{Q}_i(k) \rightarrow \bar{Q}_i$ if we take $k \rightarrow \infty$, implying the first claim of the proof. Finally, note that

$$\begin{aligned} \sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} \alpha_{il} p_{ij} \text{Tr}(C_{il} \bar{Q}_i C'_{il}) &= \langle \bar{\mathbf{C}}; \bar{\mathbf{Q}} \rangle \\ &= \langle \bar{\mathbf{P}} - \mathcal{L}(\bar{\mathbf{P}}); \bar{\mathbf{Q}} \rangle \\ &= \langle \bar{\mathbf{P}}; \bar{\mathbf{J}} \rangle \\ &= \sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} v_i \alpha_{il} \text{Tr}(J'_{il} \mathcal{E}_i(\bar{\mathbf{P}}) J_{il}) , \end{aligned}$$

and thus, the claim follows. \square

Definition 3 (\mathcal{H}_2 norm). Consider that $\mathcal{C} \in \mathfrak{C}$ and $x_0 = 0$. Let z_s be the controlled output of (3) if

$$w(k) = \begin{cases} e_s, & k = 0 \\ 0, & k > 0, \end{cases}$$

where $e_s \in \mathbb{R}^r$ is the s -th standard basis of \mathbb{R}^r . For

$$\|z_s\|_2^2 \triangleq \sum_{k=0}^{\infty} \mathbf{E}(\|z_s(k)\|^2) ,$$

the \mathcal{H}_2 norm of (3) is defined by

$$\|\mathcal{G}_c\|_2^2 \triangleq \sum_{s=1}^r \|z_s\|_2^2 .$$

■

Recalling that $\mu_i = \text{Prob}(\theta_0 = i)$, we have, after straightforward manipulations, see, for instance, [11], that

$$\|\mathcal{G}_c\|_2^2 = \sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} \alpha_{il} p_{ij} \text{Tr}(C_{il} \bar{Q}_i C'_{il}) \tag{14}$$

$$= \sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} \mu_i \alpha_{il} \text{Tr}(J'_{il} \mathcal{E}_i(\bar{\mathbf{P}}) J_{il}) , \tag{15}$$

where $\bar{P} \in \mathbb{H}^{2n+}$ is the solution of (13) and $\tilde{Q} \in \mathbb{H}^{2n+}$, the solution of

$$\tilde{Q} = \mathcal{F}(\tilde{Q}) + \mathbf{J}, \quad (16)$$

for $\mathbf{J}_j \triangleq \sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} \mu_i \alpha_{il} p_{ij} J_{il} J'_{il}$, $\mathbf{J} \in \mathbb{H}^{2n+}$. Note the similarity between (12) and (16), and the discussion in Proposition 1. It readily follows that if $\mu = \nu$, then

$$\|\mathcal{G}_c\|_2^2 = \lim_{k \rightarrow \infty} \mathbf{E}(\|z(k)\|^2).$$

We are now interested in providing a formulation in terms of matrix inequalities for obtaining the \mathcal{H}_2 norm of (3) for a given controller $\mathcal{C} \in \mathcal{C}$. By considering the last assertion of Theorem 1, we have that

$$\|\mathcal{G}_c\|_2^2 = \inf_{P > 0, W_{il} > 0, \bar{M}_{il} > 0, \gamma} \gamma^2 \quad (17)$$

subject to

$$\sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} \mu_i \alpha_{il} \text{Tr}(W_{il}) < \gamma^2, \quad (18)$$

$$\begin{bmatrix} W_{il} & \bullet \\ J_{il} & \mathcal{E}_i(P)^{-1} \end{bmatrix} > 0, \quad (19)$$

$$P_i - \sum_{l \in \mathbb{M}_i} \alpha_{il} \bar{M}_{il} > 0, \quad (20)$$

$$\begin{bmatrix} \bar{M}_{il} & \bullet & \bullet \\ A_{il} & \mathcal{E}_i(P)^{-1} & \bullet \\ C_{il} & 0 & I_q \end{bmatrix} > 0, \quad (21)$$

for all $i \in \mathbb{N}$, $l \in \mathbb{M}_i$, A_{il} , J_{il} , and C_{il} in (4). Note that $\mathcal{E}_i(P)^{-1}$ exists since $\sum_{j \in \mathbb{N}} p_{ij} = 1$ for all $i \in \mathbb{N}$. Given the previous discussions, we are now able to state the main goal of this work, that is, studying the problem,

$$\inf_{\mathcal{C} \in \mathcal{C}, P, W_{il}, \bar{M}_{il}, \gamma} \{\gamma^2 : (18) - (21)\}. \quad (22)$$

For the mode-dependent case, (22) was solved in [13] via coupled Riccati equations leading to the so-called \mathcal{H}_2 separation principle, and in [23], via the LMI formulation. However, the additional complexity induced by the *partial observation* of the Markov chain renders it hard to solve considering the previous methods. To date, there are only sub-optimal solutions in the literature for the similar problem written as in (22). As we are going to see in the next sections, we provide a type of sub-optimal *separation procedure* in which a state-feedback controller is calculated and used in an intermediary step to obtain the remaining controller matrices.

4 The \mathcal{H}_2 separation procedure for the partial observation case

For the discussion that follows, consider the following inequalities

$$\sum_{i \in \mathbb{N}} \sum_{l \in \mathbb{M}_i} \mu_i \alpha_{il} \text{Tr}(W_{il}) < \gamma^2, \tag{23}$$

$$\begin{bmatrix} W_{il} & \bullet & \bullet \\ \mathcal{E}_i(Y)J_i & \mathcal{E}_i(Y) & \bullet \\ G_l J_i + F_l H_i & 0 & \text{Her}(G_l) + \mathcal{E}_i(Y) - \mathcal{E}_i(X) \end{bmatrix} > 0, \tag{24}$$

$$\begin{bmatrix} Y_i & \bullet \\ Y_i & X_i \end{bmatrix} > \sum_{l \in \mathbb{M}_i} \alpha_{il} \begin{bmatrix} M_{il} & \bullet \\ N_{il} & S_{il} \end{bmatrix}, \tag{25}$$

$$\begin{bmatrix} M_{il} & \bullet & \bullet & \bullet & \bullet \\ N_{il} & S_{il} & \bullet & \bullet & \bullet \\ \mathcal{E}_i(Y)(A_i + B_i K_l) & \mathcal{E}_i(Y)A_i & \mathcal{E}_i(Y) & \bullet & \bullet \\ G_l(A_i + B_i K_l) + F_l L_i + R_l & G_l A_i + F_l L_i & 0 & \text{Her}(G_l) + \mathcal{E}_i(Y) - \mathcal{E}_i(X) & \bullet \\ C_i + D_i K_l & C_i & 0 & 0 & I \end{bmatrix} > 0, \tag{26}$$

for all $i \in \mathbb{N}$, $l \in \mathbb{M}_i$. We define the set of variables of (23)-(26) as

$$\xi \triangleq \{W_{il}, M_{il}, N_{il}, S_{il}, Y_i, X_i, G_l, K_l, F_l, R_l, i \in \mathbb{N}, l \in \mathbb{M}_i\} \cup \phi,$$

where $\phi = \emptyset$ if γ is given and $\phi = \gamma_a, \gamma_a \triangleq \gamma^2$ if it is a variable. The set of all solutions of (23)-(26) is represented by

$$\Xi \triangleq \{\xi : (23) - (26) \text{ holds}\}.$$

We point out that Ξ is bilinear with respect to G_l and K_l , and $\mathcal{E}_i(Y)$ and K_l , a characteristic we will exploit for tackling our problem.

Theorem 2. Consider the following statements:

- (i) There exists ξ such that $\xi \in \Xi$.
- (ii) There exists a controller \mathcal{C} such that $\mathcal{C} \in \mathfrak{C}$ and $\|\mathcal{G}_c\|_2 < \gamma$.

We have that (i) \implies (ii) by setting $A_{cl} = -G_l^{-1}R_l$, $B_{cl} = -G_l^{-1}F_l$, and $C_{cl} = K_l$ for all $l \in \mathbb{M}$. Moreover (ii) \implies (i) whenever $\hat{\theta} = \theta$. ■

Proof. The result is based on the congruence transformations laid out in [23] and [24], and used for the hidden MJLS stability problem in [43]. Consider the following partitions for P_i in (18)-(21), along with its inverse

$$P_i = \begin{bmatrix} X_i & \bullet \\ U'_i & \hat{X}_i \end{bmatrix}, P_i^{-1} = \begin{bmatrix} Y_i^{-1} & \bullet \\ V'_i & \hat{Y}_i \end{bmatrix}. \tag{27}$$

for $X_i, Y_i^{-1} \in \mathbb{B}(\mathbb{R}^n)$. Note that since $P_i > 0$ for all $i \in \mathbb{N}$, it is possible to define P_i^{-1} . Besides, X_i, \hat{X}_i, Y_i^{-1} , and \hat{Y}_i are positive-definite blocks. Furthermore,

$$\mathcal{E}_i(P) = \begin{bmatrix} \mathcal{E}_i(X) & \bullet \\ \mathcal{E}_i(U)' & \mathcal{E}_i(\hat{X}) \end{bmatrix}, \mathcal{E}_i(P)^{-1} = \begin{bmatrix} R_{1i} & \bullet \\ R'_{2i} & R_{3i} \end{bmatrix}. \tag{28}$$

By defining

$$\mathcal{T}_i \triangleq \begin{bmatrix} I & I \\ V'_i Y_i & 0 \end{bmatrix}$$

we get that

$$\mathcal{T}'_i P_i \mathcal{T}_i = \begin{bmatrix} Y_i & Y_i \\ Y_i & X_i \end{bmatrix}$$

and, similarly, by defining

$$\mathcal{H}_i \triangleq \begin{bmatrix} I & \mathcal{E}_i(X) \\ 0 & \mathcal{E}_i(U)'\end{bmatrix}$$

we get that

$$\mathcal{H}'_i \mathcal{E}_i(P)^{-1} \mathcal{H}_i = \begin{bmatrix} R_{1i} & I \\ I & \mathcal{E}_i(X) \end{bmatrix}.$$

(i) \implies (ii). Given (i), through Lemma 2 (see the Appendix), we get that $G_l \mathcal{E}_i(X - Y)^{-1} G'_l \geq \text{Her}(G_l) + \mathcal{E}_i(Y) - \mathcal{E}_i(X)$, that allows us to infer from (24) and (26) that

$$\begin{bmatrix} W_{il} & \bullet & \bullet \\ \mathcal{E}_i(Y) J_i & \mathcal{E}_i(Y) & \bullet \\ G_l J_i + F_l H_i & 0 & G_l \mathcal{E}_i(X - Y)^{-1} G'_l \end{bmatrix} > 0, \tag{29}$$

$$\begin{bmatrix} M_{il} & \bullet & \bullet & \bullet & \bullet \\ N_{il} & S_{il} & \bullet & \bullet & \bullet \\ \mathcal{E}_i(Y)(A_i + B_i K_l) & \mathcal{E}_i(Y) A_i & \mathcal{E}_i(Y) & \bullet & \bullet \\ G_l(A_i + B_i K_l) + F_l L_i + R_l & G_l A_i + F_l L_i & 0 & G_l \mathcal{E}_i(X - Y)^{-1} G'_l & \bullet \\ C_i + D_i K_l & C_i & 0 & 0 & I \end{bmatrix} > 0, \tag{30}$$

also holds for all $i \in \mathbb{N}$, $l \in \mathbb{M}_i$. By setting $U_i = -\hat{X}_i$, we get that U_i is a symmetric, negative-definite matrix. Besides, considering that $P_i P_i^{-1} = I_{2n}$ for all $i \in \mathbb{N}$, we get that $V_i = Y_i^{-1}$, $Y_i = X_i + U_i$, and $R_{1i} = [\mathcal{E}_i(X) - \mathcal{E}_i(U) \mathcal{E}_i(\hat{X})^{-1} \mathcal{E}_i(U)']^{-1} = \mathcal{E}_i(X + U)^{-1} = \mathcal{E}_i(Y)^{-1}$. By defining

$$\mathcal{D}_{il} \triangleq \begin{bmatrix} \mathcal{E}_i(Y)^{-1} & I \\ 0 & G_l^{-T} \mathcal{E}_i(X - Y) \end{bmatrix}, \tag{31}$$

setting $R_l = -G_l A_{cl}$, $F_l = -G_l B_{cl}$, $K_l = C_{cl}$, and applying the congruence transformations **diag**(I_r, \mathcal{D}_{il}) and **diag**($I_{2n}, \mathcal{D}_{il}, I_q$) to the previous inequalities, we can rewrite (25), and (29)-(30) as follows

$$\begin{aligned} & \begin{bmatrix} W_{il} & \bullet \\ \mathcal{H}'_i J_{il} & \mathcal{H}'_i \mathcal{E}'_i(P)^{-1} \mathcal{H}_i \end{bmatrix} > 0, \\ & \mathcal{J}'_i P_i \mathcal{J}_i - \sum_{l \in \mathbb{M}_i} \alpha_{il} \mathcal{J}'_i \bar{M}_{il} \mathcal{J}_i > 0, \\ & \begin{bmatrix} \mathcal{J}'_i \bar{M}_{il} \mathcal{J}_i & \bullet & \bullet \\ \mathcal{H}'_i A_{il} \mathcal{J}_i & \mathcal{H}'_i \mathcal{E}'_i(P)^{-1} \mathcal{H}_i & \bullet \\ C_{il} \mathcal{J}_i & 0 & I_q \end{bmatrix} > 0. \end{aligned}$$

By applying the congruence transformations $\mathbf{diag}(I_r, \mathcal{H}_i^{-1})$, \mathcal{J}_i^{-1} , and $\mathbf{diag}(\mathcal{J}_i^{-1}, \mathcal{H}_i^{-1}, I_q)$ respectively to the last inequalities, we get that (18)-(21) holds, that implies (ii).

(ii) \implies (i). Given that $\hat{\theta}(k) = \theta(k)$ for all k , we have that $\alpha_{ii} = 1$ for all $i \in \mathbb{N}$. In this case, given that (ii) holds, it is clear that by performing inversely the last steps of the sufficiency proof, and also considering that $\mathcal{E}_i(Y)^{-1} \geq R_{li}$ for all $i \in \mathbb{N}$ (see Lemma 3 in the Appendix), we get that (19) and (21) yield to

$$\begin{bmatrix} W_{ii} & \bullet & \bullet \\ J_i & \mathcal{E}_i(Y)^{-1} & \bullet \\ \mathcal{E}_i(X)J_i + F_i H_i & I & \mathcal{E}_i(X) \end{bmatrix} > 0, \tag{32}$$

$$\begin{bmatrix} M_{ii} & \bullet & \bullet & \bullet & \bullet \\ N_{ii} & S_{ii} & \bullet & \bullet & \bullet \\ A_i(K_i) & A_i & \mathcal{E}_i(Y)^{-1} & \bullet & \bullet \\ \mathcal{E}_i(X)A_i(K_i) + F_i L_i + R_i & \mathcal{E}_i(X)A_i + F_i L_i & I & \mathcal{E}_i(X) & \bullet \\ C_i + D_i K_i & C_i & 0 & 0 & I \end{bmatrix} > 0, \tag{33}$$

where $A_i(K_i) \triangleq A_i + B_i K_i$, $R_i \triangleq \mathcal{E}_i(U)A_{ci}V'_i Y_i$, $F_i \triangleq \mathcal{E}_i(U)B_{ci}$, and $K_i \triangleq C_{ci}V'_i Y_i$. By defining

$$\bar{D}_{ii} \triangleq \begin{bmatrix} \mathcal{E}_i(Y) & -\mathcal{E}_i(Y) \\ 0 & I \end{bmatrix},$$

and applying the congruence transformations $\mathbf{diag}(I_r, \bar{D}_{ii})$ and $\mathbf{diag}(I_{2n}, \bar{D}_{ii}, I_q)$ to (32) and (33) respectively yields to (24) and (26) with the particular choice $G_i = \mathcal{E}_i(X - Y)$, and thus the claim follows. \square

The result in Theorem 2 presents a set of sufficient bilinear conditions for obtaining the controller, that becomes necessary for the mode-dependent case. The bilinearity suggests a strategy that will become clear from the next lemma on.

Lemma 1. *If $\xi \in \Xi$, then $K \in \mathfrak{K}$.* ■

Proof. It readily follows by considering (25) and the $3n \times 3n$ block of (26) in order to get that $Y - \mathcal{L}(Y) > 0$ holds, for $\mathcal{A}_{il} = A_i + B_i K_l$, $i \in \mathbb{N}, l \in \mathbb{M}_i$. \square

We define a new variable set for (23)-(26),

$$\bar{\xi} \triangleq \{W_{il}, M_{il}, N_{il}, S_{il}, Y_i, X_i, G_l, F_l, R_l, i \in \mathbb{N}, l \in \mathbb{M}_i\} \cup \phi,$$

such that $\xi = \bar{\xi} \cup \{K_l \in \mathbb{M}\}$, where $\phi = \emptyset$ or $\phi = \gamma_a$, $\gamma_a = \gamma^2$. The set of solutions for a given $K \in \mathfrak{K}$ is defined in the following.

$$\bar{\Xi}(K) \triangleq \{ \bar{\xi} : (23) - (26) \text{ hold} \} .$$

Corollary 1. For a given $K \in \mathbb{K}$ such that $\bar{\Xi}(K) \neq \emptyset$, if $\bar{\xi} \in \bar{\Xi}(K)$, then by setting $A_{cl} = -G_l^{-1}R_l$, $B_{cl} = -G_l^{-1}F_l$, and $C_{cl} = K_l$ for all $l \in \mathbb{M}$ we have that $\mathcal{C} \in \mathfrak{C}$ and $\|\mathcal{G}_c\|_2 < \gamma$. ■

Note that by fixing K , the inequality (25) becomes an LMI in $\bar{\xi}$, and then (23)-(26) can be solved by standard solvers such as SeDuMi, see, for instance, [49]. Then, the \mathcal{H}_2 ad hoc separation procedure for MJLS with partial information on $\theta(k)$ consists in finding a stabilizing state-feedback gain obtained, for instance, through [11, 42, 51], and calculating the remaining controller matrices by Corollary 1. In this case, we can find the \mathcal{H}_2 dynamic output feedback controller that leads to the best upper bound as follows,

$$\inf_{\bar{\xi} \in \bar{\Xi}(K)} \{ \gamma_a; \text{ such that (23) - (26) hold} \} , \tag{34}$$

where $\gamma_a = \gamma^2$. The Algorithm 1 is shown below. Considering the sub-optimal char-

Algorithm 1 The \mathcal{H}_2 ad hoc separation procedure (Corollary 1)

- 1: Calculate a stochastic stabilizing state-feedback gain $K \in \mathfrak{K}$ such that $\bar{\Xi}(K) \neq \emptyset$;
 - 2: Use K_l as an input in (23)-(26) and calculate A_{cl} and B_{cl} for all $l \in \mathbb{M}$. The final controller is given by (A_{cl}, B_{cl}, C_{cl}) as shown in Corollary 1.
-

acteristics of (34), it is a fact that those conditions do not parametrize all the controllers in the form (2) for the *partial observation case*. Besides, even if $\bar{\Xi} \neq \emptyset$, we have no guarantee that $\bar{\Xi}(K) \neq \emptyset$ for our choice of K . In the case in which the conditions (23)-(26) do not hold for a given K , a new stabilizing state-feedback gain must be provided. As for the case where $\bar{\Xi} = \emptyset$, there are no similar results in the literature to date that would provide an alternative to the design of \mathcal{C} in the hidden MJLS formulation, and thus further study is required.

Remark 1. In the case of perfect observation of the Markov chain, the conditions (23)-(26) are still in the BMI formulation, since we still have to choose K . Due to the equivalence stated in Theorem 2, it is clear that the optimal \mathcal{H}_2 dynamic controller can be obtained by a suitable choice of K . What we observe in the numerical simulations is that, by choosing the optimal \mathcal{H}_2 state-feedback controller, we retrieve the optimal \mathcal{H}_2 dynamic controller, in agreement with the separation principle results of [13] (see the Appendix).

5 Illustrative Examples

In this section, we present two illustrative examples. The first one consists of a simple unstable MJLS that we want to stabilize. The second one is the unstable lateral dynamics of an unmanned aircraft subject to actuator failures.

Example 1. The system matrices in this example are taken as follows

$$A_1 = \begin{bmatrix} 1.4 & 0.1 \\ 0 & 0.5 \end{bmatrix}, A_2 = \begin{bmatrix} 0.9 & 0 \\ 0.3 & 1.2 \end{bmatrix}, B_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, B_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and

$$J_1 = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.4 & 0 \end{bmatrix}, J_2 = \begin{bmatrix} 1.0 & 0 & 0 \\ 0 & 0.8 & 0 \end{bmatrix}, C_1 = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}, C_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

as well as $D_1 = D_2 = [0 \ 1]'$. The transition probability matrix and initial distribution are given by

$$\mathbb{P} = \begin{bmatrix} 0.9 & 0.1 \\ 0.8 & 0.2 \end{bmatrix}, \mu' = \begin{bmatrix} 0.8889 \\ 0.1111 \end{bmatrix},$$

thus, we have that $\mu = \mu\mathbb{P}$ and $r_\sigma(\mathcal{L}) = 1.8460$, that is, an unstable MJLS. For the measurement equation, we set

$$L_1 = [1 \ 0], L_2 = [0 \ 1], H_1 = H_2 = [0 \ 0 \ 1].$$

The conditional probability matrix is assumed to have the following structure

$$r = \begin{bmatrix} \rho & 1-\rho \\ 1-\rho & \rho \end{bmatrix} \quad (35)$$

where $\text{Prob}(\hat{\theta}(k) = i \mid \theta(k) = i) = \rho$ for all $i \in \mathbb{N}$. We initially take $\rho = 1.0$, that is, we assume that we can perfectly measure $\theta(k)$, the so-called mode-dependent case. In this case, we calculate the optimal \mathcal{H}_2 state-feedback controller through the conditions in [42] and obtain

$$\begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = \begin{bmatrix} -0.7405 & 0.0667 \\ 0 & 0 \end{bmatrix} \quad (36)$$

with $r_\sigma(\mathcal{L}) = 0.3983$ and the optimal cost given by $\|\mathcal{G}_K^*\|_2 = 0.9870$. Note that we can calculate the same controller with the results of [13], where the solution of the Control CARE in (40) is given by

$$P_1 = \begin{bmatrix} 2.0956 & -1.3341 \\ -1.3341 & 1.1814 \end{bmatrix}, P_2 = \begin{bmatrix} 2.1577 & -0.8607 \\ -0.8607 & 1.9115 \end{bmatrix}.$$

We have that

$$Y_1 = P_1, Y_2 = P_2,$$

that is, matrices Y_i in (23)-(26) converge to the optimal solution of (40) in this example. By using K in (36) as an input to (34), we get the remaining controllers matrices

$$\left[\begin{array}{cc|c} A_{c1} & B_{c1} & \\ \hline A_{c2} & B_{c2} & \end{array} \right] = \left[\begin{array}{cc|c} -0.9326 & 0.2334 & 0.8516 \\ -0.7588 & 0.5667 & 0.0183 \\ \hline 0.9000 & -0.0878 & 0.0878 \\ 0.3000 & 0.7928 & 0.4072 \end{array} \right]$$

with $r_\sigma(\mathcal{L}) = 0.4026$ and the optimal cost given by $\|\mathcal{G}_c^*\|_2 = \gamma_c^* = 2.2507$. Conversely, by solving the Filtering CARE in (42), we get

$$S_1 = \begin{bmatrix} 1.3653 & 0.0824 \\ 0.0824 & 0.3143 \end{bmatrix}, S_2 = \begin{bmatrix} 0.1801 & 0.0158 \\ 0.0158 & 0.0511 \end{bmatrix},$$

for

$$F_1(S) = \begin{bmatrix} 0.8516 \\ 0.0183 \end{bmatrix}, F_2(S) = \begin{bmatrix} 0.0878 \\ 0.4072 \end{bmatrix}$$

that are numerically equal to B_{ci} calculated through (34). It is clear also that

$$A_{ci} = A_i + B_i K_i - B_{ci} L_i,$$

for all $i \in \mathbb{N}$, echoing the result in [13]. Finally, by computing

$$(\gamma_F^*)^2 \triangleq \sum_{i \in \mathbb{N}} \text{Tr}[O_i^{1/2} K_i (Y^*) S_i K_i (Y^*)' O_i^{1/2}] = 4.0914$$

where $O_i \triangleq [B_i' \mathcal{E}_i(P) B_i + D_i' D_i]$, we have that $\|\mathcal{G}_c^*\|_2^2 = (\gamma_c^*)^2 = \|\mathcal{G}_K^*\|_2^2 + (\gamma_F^*)^2 = 0.9741 + 4.0914 = 5.0655$, that is, $\|\mathcal{G}_*^*\|_2 = 2.2507$, as expected.

We now set $\rho = 0.7$ in (35) and calculate the \mathcal{H}_2 state-feedback gains with the result in [42] in order to obtain the following state-feedback controller

$$\begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = \begin{bmatrix} -0.6060 & -0.0232 \\ -0.5728 & -0.0072 \end{bmatrix}. \quad (37)$$

Using (37) for solving (34), we get the final dynamic controller, with A_{cl} and B_{cl} given by

$$\left[\begin{array}{cc|c} A_{c1} & B_{c1} & \\ \hline A_{c2} & B_{c2} & \end{array} \right] = \left[\begin{array}{cc|c} -0.6145 & -0.0315 & 0.9231 \\ -0.6741 & 0.6821 & 0.2050 \\ \hline 0.0205 & -0.0281 & 0.5557 \\ -0.4236 & 0.9857 & 0.2079 \end{array} \right]$$

with $r_\sigma(\mathcal{L}) = 0.7658$, $\gamma^* = 3.8978$, and $\|\mathcal{G}_c\|_2 = 2.9660$. It is clear that the introduction of the asynchronous effect renders the quadratic performance worse and adds an expected conservatism on the guaranteed cost γ^* with respect to the actual closed-loop system norm $\|\mathcal{G}_c\|_2$. For illustrating the stochastic behavior of the system driven by a wide-sense white noise sequence, we present the curves in Fig. 2, obtained by means of a Monte Carlo simulation of 4000 rounds. \square

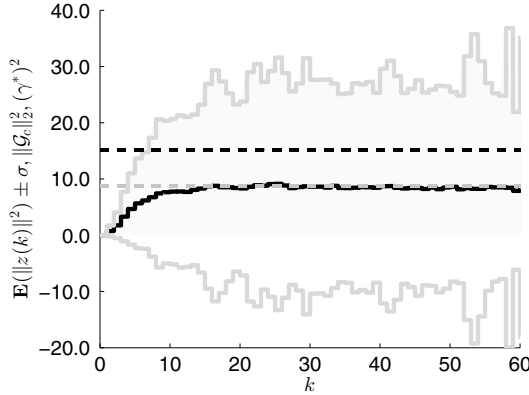


Fig. 2 $E(\|z(k)\|^2) \pm \sigma$ (full black line), $\|\mathcal{G}_c\|_2^2$ (dashed grey line), $(\gamma^*)^2$ (dashed black line).

Example 2. We consider the lateral-directional dynamics of a small unmanned aircraft in steady flight as presented in [20]. The states are variations in the roll rate (Δp), the yaw rate (Δr), the sideslip angle ($\Delta \beta$), and the roll angle ($\Delta \phi$), whereas the control inputs are the aileron ($\Delta \delta_{aileron}$) and the rudder ($\Delta \delta_{rudder}$). The actuators are subject to abrupt faults as modeled by [11]. The Markov chain state space is given by $\mathbb{N} = \{1, 2, 3\}$ and is illustrated by Fig. 1, where $\theta(k) = 1$ is the nominal mode of operation, $\theta(k) = 2$ and $\theta(k) = 3$ are modes with failures in the actuators. The transition probability matrix is given by

$$\mathbb{P} = \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0.2 & 0.7 & 0.1 \\ 0 & 0.9 & 0.1 \end{bmatrix},$$

and the initial probability is taken as $\mu = [0.3103 \ 0.6207 \ 0.0690]$, that corresponds to the stationary distribution. The conditional probability matrix has the following structure:

$$Y = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \rho & 1 - \rho \\ 0 & 1 - \rho & \rho \end{bmatrix}, \tag{38}$$

where $\text{Prob}(\hat{\theta}(k) = i \mid \theta(k) = i) = \rho$, $i \in \{2, 3\}$, that is, the probability of detecting correctly the parameter $\theta(k)$ for modes 2 and 3. Then, in the nominal mode $\theta(k) = 1$, the detector would provide correct estimates of the Markov parameters, whereas for the remaining faulty states, the probability of obtaining the correct mode of operation is ρ . In this case, the discretized (zero-order hold, $T = 0.05$ s) system matrices are given by

$$[A_i | B_1] = \begin{bmatrix} 0.5637 & 0.1133 & -0.6607 & -0.0062 & 2.9735 & -0.0618 \\ 0.0198 & 0.8368 & 1.0512 & 0.0089 & -0.1175 & 0.6414 \\ 0.0033 & -0.0450 & 0.9481 & 0.0159 & 0.0112 & -0.0165 \\ 0.0381 & 0.0073 & -0.0164 & 0.9999 & 0.0812 & -0.0006 \end{bmatrix},$$

for all $i \in \mathbb{N}$. Moreover the actuator failures are modeled by

$$B_2 = B_1 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad B_3 = -B_2,$$

that is, $\theta(k) = 2$ is the mode with an inactive aileron command, and $\theta(k) = 3$ is the mode where the aileron is not active and the rudder command is inverted. The exogenous input matrix is taken as, for all $i \in \mathbb{N}$, $J_i = [I_4 \quad 0_{4 \times 2}]$. We set

$$C_i = \begin{bmatrix} I_4 \\ 0_{2 \times 4} \end{bmatrix}, \quad D_i = \begin{bmatrix} 0_{4 \times 2} \\ I_2 \end{bmatrix},$$

as well as

$$L_i = [0_{2 \times 2} \quad I_2], \quad H_i = [0_{2 \times 4} \quad I_2],$$

for all $i \in \mathbb{N}$, that is, we consider that we can measure the variations on the sideslip and roll angles. We now want to investigate the behavior of γ and $\|\mathcal{G}_c\|_2$ with respect to $\rho \in [0, 1]$. For that, we minimize the \mathcal{H}_2 control conditions of [42] for a given $\rho \in [0, 1]$ and obtain the \mathcal{H}_2 state-feedback controllers used in (34) for calculating the dynamic controller for a given ρ . The result is shown in Fig. 3. In this example, we note the similar type of symmetry of Fig. 3 compared to the filtering control and state-feedback works [41] and [42], and more importantly, the optimal \mathcal{H}_2 dynamic output feedback controller can be obtained, and also a clustered \mathcal{H}_2 dynamic controller. That is, we note two interesting cases in Fig. 3: (1) $\rho = 1$ and $\rho = 0$; (2) $\rho = 0.5$. In the first case, we get the *mode-dependent* formulation, since $\text{Prob}(\hat{\theta}(k) = i \mid \theta(k) = i) = 1$ for $i \in \{2, 3\}$. Interestingly, the case characterized by $\text{Prob}(\hat{\theta}(k) = i \mid \theta(k) = i) = 0$ for $i \in \{2, 3\}$ also leads to the mode-dependent case, since there are only two possible outcomes of $\hat{\theta}(k)$ for $\theta(k) = 2$ and $\theta(k) = 3$. For the case where $\rho = 0.5$, we observe that Assumption 3 holds. In this case, the controller is given by

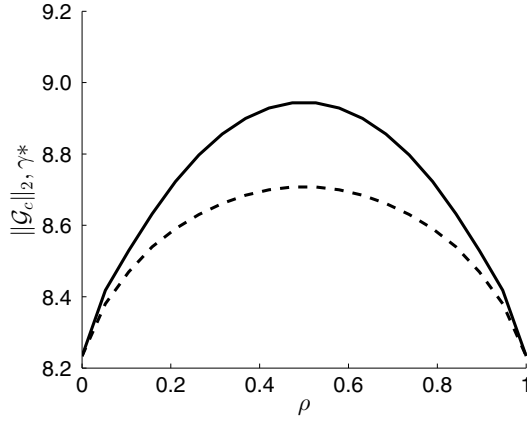


Fig. 3 γ^* (black line) and $\|\mathcal{G}_c\|_2$ (dashed black line) against ρ .

$$\begin{bmatrix} A_{c1} \\ A_{ci} \end{bmatrix} = \frac{\begin{bmatrix} 0.0121 & -0.0251 & 0.4127 & -1.6416 \\ 0.0210 & 0.4315 & 0.1302 & -0.1000 \\ 0.0016 & -0.0338 & 0.4018 & 0.0273 \\ 0.0203 & 0.0084 & -0.0693 & 0.4430 \end{bmatrix}}{\begin{bmatrix} 0.5150 & 0.0141 & -0.5141 & -0.1855 \\ 0.0502 & 0.7511 & 0.3197 & 0.0364 \\ -0.0055 & 0.0471 & 0.5637 & 0.0599 \\ 0.0413 & 0.0096 & -0.0103 & 0.3941 \end{bmatrix}},$$

for $i \in \mathbb{N}^2 \triangleq \{2, 3\}$, along with

$$\begin{bmatrix} B_{c1} | B_{ci} \end{bmatrix} = \left[\begin{array}{cc|cc} -0.4322 & 0.3141 & -0.2348 & 0.1307 \\ 0.7413 & 0.0860 & 0.4892 & -0.0724 \\ 0.5319 & -0.0195 & 0.3270 & -0.0305 \\ -0.0367 & 0.4224 & 0.0044 & 0.6086 \end{array} \right],$$

and

$$\begin{bmatrix} C_{c1} \\ C_{ci} \end{bmatrix} = \frac{\begin{bmatrix} -0.1874 & -0.0581 & 0.1750 & -0.4909 \\ -0.0334 & -0.6378 & -0.3274 & -0.1564 \\ 0 & 0 & 0 & 0 \\ 0.0002 & -0.1878 & -0.5158 & -0.1133 \end{bmatrix}}{\begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}}, \tag{39}$$

for $i \in \mathbb{N}^2$, with $\gamma^* = 8.9450$ and $\|\mathcal{G}_c\|_2 = 8.7084$. We infer that, if Assumption 3 and a clustered state-feedback controller is used as an input to Algorithm 1, we are able to get clustered A_{ci} and B_{ci} as well.

6 Conclusion

In this work, we presented a study of the \mathcal{H}_2 dynamic output feedback control for MJLS in which we cannot perfectly measure the Markov chain. The conditions we derived are given in terms of BMI, that can be recast into the LMI formulation by providing a stabilizing state-feedback controller. The final controller would be composed by the state-feedback controller provided in the first step and the “filter-like” structure calculated in the second step, suggesting a type of \mathcal{H}_2 *ad-hoc separation procedure*. The advantage of this method relies in the following characteristics. We can retrieve the optimal mode-dependent controller in the case we can measure the Markov chain, since our conditions also become necessary in that situation. Furthermore, clusterized and mode-independent controllers can be calculated as a by-product of the observation model we use in our work. On the other hand, the weak point in this approach is the initial guess of the state-feedback controller, as the calculation of the remaining structure is dependent on that choice. For future research, it is desirable to improve the separation procedure and find alternative ways of obtaining the dynamic controller structure.

Acknowledgements This work was supported in part by the São Paulo Research Foundation - FAPESP, grants FAPESP-2015/09912-8 and FAPESP-2017/06358-5 for the first author; and the National Council for Scientific and Technological Development - CNPq, grant CNPq-304091/2014-6, the FAPESP/SHELL Brasil through the Research Center for Gas Innovation, grant FAPESP/SHELL-2014/50279-4, the project INCT, grants FAPESP/INCT - 2014/50851-0 and CNPq/INCT-465755/2014-3, and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 88887.136349/2017-00, for the second author.

Appendix

6.1 Auxiliary results

Lemma 2 ([14],[44]). For $P = P' > 0$, the inequality $GP^{-1}G' \geq \text{Her}(G) - P$ holds for any square matrix G of compatible dimensions. ■

Proof. It readily follows by noting that $(G - P)P^{-1}(G - P)' \geq 0$ holds true for all square matrices G . □

Lemma 3 ([23]). For $\hat{X} \in \mathbb{H}^n$, $\hat{X} > 0$, $\hat{U} \in \mathbb{H}^n$, and $p_{ij} \geq 0$, $\sum_{j \in \mathbb{N}} p_{ij} = 1$, we have that $\mathcal{E}_i(\hat{U}\hat{X}^{-1}\hat{U}') \geq \mathcal{E}_i(\hat{U})\mathcal{E}_i(\hat{X})^{-1}\mathcal{E}_i(\hat{U}')$. ■

Proof. Note that $\hat{U}_j\hat{X}_j^{-1}\hat{U}'_j \geq \hat{U}_j\hat{X}_j^{-1}\hat{U}'_j$ for all $j \in \mathbb{N}$, and thus by taking the Schur complement, we get that

$$\begin{bmatrix} \hat{U}_j\hat{X}_j^{-1}\hat{U}'_j & \hat{U}_j \\ \hat{U}'_j & \hat{X}_j \end{bmatrix} \geq 0.$$

Multiplying the last inequality by p_{ij} , summing everything up with respect to $j \in \mathbb{N}$, and applying again the Schur complement leads to the desired result. \square

6.2 Coupled Algebraic Riccati Equations

Consider that Assumption 2 holds, and also scalar $\mu_i > 0$, $i \in \mathbb{N}$, $\sum_{i \in \mathbb{N}} \mu_i = 1$, $J \triangleq (J_1, \dots, J_N) \in \mathbb{H}^{r,n}$, $L \triangleq (L_1, \dots, L_N) \in \mathbb{H}^{n,p}$, $H \triangleq (H_1, \dots, H_N) \in \mathbb{H}^{r,p}$, $C \triangleq (C_1, \dots, C_N) \in \mathbb{H}^{n,q}$, and $D \triangleq (D_1, \dots, D_N) \in \mathbb{H}^{m,q}$, and assume that $C_i' D_i = 0$, $H_i J_i' = 0$, $D_i' D_i > 0$, and $H_i H_i' > 0$ for all $i \in \mathbb{N}$. We introduce two sets of discrete-time CARE that were presented in [13] for the study of the \mathcal{H}_2 separation principle for MJLS.

Definition 4 (Control CARE). We say that $P \triangleq (P_1, \dots, P_N) \in \mathbb{H}^{n+}$ is the stochastic stabilizing solution of

$$P_i = A_i' \mathcal{E}_i(P) A_i + C_i' C_i - A_i' \mathcal{E}_i(P) B_i [B_i' \mathcal{E}_i(P) B_i + D_i' D_i]^{-1} B_i' \mathcal{E}_i(P) A_i, \quad (40)$$

for all $i \in \mathbb{N}$ if (9) (or (10)) holds for $\mathcal{A}_{ii} = A_i + B_i K_i(P)$, where $K_i(P)$ is given by

$$K_i(P) \triangleq -[B_i' \mathcal{E}_i(P) B_i + D_i' D_i]^{-1} B_i' \mathcal{E}_i(P) A_i, \quad (41)$$

for all $i \in \mathbb{N}$. ■

Definition 5 (Filtering CARE). We say that $S \triangleq (S_1, \dots, S_N) \in \mathbb{H}^{n+}$ is the stochastic stabilizing solution of

$$S_j = \sum_{i \in \mathbb{N}} p_{ij} \{A_i S_i A_i' + \mu_i J_i J_i' - A_i S_i L_i' [L_i S_i L_i' + \mu_i H_i H_i']^{-1} L_i S_i A_i'\}, \quad (42)$$

for all $i \in \mathbb{N}$ if (9) (or (10)) holds for $\mathcal{A}_{ii} = A_i - F_i(S) L_i$, where $F_i(S)$ is given by

$$F_i(S) \triangleq A_i S_i L_i' (L_i S_i L_i' + \mu_i H_i H_i')^{-1}, \quad (43)$$

for all $i \in \mathbb{N}$. ■

Conditions for the existence of stabilizing solutions to (40) and (42) are discussed in [7] and rely in the concepts of stabilizability and detectability for MJLS.

References

1. Aberkane, S., Ponsart, J.C., Sauter, D.: Multiobjective output feedback control of a class of stochastic hybrid systems with state-dependent noise. *Mathematical Problems in Engineering* **2007**(31561), 26 (2007). DOI <https://doi.org/10.1155/2007/31561>
2. Aberkane, S., Sauter, D., Ponsart, J.C.: Output feedback robust H_∞ control of uncertain active fault tolerant control systems via convex analysis. *International Journal of Control* **81**(2), 252–263 (2008). DOI [10.1080/00207170701535959](https://doi.org/10.1080/00207170701535959)

3. Athans, M.: The role and use of the stochastic linear-quadratic-Gaussian problem in control system design. *IEEE Transactions on Automatic Control* **16**(6), 529–552 (1971). DOI 10.1109/TAC.1971.1099818
4. Boukas, E.K.: *Stochastic Switching Systems: Analysis and Design*. Birkhäuser Basel, New York, NY, USA (2006). DOI 10.1007/0-8176-4452-0
5. Boyd, S., Ghaoui, L.E., Feron, E., Balakrishnan, V.: *Linear Matrix Inequalities in System and Control Theory*. SIAM Studies in Applied and Numerical Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (1994). DOI 10.1137/1.9781611970777
6. Chizeck, H.J., Ji, Y.: Optimal quadratic control of jump linear systems with Gaussian noise in discrete-time. In: *Proceedings of the 27th IEEE Conference on Decision and Control*, vol. 3, pp. 1989–1993. IEEE, New York City, NY, USA (1988). DOI 10.1109/CDC.1988.194681
7. Costa, O.L.V.: Discrete-time coupled Riccati equations for systems with Markov switching parameters. *Journal of Mathematical Analysis and Applications* **194**(1), 197–216 (1995). DOI <https://doi.org/10.1006/jmaa.1995.1294>
8. Costa, O.L.V., Fragoso, M.D.: Discrete-time LQ-optimal control problems for infinite Markov jump parameter systems. *IEEE Transactions on Automatic Control* **40**(12), 2076–2088 (1995). DOI 10.1109/9.478328
9. Costa, O.L.V., Fragoso, M.D., Marques, R.P.: *Discrete-Time Markov Jump Linear Systems*. Springer-Verlag London, New York City, NY, USA (2005). DOI 10.1007/b138575
10. Costa, O.L.V., Fragoso, M.D., Todorov, M.G.: *Continuous-Time Markov Jump Linear Systems*. Springer-Verlag Berlin Heidelberg, New York, NY, USA (2013). DOI 10.1007/978-3-642-34100-7
11. Costa, O.L.V., Fragoso, M.D., Todorov, M.G.: A detector-based approach for the H_2 control of Markov jump linear systems with partial information. *IEEE Transactions on Automatic Control* **60**(5), 1219–1234 (2015). DOI 10.1109/TAC.2014.2366253
12. Costa, O.L.V., Marques, R.P.: Mixed H_2/H_∞ -control of discrete-time Markovian jump linear systems. *IEEE Transactions on Automatic Control* **43**(1), 95–100 (1998). DOI 10.1109/9.654895
13. Costa, O.L.V., Tuesta, E.F.: H_2 -control and the separation principle for discrete-time Markovian jump linear systems. *Mathematics of Control, Signals and Systems* **16**(4), 320–350 (2004). DOI 10.1007/s00498-004-0142-3
14. Daafouz, J., Bernussou, J.: Parameter dependent Lyapunov functions for discrete time systems with time varying parametric uncertainties. *Systems & Control Letters* **43**(5), 355–359 (2001). DOI [https://doi.org/10.1016/S0167-6911\(01\)00118-9](https://doi.org/10.1016/S0167-6911(01)00118-9)
15. Davis, M.H.A., Vinter, R.B.: *Stochastic Modelling and Control*. Springer Netherlands, New York City, NY, USA (1985). DOI 10.1007/978-94-009-4828-0
16. Doyle, J.: Guaranteed margins for LQG regulators. *IEEE Transactions on Automatic Control* **23**(4), 756–757 (1978). DOI 10.1109/TAC.1978.1101812
17. Doyle, J., Stein, G.: Multivariable feedback design: Concepts for a classical/modern synthesis. *IEEE Transactions on Automatic Control* **26**(1), 4–16 (1981). DOI 10.1109/TAC.1981.1102555
18. Doyle, J.C., Glover, K., Khargonekar, P.P., Francis, B.A.: State-space solutions to standard H_2 and H_∞ control problems. *IEEE Transactions on Automatic Control* **34**(8), 831–847 (1989). DOI 10.1109/9.29425
19. Dragan, V., Morozan, T., Stoica, A.M.: *Mathematical Methods in Robust Control of Linear Stochastic Systems*. Springer, New York, NY, USA (2013). DOI 10.1007/978-1-4614-8663-3
20. Ducard, G.J.J.: *Fault-tolerant Flight Control and Guidance Systems*. Springer-Verlag London, New York City, NY, USA (2009). DOI 10.1007/978-1-84882-561-1
21. Fioravanti, A.R., Gonçalves, A.P.C., Geromel, J.C.: Discrete-time H_∞ output feedback for Markov jump systems with uncertain transition probabilities. *International Journal of Robust and Nonlinear Control* **23**(8), 894–902 (2013). DOI 10.1002/rnc.2807
22. Georgiou, T.T., Lindquist, A.: The separation principle in stochastic control, redux. *IEEE Transactions on Automatic Control* **58**(10), 2481–2494 (2013). DOI 10.1109/TAC.2013.2259207

23. Geromel, J.C., Gonçalves, A.P.C., Fioravanti, A.R.: Dynamic output feedback control of discrete-time Markov jump linear systems through linear matrix inequalities. *SIAM Journal on Control and Optimization* **48**(2), 573–593 (2009). DOI 10.1137/080715494
24. Gonçalves, A.P., Fioravanti, A.R., Geromel, J.C.: Markov jump linear systems and filtering through network transmitted measurements. *Signal Processing* **90**(10), 2842–2850 (2010). DOI <http://dx.doi.org/10.1016/j.sigpro.2010.04.007>
25. Gunckel, T.L., Franklin, G.F.: A general solution for linear, sampled-data control. *ASME. Journal of Basic Engineering* **85**(2), 197–201 (1963). DOI 10.1115/1.3656559
26. Ji, Y., Chizeck, H.J.: Jump linear quadratic Gaussian control: Steady-state solution and testable conditions. *Control, Theory and Advanced Technology* **6**, 289–319 (1990)
27. Joseph, D.P., Tou, T.J.: On linear control theory. *Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry* **80**(4), 193–196 (1961). DOI 10.1109/TAI.1961.6371743
28. K. Zhou, J.C.D., Glover, K.: *Robust and Optimal Control*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1996)
29. Kalman, R.E.: Contributions to the theory of optimal control. *Boletín de la Sociedad Matemática Mexicana* **5**, 102–149 (1960)
30. Kalman, R.E.: A new approach to linear filtering and prediction problems. *ASME. Journal of Basic Engineering* **85**, 35–45 (1960)
31. Kalman, R.E., Bertram, J.E.: Control system analysis and design via the “second method” of Lyapunov: I continuous-time systems. *ASME. Journal of Basic Engineering* **82**(2), 371–393 (1960). DOI 10.1115/1.3662604
32. Kalman, R.E., Bertram, J.E.: Control system analysis and design via the “second method” of Lyapunov: II discrete-time systems. *ASME. Journal of Basic Engineering* **82**(2), 394–400 (1960). DOI 10.1115/1.3662605
33. Liu, X., Ma, G., Pagilla, P.R., Ge, S.S.: Dynamic output feedback asynchronous control of networked Markovian jump systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* pp. 1–11 (2018). DOI 10.1109/TSMC.2018.2827166
34. Mahmoud, M., Jiang, J., Zhang, Y.: *Active Fault Tolerant Control Systems - Stochastic Analysis and Synthesis*. Springer, Germany (2003)
35. Mariton, M.: Detection delays, false alarm rates and the reconfiguration of control systems. *International Journal of Control* **49**, 981–992 (1989). DOI 10.1080/00207178908559680
36. Mariton, M.: *Jump Linear Systems in Automatic Control*. CRC Press, New York, NY, USA (1990)
37. Mehdi, D., Boukas, E.K., Bachelier, O.: Static output feedback design for uncertain linear discrete time systems. *IMA Journal of Mathematical Control and Information* **21**(1), 1–13 (2004). DOI 10.1093/imamci/21.1.1
38. Morais, C.F., Braga, M.F., Oliveira, R.C.L.F., Peres, P.L.D.: An LMI approach for H_2 dynamic output feedback control of discrete-time Markov systems with uncertain probabilities. In: 2016 IEEE Conference on Computer Aided Control System Design (CACSD), pp. 1–6. IEEE, Buenos Aires, Argentina (2016). DOI 10.1109/CACSD.2016.7602559
39. Morais, C.F., Braga, M.F., Oliveira, R.C.L.F., Peres, P.L.D.: LMI-based design of H_∞ dynamic output feedback controllers for mjls with uncertain transition probabilities. In: 2016 American Control Conference (ACC), pp. 5650–5655. IEEE, Boston, MA, USA (2016). DOI 10.1109/ACC.2016.7526556
40. Morais, C.F., Braga, M.F., Oliveira, R.C.L.F., Peres, P.L.D.: Reduced-order dynamic output feedback control of uncertain discrete-time Markov jump linear systems. *International Journal of Control* **90**(11), 2368–2383 (2017). DOI 10.1080/00207179.2016.1245871
41. de Oliveira, A.M., Costa, O.L.V.: H_2 -filtering for discrete-time hidden Markov jump systems. *International Journal of Control* **90**(3), 599–615 (2017). DOI <https://doi.org/10.1080/00207179.2016.1186844>
42. de Oliveira, A.M., Costa, O.L.V.: Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control of hidden Markov jump systems. *International Journal of Robust and Nonlinear Control* **28**(4), 1261–1280 (2018). DOI 10.1002/rnc.3952

43. de Oliveira, A.M., Costa, O.L.V., Daafouz, J.: Design of stabilizing dynamic output feedback controllers for hidden Markov jump linear systems. *IEEE Control Systems Letters* **2**(2), 278–283 (2018). DOI [10.1109/LCSYS.2018.2829883](https://doi.org/10.1109/LCSYS.2018.2829883)
44. de Oliveira, M.C., Bernussou, J., Geromel, J.C.: A new discrete-time robust stability condition. *Systems & Control Letters* **37**(4), 261–265 (1999). DOI [https://doi.org/10.1016/S0167-6911\(99\)00035-3](https://doi.org/10.1016/S0167-6911(99)00035-3)
45. de Oliveira, M.C., Geromel, J.C., Bernussou, J.: Design of dynamic output feedback decentralized controllers via a separation procedure. *International Journal of Control* **73**(5), 371–381 (2000). DOI [10.1080/002071700219551](https://doi.org/10.1080/002071700219551)
46. Peaucelle, D., Arzelier, D.: An efficient numerical solution for H_2 static output feedback synthesis. In: 2001 European Control Conference (ECC), pp. 3800–3805. IEEE, Porto, Portugal (2001). DOI [10.23919/ECC.2001.7076526](https://doi.org/10.23919/ECC.2001.7076526)
47. Ross, S.M.: *Introduction to Probability Models*, tenth edn. Academic Press, Inc., Orlando, FL, USA (2010)
48. Skogestad, S., Postlethwaite, I.: *Multivariable Feedback Control: Analysis and Design*. John Wiley & Sons, Inc., USA (2005)
49. Sturm, J.F.: Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software* **11**, 625–653 (1999)
50. by Tamer Basar, E.: *Control Theory: Twenty-Five Seminal Papers (1932-1981)*. Wiley-IEEE Press (2001). DOI [10.1109/9780470544334.ch8](https://doi.org/10.1109/9780470544334.ch8)
51. Todorov, M.G., Fragoso, M.D., Costa, O.L.V.: Detector-based H_∞ results for discrete-time Markov jump linear systems with partial observations. *Automatica* **91**, 159–172 (2018). DOI <https://doi.org/10.1016/j.automatica.2018.01.034>
52. do Val, J.B., Geromel, J.C., Gonçalves, A.P.C.: The H_2 -control for jump linear systems: cluster observations of the Markov state. *Automatica* **38**(2) (2002). DOI [http://dx.doi.org/10.1016/S0005-1098\(01\)00210-2](https://doi.org/10.1016/S0005-1098(01)00210-2)
53. Vargas, A.N., Costa, E.F., do Val, J.: *Advances in the Control of Markov Jump Linear Systems with No Mode Observation*. Springer International Publishing, New York City, NY, USA (2016)
54. Wonham, W.H.: On the separation theorem of stochastic control. *SIAM Journal on Control* **6**, 312–326 (1968)
55. Wu, Z., Shi, P., Shu, Z., Su, H., Lu, R.: Passivity-based asynchronous control for Markov jump systems. *IEEE Transactions on Automatic Control* **62**(4), 2020–2025 (2017). DOI [10.1109/TAC.2016.2593742](https://doi.org/10.1109/TAC.2016.2593742)



Time-Inconsistent Optimal Control Problems and Related Issues

Wei Yan and Jiongmin Yong

Abstract Classical stochastic optimal control problems are time-consistent, by which it means that an optimal control selected at a given initial pair remains optimal thereafter, along the optimal pair. When the discount is non-exponential and/or the probability is subjective, the corresponding optimal control problem is time-inconsistent, in general. In this paper, we survey recent results in the area and briefly present some of our on-going works.

1 Introduction

In daily life, people frequently face various decision-making situations. When a decision has to be made by an individual, among all possible choices, the individual selects the one he/she feels the best. However, more than often, he/she will regret the selected decision sometime later. Such a phenomenon is called the *time-inconsistency* of the problem under consideration. It is easy to find examples around us. For instance, when there is a big holiday sale, one might buy some good-looking stuff (to meet the instant satisfaction). After a while, one might realize that these stuff are actually not necessary, might never be used, and only deserve a “garage sale”, therefore regret the decision made previously. Also, in trading stocks, people often regret: I should buy certain stocks long time ago, and/or I should sell certain stocks which later became no-value much earlier, and so on.

To study such kind of problems, we first need to find out what are the main reasons causing the time-inconsistency. Careful investigations showed that ([6])

Wei Yan
University of Central Florida, Orlando, FL 32816, e-mail: wei.yan@ucf.edu

Jiongmin Yong
University of Central Florida, Orlando, FL 32816, e-mail: jiongmin.yong@ucf.edu. This author is partially supported by NSF Grant DMS-1812921.

there are two main reasons causing the time-inconsistency of the problem: (i) people’s time-preferences, and (ii) people’s risk-preferences. Mathematically, time-preferences can be described by discounting (exponential or non-exponential), and risk-preferences can be described by subjective or objective probabilities. Most people over-weight the current/near future utility (level of satisfaction). Buying unnecessary things at big holiday sale is a typical example of such which is exactly due to the time-preferences. The stock buying and selling indicate that different people will have different opinions about the risks involved in the coming events. This is the risk-preference. It is known that if the discounting is exponential and the probability is objective, then the problem is time-consistent. Otherwise, the problem is generally time-inconsistent.

The purpose of this paper is to exhibit mathematical formulations of various time-inconsistent optimal control problems, survey some results obtained by us in the recent years, present some new results, and pose some open problems.

2 Time-Consistent Situations

In this section, let us look at the time-consistency of some typical optimal control problems. In the rest of the paper, we let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a complete filtered probability space on which a d -dimensional standard Brownian motion $W(\cdot)$ is defined, whose natural filtration, augmented by all the \mathbb{P} -null sets, is given by $\mathbb{F} = \{\mathcal{F}_t\}_{t \geq 0}$. We introduce the following spaces. For $p, q \in [1, \infty)$,

$$L^p_{\mathcal{F}_T}(\Omega; \mathbb{R}^n) = \left\{ \xi : \Omega \rightarrow \mathbb{R}^n \mid \xi \text{ is } \mathcal{F}_T\text{-measurable, } \mathbb{E}|\xi|^p < \infty \right\},$$

$$L^p_{\mathcal{F}_T}(\Omega; L^q(t, T; \mathbb{R}^n)) = \left\{ \varphi : [t, T] \times \Omega \rightarrow \mathbb{R}^n \mid \varphi(\cdot) \text{ is } \mathcal{B}[t, T] \otimes \mathcal{F}_T\text{-measurable,} \right.$$

$$\left. \mathbb{E} \left(\int_t^T |\varphi(s)|^q ds \right)^{\frac{p}{q}} < \infty \right\},$$

$$L^p_{\mathbb{F}}(\Omega; L^q(t, T; \mathbb{R}^n)) = \left\{ \varphi(\cdot) \in L^p_{\mathcal{F}_T}(\Omega; L^q(t, T; \mathbb{R}^n)) \mid \right.$$

$$\left. \varphi(\cdot) \text{ is } \mathbb{F}\text{-progressively measurable} \right\},$$

$$L^p_{\mathbb{F}}(\Omega; C([t, T]; \mathbb{R}^n)) = \left\{ \varphi : [t, T] \times \Omega \mid \varphi(\cdot) \text{ is } \mathbb{F}\text{-adapted and continuous,} \right.$$

$$\left. \mathbb{E} \left[\sup_{s \in [t, T]} |\varphi(s)|^p \right] < \infty \right\}.$$

For $p = \infty$ and/or $q = \infty$, we can obviously define the corresponding spaces. We denote

$$L^p_{\mathbb{F}}(\Omega; L^p(0, T; \mathbb{R}^n)) = L^p_{\mathbb{F}}(0, T; \mathbb{R}^n), \quad 1 \leq p \leq \infty.$$

In the case, $n = 1$, we will omit \mathbb{R}^n in the notation, for example, $L^p_{\mathcal{F}_T}(\Omega)$, etc. Next, we introduce the *admissible control set*:

$$\mathcal{U}^p[t, T] = \left\{ u : [t, T] \times \Omega \rightarrow U \mid u(\cdot) \in L^p_{\mathbb{F}}(\Omega; L^2(t, T; \mathbb{R}^m)) \right\}.$$

Now, we consider the following controlled stochastic differential equation (SDE, for short)

$$(1) \quad \begin{cases} dX(s) = b(s, X(s), u(s))ds + \sigma(s, X(s), u(s))dW(s), & s \in [t, T], \\ X(t) = x, \end{cases}$$

where $b : [0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}^n$, $\sigma : [0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}^{n \times d}$ are given (deterministic) maps, with $U \subseteq \mathbb{R}^m$ being a non-empty set, and $(t, x) \in [0, T] \times \mathbb{R}^n$ being called the *initial pair*. In the above, $X(\cdot)$ is called the *state process* and $u(\cdot)$ is called the *control process*. We introduce the following assumption for the state equation (1).

(H1) The map $b : [0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}^n$, $\sigma : [0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}^{n \times d}$ are continuous and there exists a constant $L > 0$ such that

$$(2) \quad \begin{aligned} |b(t, x, u) - b(t, x', u)| + |\sigma(t, x, u) - \sigma(t, x', u)| &\leq L|x - x'|, \\ \forall(t, u) \in [0, T] \times U, x, x' \in \mathbb{R}^n, \\ |b(t, 0, u)| + |\sigma(t, 0, u)| &\leq L(1 + |u|), \quad \forall(t, u) \in [0, T] \times U. \end{aligned}$$

Under (H1), for any $(t, x) \in [0, T] \times \mathbb{R}^n$, and any $u(\cdot) \in \mathcal{U}^p[t, T]$, there exists a unique solution $X(\cdot) = X(\cdot; t, x, u(\cdot)) \in L^p_{\mathbb{F}}(\Omega; C([t, T]; \mathbb{R}^n))$ to the *state equation* (1). Moreover, the following estimate holds:

$$(3) \quad \mathbb{E}_t \left[\sup_{s \in [t, T]} |X(s)|^p \right] \leq K \mathbb{E}_t \left[1 + |x|^p + \left(\int_t^T |u(s)|^2 ds \right)^{\frac{p}{2}} \right].$$

Hereafter, $K > 0$ represents a generic constant which can be different from line to line. To measure the performance of the control, we may introduce the following cost functional:

$$(4) \quad \bar{J}(t, x; u(\cdot)) = \mathbb{E} \left[e^{-\delta(T-t)} h(X(T)) + \int_t^T e^{-\delta(s-t)} g(s, X(s), u(s)) ds \right],$$

for some maps $h : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : [0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}$ and a constant $\delta \geq 0$. We introduce the following assumption.

(H2) The map $g : [0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}$ and $h : \mathbb{R}^n \rightarrow \mathbb{R}$ are continuous and there exist constants $L, q > 0$ such that

$$(5) \quad |g(t, x, u)| + |h(x)| \leq L(1 + |x|^q + |u|^q), \quad (t, x, u) \in [0, T] \times \mathbb{R}^n \times U.$$

It is clear that for any $(t, x) \in [0, T] \times \mathbb{R}^n$, and any $u(\cdot) \in \mathcal{U}^p[t, T]$ with $p \geq q$, the state $X(\cdot) \in L^p_{\mathbb{F}}(\Omega; C([t, T]; \mathbb{R}^n))$. Then by (H2),

$$|g(s, X(s), u(s))| \leq L(1 + |X(s)|^q + |u(s)|^q), \quad |h(X(T))| \leq L(1 + |X(T)|^q).$$

Hence, $\bar{J}(t, x; u(\cdot))$ is well-defined. We refer to the function $t \mapsto e^{-\delta t}$ as the *exponential discounting*. The following is a classical stochastic optimal control problem.

Problem (C̄). For given $(t, x) \in [0, T] \times \mathbb{R}^n$, find a $\bar{u}(\cdot) \in \mathcal{U}^p[t, T]$ such that

$$(6) \quad \bar{J}(t, x; \bar{u}(\cdot)) = \inf_{u(\cdot) \in \mathcal{U}^p[t, T]} \bar{J}(t, x; u(\cdot)) = V(t, x).$$

Any $\bar{u}(\cdot) \in \mathcal{U}^p[t, T]$ satisfying (6) is called an *optimal control* of Problem (C̄), and $(t, x) \mapsto V(t, x)$ is called the *value function* of Problem (C̄). One can show that

$$(7) \quad V(t, x) = \inf_{u(\cdot) \in \mathcal{U}[t, \tau]} \mathbb{E} \left[e^{-\delta(\tau-t)} \left(V(\tau, X(\tau; t, x)) + \int_t^\tau e^{-\delta(s-\tau)} g(s, X(s; t, x, u(\cdot)), u(s)) ds \right) \right],$$

for $(t, x) \in [0, T] \times \mathbb{R}^n$ and $\tau \in [t, T]$. This is called the *Bellman's principle of optimality*. If $\bar{u}(\cdot) \in \mathcal{U}[t, T]$ is an optimal control of Problem (C̄) for the initial pair (t, x) , with $\bar{X}(\cdot)$ being the optimal state process, then for any $\tau \in (t, T)$,

$$\begin{aligned} V(t, x) &= \mathbb{E} \left[e^{-\delta(T-t)} h(\bar{X}(T)) + \int_t^T e^{-\delta(s-t)} g(s, \bar{X}(s), \bar{u}(s)) ds \right] \\ &= \mathbb{E} \left[e^{-\delta(\tau-t)} \left(J(\tau, \bar{X}(\tau); \bar{u}(\cdot)) \Big|_{[t, \tau]} + \int_t^\tau e^{-\delta(s-\tau)} g(s, \bar{X}(s), \bar{u}(s)) ds \right) \right] \\ &\geq \mathbb{E} \left[e^{-\delta(\tau-t)} \left(V(\tau, \bar{X}(\tau)) + \int_t^\tau e^{-\delta(s-\tau)} g(s, \bar{X}(s), \bar{u}(s)) ds \right) \right] \\ &\geq \inf_{u(\cdot) \in \mathcal{U}[t, \tau]} \mathbb{E} \left[e^{-\delta(\tau-t)} \left(V(\tau, X(\tau; t, x)) + \int_t^\tau e^{-\delta(s-\tau)} g(s, X(s; t, x, u(\cdot)), u(s)) ds \right) \right] = V(t, x). \end{aligned}$$

The above leads to

$$(8) \quad \bar{J}(\tau, \bar{X}(\tau); \bar{u}(\cdot)) \Big|_{[t, \tau]} = V(\tau, \bar{X}(\tau)).$$

This means that the restriction $\bar{u}(\cdot) \Big|_{[t, \tau]}$ of the optimal control $\bar{u}(\cdot)$ for the initial pair (t, x) on a later interval $[\tau, T]$ is an optimal control for the initial pair $(\tau, \bar{X}(\tau))$. This is called the *time-consistency* of Problem (C̄).

By the way, for later comparison purpose, we state the following result.

Proposition 1. *Let (H1)–(H2) hold. Then the value function $V(\cdot, \cdot)$ is continuous and is the unique viscosity solution to the following Hamilton-Jacobi-Bellman (HJB, for short) equation:*

$$(9) \quad \begin{cases} V_t(t,x) + \inf_{u \in U} \left\{ \frac{1}{2} \text{tr} [V_{xx}(t,x) \sigma(t,x,u) \sigma(t,x,u)^\top] \right. \\ \quad \left. + V_x(t,x)b(t,x,u) + g(t,x,u) \right\} - \delta V(t,x) = 0, & (t,x) \in [0,T] \times \mathbb{R}^n, \\ V(T,x) = h(x), & x \in \mathbb{R}^n. \end{cases}$$

On the other hand, we may introduce the following cost functional:

$$(10) \quad J(t,x;u(\cdot)) = \mathbb{E}_t \left[e^{-\delta(T-t)} h(X(T)) + \int_t^T e^{-\delta(s-t)} g(s,X(s),u(s)) ds \right],$$

where $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_t]$. Note that $J(t,x;u(\cdot))$ is an \mathcal{F}_t -measurable random variable, satisfying

$$(11) \quad \mathbb{E}[J(t,x;u(\cdot))] = \bar{J}(t,x;u(\cdot)).$$

We may formulate an optimal control problem similar to Problem (\bar{C}) , replacing $\bar{J}(t,x;u(\cdot))$ by $J(t,x;u(\cdot))$. For convenience, such a problem is called Problem (C). We have the following simple result.

Proposition 2. *Let (H1) – (H2) hold. Then Problems (C) and (\bar{C}) are equivalent in the sense that for any $(t,x) \in [0,T] \times \mathbb{R}^n$, a $\bar{u}(\cdot) \in \mathcal{U}^p[t,T]$ is optimal for one of the problems, if and only if it is optimal for the other.*

Proof. If $\bar{u}(\cdot) \in \mathcal{U}[t,T]$ is optimal at (t,x) for Problem (C), i.e.,

$$J(t,x;\bar{u}(\cdot)) \leq J(t,x;u(\cdot)), \quad \forall u(\cdot) \in \mathcal{U}^p[t,T], \text{ a.s.},$$

then $\bar{u}(\cdot)$ is optimal for Problem (\bar{C}) . Conversely, suppose $\bar{u}(\cdot) \in \mathcal{U}^p[t,T]$ is optimal at (t,x) for Problem (\bar{C}) , but it is not optimal for Problem (C). Then, there exists a $\hat{u}(\cdot) \in \mathcal{U}^p[t,T]$ such that

$$(12) \quad \mathbb{P}(\hat{\Omega}) > 0, \quad \hat{\Omega} = \left(J(t,x;\hat{u}(\cdot)) < J(t,x;\bar{u}(\cdot)) \right) \in \mathcal{F}_t.$$

Then set

$$\tilde{u}(\cdot) = \bar{u}(\cdot) \mathbf{1}_{\Omega \setminus \hat{\Omega}} + \hat{u}(\cdot) \mathbf{1}_{\hat{\Omega}} \in \mathcal{U}^p[t,T].$$

It is clear that

$$(13) \quad \begin{aligned} X(\cdot; t,x,\tilde{u}(\cdot)) &= X(\cdot; t,x,\bar{u}(\cdot)) \mathbf{1}_{\Omega \setminus \hat{\Omega}} + X(\cdot; t,x,\hat{u}(\cdot)) \mathbf{1}_{\hat{\Omega}}, \\ J(t,x,\tilde{u}(\cdot)) &= J(t,x,\bar{u}(\cdot)) \mathbf{1}_{\Omega \setminus \hat{\Omega}} + J(t,x,\hat{u}(\cdot)) \mathbf{1}_{\hat{\Omega}}. \end{aligned}$$

In fact,

$$\begin{aligned} X(s;t,x,\hat{u}(\cdot)) \mathbf{1}_{\hat{\Omega}} &= \mathbf{1}_{\hat{\Omega}} \left[x + \int_t^s b(r,X(r;t,x,\hat{u}(\cdot)),\hat{u}(r)) dr \right. \\ &\quad \left. + \int_t^s \sigma(r,X(r;t,x,\hat{u}(\cdot)),\hat{u}(r)) dW(r) \right] \end{aligned}$$

$$\begin{aligned}
 &= \mathbf{1}_{\widehat{\Omega}}x + \int_t^s \mathbf{1}_{\widehat{\Omega}}b(r, X(r; t, x, \widehat{u}(\cdot)), \widehat{u}(r))dr \\
 &\quad + \int_t^s \mathbf{1}_{\widehat{\Omega}}\sigma(r, X(r; t, x, \widehat{u}(\cdot)), \widehat{u}(r))dW(r) \\
 &= \mathbf{1}_{\widehat{\Omega}}x + \int_t^s b(r, X(r; t, x, \widehat{u}(\cdot))\mathbf{1}_{\widehat{\Omega}}, \widehat{u}(r)\mathbf{1}_{\widehat{\Omega}})dr \\
 &\quad + \int_t^s \sigma(r, X(r; t, x, \widehat{u}(\cdot))\mathbf{1}_{\widehat{\Omega}}, \widehat{u}(r)\mathbf{1}_{\widehat{\Omega}})dW(r).
 \end{aligned}$$

Likewise,

$$\begin{aligned}
 X(s; t, x, \bar{u}(\cdot))\mathbf{1}_{\Omega \setminus \widehat{\Omega}} &= \mathbf{1}_{\Omega \setminus \widehat{\Omega}}x + \int_t^s b(r, X(r; t, x, \bar{u}(\cdot))\mathbf{1}_{\Omega \setminus \widehat{\Omega}}, \bar{u}(r)\mathbf{1}_{\Omega \setminus \widehat{\Omega}})dr \\
 &\quad + \int_t^s \sigma(r, X(r; t, x, \bar{u}(\cdot))\mathbf{1}_{\Omega \setminus \widehat{\Omega}}, \bar{u}(r)\mathbf{1}_{\Omega \setminus \widehat{\Omega}})dW(r).
 \end{aligned}$$

Hence,

$$\begin{aligned}
 &X(s; t, x, \bar{u}(\cdot))\mathbf{1}_{\Omega \setminus \widehat{\Omega}} + X(s; t, x, \widehat{u}(\cdot))\mathbf{1}_{\widehat{\Omega}} \\
 &= x + \int_t^s b(r, X(r; t, x, \bar{u}(\cdot))\mathbf{1}_{\Omega \setminus \widehat{\Omega}} + X(r; t, x, \widehat{u}(\cdot))\mathbf{1}_{\widehat{\Omega}}, \bar{u}(r)\mathbf{1}_{\Omega \setminus \widehat{\Omega}} + \widehat{u}(r)\mathbf{1}_{\widehat{\Omega}})dr \\
 &\quad + \int_t^s \sigma(r, X(r; t, x, \bar{u}(\cdot))\mathbf{1}_{\Omega \setminus \widehat{\Omega}} + X(r; t, x, \widehat{u}(\cdot))\mathbf{1}_{\widehat{\Omega}}, \bar{u}(r)\mathbf{1}_{\Omega \setminus \widehat{\Omega}} + \widehat{u}(r)\mathbf{1}_{\widehat{\Omega}})dW(r) \\
 &= x + \int_t^s b(r, X(r; t, x, \bar{u}(\cdot))\mathbf{1}_{\Omega \setminus \widehat{\Omega}} + X(r; t, x, \widehat{u}(\cdot))\mathbf{1}_{\widehat{\Omega}}, \bar{u}(r))dr \\
 &\quad + \int_t^s \sigma(r, X(r; t, x, \bar{u}(\cdot))\mathbf{1}_{\Omega \setminus \widehat{\Omega}} + X(r; t, x, \widehat{u}(\cdot))\mathbf{1}_{\widehat{\Omega}}, \bar{u}(r))dW(r).
 \end{aligned}$$

Then, by uniqueness, we obtain the first equality in(13). Similarly, we have the second one. Now, (12) leads to

$$\begin{aligned}
 \bar{J}(t, x; \bar{u}(\cdot)) &= \mathbb{E}[J(t, x; \bar{u}(\cdot))] = \mathbb{E}\left[J(t, x; \bar{u}(\cdot))\mathbf{1}_{\Omega \setminus \widehat{\Omega}} + J(t, x; \widehat{u}(\cdot))\mathbf{1}_{\widehat{\Omega}}\right] \\
 &< \mathbb{E}\left[J(t, x; \bar{u}(\cdot))\mathbf{1}_{\Omega \setminus \widehat{\Omega}} + J(t, x; \bar{u}(\cdot))\mathbf{1}_{\widehat{\Omega}}\right] = \mathbb{E}\left[J(t, x; \bar{u}(\cdot))\right] = \bar{J}(t, x; \bar{u}(\cdot)),
 \end{aligned}$$

contradicting the optimality of $\bar{u}(\cdot)$. Hence, Problems (C) and (\bar{C}) are equivalent. □

Next, let $c(\cdot)$ be a consumption rate and $\xi \in L^p_{\mathcal{F}_T}(\Omega)$ be a payoff/reward at $t = T$. Let $Y(\cdot)$ solve the following equation:

$$(14) \quad Y(t) = \mathbb{E}_t \left[\xi + \int_t^T g(c(s), Y(s))ds \right], \quad t \in [0, T],$$

for some proper map $g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$. Process $Y(\cdot)$ is called a *recursive utility process* (which is also called a *stochastic differential utility*, introduced by Duffie and Epstein [21]), and $g(\cdot)$ is called an *aggregator*. It turns out that if $(Y(\cdot), Z(\cdot))$ is the

adapted solution to the following so-called *backward stochastic differential equation* (BSDE, for short) in its integral form:

$$(15) \quad Y(t) = \xi + \int_t^T g(c(s), Y(s)) ds - \int_t^T Z(s) dW(s), \quad t \in [0, T],$$

then (14) holds. This suggests us to consider the following BSDE:

$$(16) \quad \begin{aligned} Y(r) = & h(X(T)) + \int_r^T g(s, X(s), u(s), Y(s), Z(s)) ds \\ & - \int_r^T Z(s) dW(s), \quad s \in [t, T], \end{aligned}$$

with $t \in [0, T)$ being a parameter, and $X(\cdot) \equiv X(\cdot; t, x, u(\cdot))$ being the state process. Under proper conditions, the above BSDE admits a unique adapted solution $(Y(\cdot), Z(\cdot)) \equiv (Y(\cdot; t, x, u(\cdot)), Z(\cdot; t, x, u(\cdot)))$. Then we may let

$$(17) \quad \begin{aligned} J(t, x; u(\cdot)) = & Y(t) \equiv Y(t; t, x, u(\cdot)) \\ = & \mathbb{E}_t \left[h(X(T)) + \int_t^T g(s, X(s), u(s), Y(s), Z(s)) ds \right]. \end{aligned}$$

This is called a *recursive cost functional*. With such a cost functional, we can pose an optimal control problem.

Note that one might also define

$$(18) \quad \begin{aligned} \bar{J}(t, x; u(\cdot)) = & \mathbb{E}[Y(t)] \equiv \mathbb{E}[Y(t; t, x, u(\cdot))] \\ = & \mathbb{E} \left[h(X(T)) + \int_t^T g(s, X(s), u(s), Y(s), Z(s)) ds \right]. \end{aligned}$$

But, $J(t, x; u(\cdot))$ seems to be more natural, as the involved BSDE is only solved on $[t, T]$.

We recall that for the following general BSDE:

$$Y(t) = \xi + \int_t^T \bar{g}(s, Y(s), Z(s)) ds - \int_t^T Z(s) dW(s), \quad t \in [0, T],$$

with some proper map $\bar{g} : [0, T] \times \mathbb{R} \times \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}$ satisfying some suitable condition, for any $\tau \in [0, T)$, one has

$$Y(t) = Y(\tau) + \int_t^\tau \bar{g}(s, Y(s), Z(s)) ds - \int_t^\tau Z(s) dW(s), \quad t \in [0, \tau].$$

Thus, if we denote the adapted solution of the above by $(Y(\cdot; \tau, Y(\tau)), Z(\cdot; \tau, Y(\tau)))$, then the following semigroup property holds:

$$(Y(t; T, \xi), Z(t; T, \xi)) = (Y(t; \tau, Y(\tau; T, \xi)), Z(t; \tau, Y(\tau; T, \xi))), \quad 0 \leq t < \tau \leq T.$$

Having this, similar to the formal derivation for Problem (\bar{C}) , we can show that the problem with the recursive cost functional (18) is also time-consistent. If we still denote the value function by $V(\cdot, \cdot)$. Then the corresponding HJB equation takes the following form (comparing with (9)):

$$(19) \quad \begin{cases} V_t(t, x) + \inf_{u \in U} \left\{ \frac{1}{2} \text{tr} [V_{xx}(t, x) \sigma(t, x, u) \sigma(t, x, u)^\top] \right. \\ \quad \left. + V_x(t, x) b(t, x, u) + g(t, x, u, V(t, x), V_x(t, x) \sigma(t, x, u)) \right\} = 0, \\ \quad \quad \quad (t, x) \in [0, T] \times \mathbb{R}^n, \\ V(T, x) = h(x), \quad x \in \mathbb{R}^n. \end{cases}$$

We will have some interesting comparisons later.

3 Time-Inconsistent Optimal Control Problems

We now, present several possible formulations of time-inconsistent stochastic optimal control problems.

3.1 Time-preferences and general discounting

Let state equation be (1). We consider several kinds of cost functionals.

1. *Non-exponential discounting.* Let us replace $t \mapsto e^{-\delta t}$ by some general decreasing non-negative function $t \mapsto \lambda(t)$ in the cost functional (10). Then the new cost functional will take the following form:

$$J(t, x; u(\cdot)) = \mathbb{E}_t \left[\lambda(T-t) h(X(T)) + \int_t^T \lambda(s-t) g(s, X(s), u(s)) ds \right].$$

This suggests us consider the following more general cost functional:

$$(20) \quad J(t, x; u(\cdot)) = \mathbb{E}_t \left[h(t, X(T)) + \int_t^T g(t, s, X(s), u(s)) ds \right],$$

where $g : \Delta[0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}$ with

$$\Delta[0, T] = \{(t, s) \in [0, T] \times [0, T] \mid t \leq s\}.$$

We see that due to the general dependence of $h(t, x)$ and $g(t, s, x, u)$ on t , one can use such a cost functional to describe the problems with general discounting. With the above cost functional, we could formulate the following optimal control problem.

Problem (N). For any give initial pair $(t, x) \in [0, T] \times \mathbb{R}^n$, find a control $\bar{u}(\cdot) \in \mathcal{U}[t, T]$ such that with the cost functional (20),

$$J(t, x; \bar{u}(\cdot)) = \inf_{u(\cdot) \in \mathcal{U}[t, T]} J(t, x; u(\cdot)).$$

One can present examples that in general the above problem is time-inconsistent, see [112]. It is also possible to introduce the following cost functional:

$$(21) \quad \bar{J}(t, x; u(\cdot)) = \mathbb{E}\left[h(t, X(T)) + \int_t^T g(t, s, X(s), u(s))ds\right] \equiv \mathbb{E}[J(t, x; u(\cdot))].$$

With such a cost functional, the optimal control problem is also time-inconsistent.

2. *Recursive cost functional with non-exponential discounting.* We have seen from the previous section that one could consider optimal control problems with recursive cost functionals. It is not hard to understand one could introduce the exponential discounting into the recursive utility (14), so that it takes the following form:

$$(22) \quad Y(t) = \mathbb{E}_t \left[e^{-\delta(T-t)} \xi + \int_t^T e^{-\delta(s-t)} g(c(s), Y(s)) ds \right], \quad t \in [0, T],$$

for some $\delta > 0$. If the discounting $t \mapsto e^{-\delta t}$ is replaced by a general discounting $t \mapsto \lambda(t)$, then one has

$$Y(t) = \mathbb{E}_t \left[\lambda(T-t)\xi + \int_t^T \lambda(s-t)g(c(s), Y(s))ds \right], \quad t \in [0, T].$$

This suggests us to consider the following BSDE:

$$(23) \quad \begin{cases} dY(t, s) = -g(t, s, X(t), X(s), u(s), Y(t, s), Z(t, s))ds \\ \qquad\qquad\qquad + Z(t, s)dW(s), \quad s \in [t, T], \\ Y(t, T) = h(t, X(t), X(T)), \end{cases}$$

with $t \in [0, T)$ being a parameter, and $X(\cdot) \equiv X(\cdot; t, x, u(\cdot))$ being the state process. Thus, the map $g : \Delta[0, T] \times \mathbb{R}^n \times \mathbb{R}^n \times U \times \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$. Then set

$$(24) \quad J(t, x; u(\cdot)) = Y(t, t).$$

This is called a *recursive cost functional* with general discounting. With the cost functional (24), we can formulation Problem (N) exactly same as before. Since cost functional (24) is more general than (20), we expect that the Problem (N) associated with cost functional (24) should be time-inconsistent as well. A treatment of a similar problem was carried out in [104]. Some further extension for the problem with state equation containing regime switching was discussed in [61].

3. *Equilibrium recursive cost functional.* We may write (23) in its integral form as follows

$$(25) \quad Y(t, r) = h(t, X(t), X(T)) + \int_r^T g(t, s, X(t), X(s), u(s), Y(t, s), Z(t, s)) ds - \int_r^T Z(t, s) dW(s), \quad r \in [t, T].$$

If we take $r = t$, then it looks like

$$(26) \quad Y(t, t) = h(t, X(t), X(T)) + \int_t^T g(t, s, X(t), X(s), u(s), Y(t, s), Z(t, s)) ds - \int_t^T Z(t, s) dW(s), \quad r \in [t, T].$$

Note that this is not an equation for $(Y(t, t), Z(t, s))$ since $Y(t, s)$ appears on the right-hand side which cannot be determined by such an equation, in general. However, the above observation suggests us to introduce the following equation:

$$(27) \quad Y(t) = h(t, X(t), X(T)) + \int_t^T g(t, s, X(t), X(s), u(s), Y(s), Z(t, s)) ds - \int_t^T Z(t, s) dW(s), \quad t \in [0, T],$$

with $(Y(\cdot), Z(\cdot, \cdot))$ being its unknown to be found. Such an equation is called a *backward stochastic Volterra integral equation* (BSVIE, for short), see [107, 108]. If $(Y(\cdot), Z(\cdot, \cdot))$ is the solution to the above, we may define

$$(28) \quad J(t, x; u(\cdot)) = Y(t) \equiv Y(t; u(\cdot)),$$

and call it the *equilibrium recursive cost functional*. Note that the process $(t, r) \mapsto Y(t, r)$ has a hidden nature of time-inconsistency. Whereas, the process $Y(\cdot)$ determined by BSVIE (27) removes such a hidden time-inconsistency. Now, we may pose Problem (N) with the cost functional given by (28). Although we do not have comparison between (24) and (28), since (28) is more general than (20), the corresponding Problem (N) is also time-inconsistent.

3.2 Risk-preferences and subjective expectation

We now look at the risk-preference aspect. We still consider the state equation (1). Risk-preference is referred to the following: Different (groups of) people will have different opinions on the risks that contained in the coming event. Optimistic people will think that the risk will not be large, and pessimistic people will feel that the risk will be big. Therefore, one could understand that when expectation is calculated, the actual probability used by different people should be *subjective*, rather than *objective*. Rigorously, one should accept that in the cost functional, say, (10), the operator

\mathbb{E} might be better replace by some \mathcal{E} , representing the expectation with respect to some subjective probability. Let us look at some possibilities.

1. *Nonlinear expectation by BSDEs.* In 1997, Peng introduced a nonlinear expectation determined by the adapted solutions ([81]). More precisely, consider the following BSDE:

$$(29) \quad Y(t) = \xi + \int_t^T \bar{g}(s, Y(s), Z(s)) ds - \int_t^T Z(s) dW(s), \quad t \in [0, T],$$

where $\bar{g} : [0, T] \times \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies the following:

(B) For all $(y, z) \in \mathbb{R} \times \mathbb{R}^d, t \mapsto \bar{g}(t, y, z)$ is continuous, with

$$\bar{g}(t, y, 0) = 0, \quad \forall (t, y) \in [0, T] \times \mathbb{R},$$

and there exists a constant $L > 0$ such that

$$|\bar{g}(t, y, z) - \bar{g}(t, y', z')| \leq L(|y - y'| + |z - z'|), \\ \forall t \in [0, T], y, y' \in \mathbb{R}, z, z' \in \mathbb{R}^d.$$

It is standard that under (B), for any $\xi \in L^p_{\mathcal{F}_T}(\Omega), p > 1$, BSDE (29) admits a unique adapted solution $(Y(\cdot), Z(\cdot))$ [78, 63]. By [81], one can define

$$\mathcal{E}_t[\xi] = Y(t) \equiv Y(t; \xi), \quad t \in [0, T]; \quad \mathcal{E}[\xi] \equiv \mathcal{E}_0[\xi].$$

We call $\mathcal{E}_t[\xi]$ the *nonlinear conditional expectation* of ξ given \mathcal{F}_t , associated with $\bar{g}(\cdot)$, and $\mathcal{E}[\xi] \equiv \mathcal{E}_0[\xi]$ the corresponding *nonlinear expectation* of ξ associated with $\bar{g}(\cdot)$. It was called *g-expectation*. Clearly, different choice of $\bar{g}(\cdot)$ (satisfying (B)) will give different nonlinear (conditional) expectations. Therefore, we could replace \mathbb{E}_t in (10) (with $\delta = 0$, for simplicity) by \mathcal{E}_t determined by some $\bar{g}(\cdot)$:

$$(30) \quad J(t, x; u(\cdot)) = \mathcal{E}_t \left[\int_t^T g(s, X(s), u(s)) ds + h(X(T)) \right],$$

and replace \mathbb{E} in (4) by \mathcal{E} (also with $\delta = 0$):

$$(31) \quad \bar{J}(t, x; u(\cdot)) = \mathcal{E} \left[\int_t^T g(s, X(s), u(s)) ds + h(X(T)) \right].$$

By the semigroup property of the adapted solutions to BSDEs, we have (comparing with (11))

$$(32) \quad \mathcal{E} [J(t, x; u(\cdot))] = \bar{J}(t, x; u(\cdot)).$$

Now, let us take a closer look at the above. According to the definition of \mathcal{E}_t , we need to solve the following BSDE (parameterized by $(t, x, u(\cdot))$):

$$\begin{aligned}
 (33) \quad Y(t, r) &= h(X(T)) + \int_t^T g(s, X(s), u(s)) ds \\
 &\quad + \int_r^T \bar{g}(s, Y(t, s), Z(t, s)) ds - \int_r^T Z(t, s) dW(s), \quad r \in [t, T],
 \end{aligned}$$

where the dependence of the adapted solution on the parameter t is emphasized and the dependence on $(x, u(\cdot))$ is suppressed. Then

$$J(t, x; u(\cdot)) = Y(t, t).$$

If we take $r = t$ in (33), then

$$\begin{aligned}
 (34) \quad Y(t, t) &= h(X(T)) + \int_t^T g(s, X(s), u(s)) ds \\
 &\quad + \int_t^T \bar{g}(s, Y(t, s), Z(t, s)) ds - \int_t^T Z(t, s) dW(s), \quad t \in [0, T].
 \end{aligned}$$

We may mimic item 3 from the previous subsection to introduce the following *equilibrium* integral equation:

$$\begin{aligned}
 (35) \quad Y(t) &= h(X(T)) + \int_t^T [g(s, X(s), u(s)) + \bar{g}(s, Y(s), Z(s))] ds \\
 &\quad - \int_t^T Z(s) dW(s), \quad t \in [0, T],
 \end{aligned}$$

which turns out to be a BSDE, and the cost functional becomes a recursive cost functional. This shows that for the nonlinear expectation determined by BSDEs, there is a natural way to transform the cost function to a standard recursive cost functional so that the corresponding optimal control problem becomes time-consistent. However, we should point out that this can be done only if there is not general discounting.

2. *Distortion of probability.* Let $\rho : [0, 1] \rightarrow [0, 1]$ be continuous, increasing such that $\rho(0) = 0$ and $\rho(1) = 1$. Such a $\rho(\cdot)$ is called a *distortion function*. Now, for any random variable, we define the following *distorted expectation*:

$$\begin{aligned}
 (36) \quad \mathcal{E}^\rho[\xi] &= \int_\Omega \xi(\omega) d(\rho \circ \mathbb{P})(\omega) \\
 &\triangleq \int_{-\infty}^0 [\rho(\mathbb{P}(\xi \geq t)) - 1] dt + \int_0^\infty \rho(\mathbb{P}(\xi \geq t)) dt.
 \end{aligned}$$

Note that in the case $\rho(r) = r, r \in [0, 1]$, we can check that $\mathcal{E}^\rho = \mathbb{E}$. We also call $\rho \circ \mathbb{P}$ is a *distorted probability*, which is not a probability measure. Note that different groups of people will have different distortion function $\rho(\cdot)$. For a given event $A \in \mathcal{F}$, the group associated with $\rho(\cdot)$ have the opinion that $\rho(\mathbb{P}(A))$ should be the probability of A , although it is not. Typical shapes of $\rho(\cdot)$ could be convex, concave, ‘‘S-shaped’’, or ‘‘backward S-shaped’’. Let us elaborate this a little more.

(i) If $\rho(\cdot)$ is convex, then

$$\rho(r) < r, \quad r \in (0, 1).$$

Thus, for any event $A \in \mathcal{F}$, these people think that the probability of A is smaller than $\mathbb{P}(A)$. Hence, these people feel that “nothing will happen”.

(ii) If $\rho(\cdot)$ is concave, then

$$\rho(r) > r, \quad \forall r \in (0, 1).$$

Thus, these people exaggerate everything.

(iii) If $\rho(\cdot)$ is “S-shaped”, then there exists a $\beta \in (0, 1)$ such that $\rho(\cdot)$ is convex on $(0, \beta)$ and concave on $(\beta, 1)$. Thus, these people exaggerate large probability events and understate small probability events.

(iv) If $\rho(\cdot)$ is “backward S-shaped”, then there exists a $\beta \in (0, 1)$ such that $\rho(\cdot)$ is concave on $(0, \beta)$ and convex on $(\beta, 1)$. Thus, these people exaggerate small probability events and understate large probability events. Buy insurance and buy lottery are kind of behavior that exaggerates small probability events.

Now, for any initial pair $(t, x) \in [0, T] \times \mathbb{R}^n$ and control $u(\cdot) \in \mathcal{U}[t, T]$, let $X(\cdot; t, x, u(\cdot))$ be the state process. We introduce the following cost functional under distortion $\rho(\cdot)$:

$$(37) \quad \bar{J}(t, x; u(\cdot)) = \mathcal{E}^\rho \left[\int_t^T g(s, X(s), u(s)) ds + h(X(T)) \right].$$

Then we may pose an optimal control problem associated with such a cost functional. Such a problem should be time-inconsistent. To our best knowledge, such a problem has not been carefully studied from the time-inconsistent point of view. There are some investigations by means of quantiles ([48, 38, 119]).

3. *Conditional expectations appear nonlinearly.* Another possibility of expressing risk-preferences is to allow the conditional expectation nonlinearly appear in the cost functional. More precisely, for state equation (1), we may introduce the following cost functional

$$(38) \quad J(t, x; u(\cdot)) = \mathbb{E}_t \left[\int_t^T g(s, X(s), \mathbb{E}_t[X(s)], u(s), \mathbb{E}_t[u(s)]) ds + h(X(T), \mathbb{E}_t[X(T)]) \right].$$

With such a cost functional, we can pose an optimal control problem. It turns out that such a problem is time-inconsistent.

3.3 Mixed situations

We now look at the case that both time-preferences and risk-preferences appear together. According to the previous subsections, we may have the following type cost functional:

$$(39) \quad J(t, x; u(\cdot)) = \mathcal{E}^\rho \left[h(t, X(T)) + \int_t^T g(t, s, X(s), u(s)) ds \right],$$

with some distortion function ρ and some maps $h : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \Delta [0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}$.

We now describe the next possibility. The following BSVIE with mean-field (MF-BSVIE, for short) was studied in [94]:

$$(40) \quad \begin{aligned} Y(t) &= h(t, X(t), X(T)) \\ &+ \int_t^T g(t, s, X(t), X(s), \mathbb{E}_t[X(s)], u(s), \mathbb{E}_t[u(s)], Y(s), \mathbb{E}_t[Y(s)], \\ &\quad Z(t, s), \mathbb{E}_t[Z(t, s)]) ds - \int_t^T Z(t, s) dW(s), \quad t \in [0, T], \end{aligned}$$

and define

$$(41) \quad J(t, x; u(\cdot)) = Y(t).$$

Then we can pose optimal control problem with the state equation (1) and the cost functional (41). Such a problem is, of course, time-inconsistent. In the case that the state equation is linear, with the cost functional containing general discounting and quadratic forms of conditional expectations of the state and the control, equilibrium strategy was constructed in [113].

4 Equilibrium Strategies, Variational Method and Necessary Conditions

For time-inconsistent optimal control problems, it is not wise to find optimal control for any given initial pair. In fact, if $(t, x) \in [0, T] \times \mathbb{R}^n$ is a given initial pair, and we have found an optimal control $\bar{u}(\cdot) \equiv \bar{u}(\cdot; t, x)$, then at a later time $\tau \in (t, T]$, one may find optimal control at $(\tau, \bar{X}(\tau))$ denoted by $\bar{u}(\cdot; \tau, \bar{X}(\tau))$, and one could have

$$J(\tau, \bar{X}(\tau); \bar{u}(\cdot; \tau, \bar{X}(\tau))) > J(\tau, \bar{X}(\tau); \bar{u}(\cdot; t, x)|_{[\tau, T]}).$$

Hence, formally, one has to have a control $\hat{u}(s, t, \bar{X}(t))$ of two time variables (t, s) to keep it “optimal”. Clearly, this is not practically feasible. Hence, instead, we need to find some time-consistent control/strategies which still keep certain type

of optimality. For definiteness, let us use Problem (N) to represent a general time-inconsistent optimal control problem with the state equation (1) and with the cost functional denoted by $J(t, x; u(\cdot))$ which could be any of the ones described above. We introduce the following definition.

Definition 1. (i) For given $x \in \mathbb{R}^n$, a $\bar{u}(\cdot)$ is called an *open-loop equilibrium control* if for any $t \in [0, T]$, and any $u \in U$,

$$(42) \quad \lim_{\varepsilon \downarrow 0} \frac{J(t, \bar{X}(t); u^\varepsilon(\cdot)) - J(t, \bar{X}(t); \bar{u}(\cdot))}{\varepsilon} \geq 0,$$

where $\bar{X}(\cdot) = X(\cdot; 0, x, \bar{u}(\cdot))$ and

$$(43) \quad u^\varepsilon(s) = \begin{cases} u, & s \in [t, t + \varepsilon), \\ \bar{u}(s), & s \in [t + \varepsilon, T], \end{cases}$$

(ii) A map $\Psi : [0, T] \times \mathbb{R}^n \rightarrow U$ is called a *closed-loop equilibrium strategy* if for every $x \in \mathbb{R}^n$ the following equation

$$(44) \quad \begin{cases} d\bar{X}(s) = b(s, \bar{X}(s), \Psi(s, \bar{X}(s)))ds + \sigma(s, \bar{X}(s), \Psi(s, \bar{X}(s)))dW(s), & s \in [0, T], \\ \bar{X}(0) = x, \end{cases}$$

admits a unique solution $\bar{X}(\cdot)$ and for each $(t, u) \in [0, T] \times U$, let $X^\varepsilon(\cdot)$ be the solution to the following:

$$(45) \quad \begin{cases} dX^\varepsilon(s) = b(s, X^\varepsilon(s), u)ds + \sigma(s, X^\varepsilon(s), u)dW(s), & s \in [t, t + \varepsilon), \\ dX^\varepsilon(s) = b(s, X^\varepsilon(s), \Psi(s, X^\varepsilon(s)))ds + \sigma(s, X^\varepsilon(s), \Psi(s, X^\varepsilon(s)))dW(s), & s \in [t + \varepsilon, T], \\ X^\varepsilon(t) = \bar{X}(t). \end{cases}$$

and the following holds:

$$(46) \quad \lim_{\varepsilon \downarrow 0} \frac{J(t, \bar{X}(t); u\mathbf{1}_{[t, t+\varepsilon)} \oplus \Psi) - J(t, \bar{X}(t); \Psi)}{\varepsilon} \geq 0,$$

where

$$(47) \quad (u\mathbf{1}_{[t, t+\varepsilon)} \oplus \Psi)(s) = \begin{cases} u, & s \in [t, t + \varepsilon), \\ \Psi(s, X^\varepsilon(s)), & s \in [t + \varepsilon, T], \end{cases}$$

We note that the equilibrium controls/strategies have to major features: time-consistency (represented by $\bar{u}(s)$ and $\Psi(s, \bar{X}(s))$ only depend on one time variable), and local optimality (exhibited by (42) and (87), respectively). open-loop equilibrium strategy $\bar{u}(\cdot)$ is initial state x dependent (through the process $\bar{X}(\cdot)$). Whereas, closed-loop equilibrium strategy is initial state x independent.

For linear quadratic problems, open-loop equilibrium control was studied in [43, 44]. For nonlinear case, inspired by the general stochastic optimal control theory, one could expect to have Pontryagin type maximum principle. We now would like to derive necessary conditions of open-loop equilibrium strategies for the time-inconsistent optimal control problem associated with some cost functional. The purpose is to present the main idea. We do not pursue the most generality. However, even such a result seems to be new, to our best knowledge. With our idea below, it is possible to derive necessary conditions for open-loop equilibrium controls of problems associated with other type of cost functionals. See [80, 116, 28, 74, 109] for relevant works.

Consider cost functional (20) which is rewritten below for convenience:

$$(48) \quad J(t, x; u(\cdot)) = \mathbb{E}_t \left[h(t, X(T)) + \int_t^T g(t, s, X(s), u(s)) ds \right].$$

Namely, we are now only concerned with the general discounting case. We introduce the following assumption.

(H3) Let $U \subseteq \mathbb{R}^m$ be compact, $d = 1$. Let $b, \sigma : [0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}^n, g : \Delta[0, T] \times \mathbb{R}^n \times U \rightarrow \mathbb{R}, h : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$ be continuous and

$$x \mapsto (b(s, x, u), \sigma(s, x, u), g(t, s, x, u), h(t, x))$$

is C^2 with all derivatives up to order 2 being bounded and Lipschitz continuous in x .

The conditions assumed in (H3) can be relaxed. But, we do not pursue the most generality. Note that since U is assumed to be compact, $\mathcal{U}^p[t, T]$ it is independent of $p \in [1, \infty]$. Thus, we will simply write $\mathcal{U}[t, T]$ instead below.

The following gives a necessary condition for open-loop equilibrium control for our corresponding time-inconsistent optimal control problem.

Theorem 1. *Let (H3) hold. Suppose $(\bar{X}(\cdot), \bar{u}(\cdot))$ is an equilibrium pair on $[0, T]$, associated with some initial state $x \in \mathbb{R}^n$. Suppose for any given $t \in [0, T]$, $(Y(\cdot; t), Z(\cdot; t))$ and $(P(\cdot; t), \Lambda(\cdot; t))$ are the adapted solutions to the following BSDEs, respectively:*

$$(49) \quad \begin{cases} Y(s; t) = -[b_x(s, \bar{X}(s), \bar{u}(s))^\top Y(s; t) + \sigma_x(s, \bar{X}(s), \bar{u}(s))^\top Z(s; t) \\ \quad + g_x(t, s, \bar{X}(s), \bar{u}(s))^\top] ds + Z(s; t) dW(s), \quad s \in [t, T], \\ Y(T; t) = h_x(t, \bar{X}(T))^\top, \end{cases}$$

and

$$(50) \quad \begin{cases} dP(s;t) = - \left(b_x(s, \bar{X}(s), \bar{u}(s))^\top P(s;t) + P(s;t) b_x(s, \bar{X}(s), \bar{u}(s)) \right. \\ \quad \left. + \sigma_x(s, \bar{X}(s), \bar{u}(s))^\top P(s;t) \sigma_x(s, \bar{X}(s), \bar{u}(s)) \right. \\ \quad \left. + \sigma_x(s, \bar{X}(s), \bar{u}(s))^\top \Lambda(s;t) + \Lambda(s;t) \sigma_x(s, \bar{X}(s), \bar{u}(s)) \right. \\ \quad \left. + g_{xx}(t, s, \bar{X}(s), \bar{u}(s)) \right) ds + \Lambda(s;t) dW(s), \quad s \in [t, T], \\ P(T;t) = h_{xx}(t, \bar{X}(T)). \end{cases}$$

Then almost surely, for any $u \in U$,

$$(51) \quad \begin{aligned} 0 &\leq \langle Y(t;t), b(t, \bar{X}(t), u) - b(t, \bar{X}(t), \bar{u}(t)) \rangle \\ &+ \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \mathbb{E}_t \int_t^{t+\varepsilon} \langle Z(s;t), \sigma(s, \bar{X}(s), u) - \sigma(s, \bar{X}(s), \bar{u}(s)) \rangle ds \\ &+ g(t, t, \bar{X}(t), u) - g(t, t, \bar{X}(t), \bar{u}(t)) \\ &+ \text{tr} \{ [\sigma(t, \bar{X}(t), u) - \sigma(t, \bar{X}(t), \bar{u}(t))]^\top P(t) [\sigma(t, \bar{X}(t), u) - \sigma(t, \bar{X}(t), \bar{u}(t))] \}. \end{aligned}$$

This is basically a Pontryagin's type maximum principle.

Proof. Suppose $\bar{u}(\cdot) \in \mathcal{U}[0, T]$ is an open-loop equilibrium control on $[0, T]$ for a given initial state $x \in \mathbb{R}^n$, and denote the corresponding open-loop equilibrium state process by $\bar{X}(\cdot)$. For each $t \in [0, T)$ and $u \in U$, let $X^\varepsilon(\cdot)$ be the solution of the state equation on $[t, T]$ corresponding to $(t, \bar{X}(t), u^\varepsilon(\cdot))$. Introduce the following:

$$\begin{aligned} b_x^\varepsilon(s) &= \int_0^1 b_x(s, \bar{X}(s) + \beta \widehat{X}^\varepsilon(s), u^\varepsilon(s)) d\beta, \\ \sigma_x^\varepsilon(s) &= \int_0^1 \sigma_x(s, \bar{X}(s) + \beta \widehat{X}^\varepsilon(s), u^\varepsilon(s)) d\beta, \\ \widehat{b}(s) &= b(s, \bar{X}(s), u) - b(s, \bar{X}(s), \bar{u}(s)), \quad \widehat{b}^\varepsilon(s) = \widehat{b}(s) \mathbf{1}_{[t, t+\varepsilon)}(s), \\ \widehat{\sigma}(s) &= \sigma(s, \bar{X}(s), u) - \sigma(s, \bar{X}(s), \bar{u}(s)), \quad \widehat{\sigma}^\varepsilon(s) = \widehat{\sigma}(s) \mathbf{1}_{[t, t+\varepsilon)}(s). \end{aligned}$$

Let $X_1^\varepsilon(\cdot)$ and $X_2^\varepsilon(\cdot)$ be the solution to the following:

$$\begin{cases} dX_1^\varepsilon(s) = b_x^\varepsilon(s) X_1^\varepsilon(s) ds + [\sigma_x^\varepsilon(s) X_1^\varepsilon(s) + \widehat{\sigma}^\varepsilon(s)] dW(s), & s \in [t, T], \\ X_1^\varepsilon(t) = 0, \end{cases}$$

$$\begin{cases} dX_2^\varepsilon(s) = [b_x^\varepsilon(s) X_2^\varepsilon(s) + \widehat{b}^\varepsilon(s)] ds + \sigma_x^\varepsilon(s) X_2^\varepsilon(s) dW(s), & s \in [t, T], \\ X_2^\varepsilon(t) = 0. \end{cases}$$

Then

$$\begin{aligned} \mathbb{E}_t \left[\sup_{s \in [t, T]} |X_1^\varepsilon(s)|^2 \right] &\leq K \mathbb{E}_t \int_t^T |\widehat{\sigma}^\varepsilon(s)|^2 ds = K \mathbb{E}_t \int_t^{t+\varepsilon} |\widehat{\sigma}(s)|^2 ds \\ &\leq K \mathbb{E}_t \int_t^{t+\varepsilon} (1 + |\bar{X}(s)|)^2 ds \leq K\varepsilon \left(1 + \mathbb{E}_t \left[\sup_{s \in [t, T]} |\bar{X}(s)| \right] \right) \leq K\varepsilon, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_t \left[\sup_{s \in [t, T]} |X_2^\varepsilon(s)|^2 \right] &\leq K \mathbb{E}_t \left(\int_t^T |\widehat{b}^\varepsilon(s)| ds \right)^2 = K \mathbb{E}_t \left(\int_t^{t+\varepsilon} |\widehat{b}(s)| ds \right)^2 \\ &\leq K \mathbb{E}_t \left(\int_t^{t+\varepsilon} (1 + |\bar{X}(s)|) ds \right)^2 \leq K\varepsilon^2 \left(1 + \mathbb{E}_t \left[\sup_{s \in [t, T]} |\bar{X}(s)| \right] \right)^2 \leq K\varepsilon^2. \end{aligned}$$

If we denote

$$\widehat{X}^\varepsilon(\cdot) \equiv X^\varepsilon(\cdot) - \bar{X}(\cdot) = X_1^\varepsilon(\cdot) + X_2^\varepsilon(\cdot),$$

then the following holds:

$$\begin{cases} d\widehat{X}^\varepsilon(s) = [b_x^\varepsilon(s)\widehat{X}^\varepsilon(s) + \widehat{b}^\varepsilon(s)]ds + [\sigma_x^\varepsilon(s)\widehat{X}^\varepsilon(s) + \widehat{\sigma}^\varepsilon(s)]dW(s), & s \in [t, T], \\ \widehat{X}^\varepsilon(t) = 0, \end{cases}$$

Further, denoting $\mathbb{X}^\varepsilon(\cdot) = X_1^\varepsilon(\cdot)X_1^\varepsilon(\cdot)^\top$, we have

$$\begin{aligned} d\mathbb{X}^\varepsilon(s) &= [b_x^\varepsilon(s)\mathbb{X}^\varepsilon(s) + \mathbb{X}^\varepsilon(s)b_x^\varepsilon(s)^\top + \sigma_x^\varepsilon(s)\mathbb{X}^\varepsilon(s)\sigma_x^\varepsilon(s)^\top \\ &\quad + \sigma_x^\varepsilon(s)X_1^\varepsilon(s)\widehat{\sigma}^\varepsilon(s)^\top + \widehat{\sigma}^\varepsilon(s)X_1^\varepsilon(s)^\top\sigma_x^\varepsilon(s)^\top + \widehat{\sigma}^\varepsilon(s)\widehat{\sigma}^\varepsilon(s)^\top]ds \\ &\quad + [\sigma_x^\varepsilon(s)\mathbb{X}^\varepsilon(s) + \mathbb{X}^\varepsilon(s)\sigma_x^\varepsilon(s)^\top]dW(s). \end{aligned}$$

Note that for any C^2 function $\varphi(\cdot)$, one has

$$\begin{aligned} \varphi(x) - \varphi(\bar{x}) &= \left[\int_0^1 \varphi_x(\bar{x} + \alpha(x - \bar{x})) d\alpha \right] (x - \bar{x}) \\ &= \varphi_x(\bar{x})(x - \bar{x}) + \left\langle \int_0^1 \int_0^1 \left(\varphi_{xx}(\bar{x} + \alpha\beta(x - \bar{x})) \alpha d\alpha d\beta \right) (x - \bar{x}), x - \bar{x} \right\rangle \\ &= \varphi_x(\bar{x})(x - \bar{x}) + \frac{1}{2} \langle \varphi_{xx}(\bar{x})(x - \bar{x}), x - \bar{x} \rangle \\ &\quad + \left\langle \left(\int_0^1 \int_0^1 [\varphi_{xx}(\bar{x} + \alpha\beta(x - \bar{x})) - \varphi_{xx}(\bar{x})] \alpha d\alpha d\beta \right) (x - \bar{x}), x - \bar{x} \right\rangle \end{aligned}$$

Hence, one has

$$\begin{aligned}
 & J(t, \bar{X}(t); u^\varepsilon(\cdot)) - J(t, \bar{X}(t); \bar{u}(\cdot)) \\
 &= \mathbb{E}_t \left[h(t, X^\varepsilon(T)) - h(t, \bar{X}(T)) + \int_t^T [g(t, s, X^\varepsilon(s), u^\varepsilon(s)) - g(t, s, \bar{X}(s), \bar{u}(s))] ds \right] \\
 &= \mathbb{E}_t \left\{ h_x(t) \widehat{X}^\varepsilon(T) + \frac{1}{2} \langle h_{xx}(t) \widehat{X}^\varepsilon(T), \widehat{X}^\varepsilon(T) \rangle \right. \\
 &\quad + \left\langle \left(\int_0^1 \int_0^1 [h_{xx}(t, \bar{X}(T) + \alpha\beta \widehat{X}^\varepsilon(T)) - h_{xx}(t)] \alpha d\alpha d\beta \right) \widehat{X}^\varepsilon(T), \widehat{X}^\varepsilon(T) \right\rangle \\
 &\quad + \int_t^T \left[\widehat{g}(t, s) \mathbf{1}_{[t, t+\varepsilon]}(s) + g_x^\varepsilon(t, s) \widehat{X}^\varepsilon(s) + \frac{1}{2} \langle g_{xx}^\varepsilon(t, s) \widehat{X}^\varepsilon(s), \widehat{X}^\varepsilon(s) \rangle \right. \\
 &\quad \left. + \left\langle \left(\int_0^1 \int_0^1 g_{xx}(t, s, \bar{X}(s) + \alpha\beta \widehat{X}^\varepsilon(s), u^\varepsilon(s)) \right. \right. \right. \\
 &\quad \quad \left. \left. \left. - g_{xx}(t, s, \bar{X}(s), u^\varepsilon(s)) \right] \alpha d\alpha d\beta \right) \widehat{X}^\varepsilon(s), \widehat{X}^\varepsilon(s) \right\rangle ds \Big\} \\
 &= \mathbb{E}_t \left\{ \int_t^{t+\varepsilon} \widehat{g}(t, s) ds + h_x(t) \widehat{X}^\varepsilon(T) + \int_t^T g_x(t, s) \widehat{X}^\varepsilon(s) ds \right. \\
 &\quad + \frac{1}{2} \text{tr} \left(h_{xx}(t) [\widehat{X}^\varepsilon(T) \widehat{X}^\varepsilon(T)^\top] \right) + \frac{1}{2} \int_t^T \text{tr} \left(g_{xx}(t, s) [\widehat{X}^\varepsilon(s) \widehat{X}^\varepsilon(s)^\top] \right) ds \\
 &\quad + \left\langle \left(\int_0^1 \int_0^1 [h_{xx}(t, \bar{X}(T) + \alpha\beta \widehat{X}^\varepsilon(T)) - h_{xx}(t)] \alpha d\alpha d\beta \right) \widehat{X}^\varepsilon(T), \widehat{X}^\varepsilon(T) \right\rangle \\
 &\quad + \int_t^{t+\varepsilon} \left[\widehat{g}_x(t, s) \widehat{X}^\varepsilon(s) + \frac{1}{2} \text{tr} \left(\widehat{g}_{xx}(t, s) [\widehat{X}^\varepsilon(s) \widehat{X}^\varepsilon(s)^\top] \right) \right] ds \\
 &\quad + \int_t^T \left\langle \left(\int_0^1 \int_0^1 g_{xx}(t, s, \bar{X}(s) + \alpha\beta \widehat{X}^\varepsilon(s), u^\varepsilon(s)) \right. \right. \\
 &\quad \quad \left. \left. - g_{xx}(t, s, \bar{X}(s), u^\varepsilon(s)) \right] \alpha d\alpha d\beta \right) \widehat{X}^\varepsilon(s), \widehat{X}^\varepsilon(s) \right\rangle ds \Big\},
 \end{aligned}$$

where

$$\begin{aligned}
 h_x(t) &= h_x(t, \bar{X}(T)), & h_{xx}(t) &= h_{xx}(t, \bar{X}(T)), \\
 g_x^\varepsilon(t, s) &= g_x(t, s, \bar{X}(s), u^\varepsilon(s)), & g_{xx}^\varepsilon(t, s) &= g_{xx}(t, s, \bar{X}(s), u^\varepsilon(s)), \\
 g_x(t, s) &= g_x(t, s, \bar{X}(s), u), & g_{xx}(t, s) &= g_{xx}(t, s, \bar{X}(s), u), \\
 \widehat{g}(t, s) &= g(t, s, \bar{X}(s), u) - g(t, s, \bar{X}(s), \bar{u}(s)), \\
 \widehat{g}_x(t, s) &= g_x(t, s, \bar{X}(s), u) - g_x(t, s, \bar{X}(s), \bar{u}(s)).
 \end{aligned}$$

We now estimate some terms.

$$\begin{aligned}
 & \mathbb{E}_t \left| \left\langle \left(\int_0^1 \int_0^1 [h_{xx}(t, \bar{X}(T) + \alpha\beta \widehat{X}^\varepsilon(T)) - h_{xx}(t)] \alpha d\alpha d\beta \right) \widehat{X}^\varepsilon(T), \widehat{X}^\varepsilon(T) \right\rangle \right| \\
 & \leq \mathbb{E}_t |\widehat{X}^\varepsilon(T)|^3 \leq K\varepsilon^{\frac{3}{2}},
 \end{aligned}$$

$$\begin{aligned} & \mathbb{E}_t \left| \int_t^{t+\varepsilon} \left[\widehat{g}_x(t,s)\widehat{X}^\varepsilon(s) + \frac{1}{2} \text{tr} \left(\widehat{g}_{xx}(t,s)[\widehat{X}^\varepsilon(s)\widehat{X}^\varepsilon(s)^\top] \right) \right] ds \right| \\ & \leq K \mathbb{E}_t \int_t^{t+\varepsilon} |\widehat{X}^\varepsilon(s)| ds \leq K \varepsilon \mathbb{E}_t \left[\sup_{s \in [t,T]} |\widehat{X}^\varepsilon(s)| \right] \leq K \varepsilon^{\frac{3}{2}}, \\ & \mathbb{E}_t \left| \int_t^T \left\langle \left(\int_0^1 \int_0^1 g_{xx}(t,s,\bar{X}(s) + \alpha\beta\widehat{X}^\varepsilon(s), u^\varepsilon(s)) \right. \right. \right. \\ & \quad \left. \left. \left. - g_{xx}(t,s,\bar{X}(s), u^\varepsilon(s)) \right) \alpha d\alpha d\beta \right) \widehat{X}^\varepsilon(s), \widehat{X}^\varepsilon(s) \right\rangle ds \right| \\ & \leq K \mathbb{E}_t \int_t^T |\widehat{X}^\varepsilon(s)|^3 ds \leq K \varepsilon^{\frac{3}{2}}. \end{aligned}$$

Moreover,

$$\begin{aligned} & [\widehat{X}^\varepsilon(\cdot)\widehat{X}^\varepsilon(\cdot)^\top] = [X_1^\varepsilon(\cdot) + X_2^\varepsilon(\cdot)][X_1^\varepsilon(\cdot) + X_2^\varepsilon(\cdot)]^\top \\ & = [X_1^\varepsilon(\cdot)X_1^\varepsilon(\cdot)^\top] + [X_1^\varepsilon(\cdot)X_2^\varepsilon(\cdot)^\top] + [X_2^\varepsilon(\cdot)X_1^\varepsilon(\cdot)^\top] + [X_2^\varepsilon(\cdot)X_2^\varepsilon(\cdot)^\top] \\ & = \mathbb{X}^\varepsilon(\cdot) + O(\varepsilon^{\frac{3}{2}}). \end{aligned}$$

Consequently, we obtain

$$\begin{aligned} & J(t, \bar{X}(t); u^\varepsilon(\cdot)) - J(t, \bar{X}(t); \bar{u}(\cdot)) \\ & = \mathbb{E}_t \left\{ \int_t^{t+\varepsilon} \widehat{g}(t,s) ds + h_x(t)\widehat{X}^\varepsilon(T) + \int_t^T g_x(t,s)\widehat{X}^\varepsilon(s) ds \right. \\ & \quad \left. + \frac{1}{2} \left[\text{tr} \left(h_{xx}(t)\mathbb{X}^\varepsilon(T) \right) + \int_t^T \text{tr} \left(g_{xx}(t,s)\mathbb{X}^\varepsilon(s) \right) ds \right] \right\} + O(\varepsilon^{\frac{3}{2}}). \end{aligned}$$

Let $(Y^\varepsilon(\cdot), Z^\varepsilon(\cdot)) \equiv (Y^\varepsilon(\cdot; t), Z^\varepsilon(\cdot; t))$ be the adapted solution to the following:

$$\begin{cases} Y^\varepsilon(s) = -[b_x^\varepsilon(s)^\top Y^\varepsilon(s) + \sigma_x^\varepsilon(s)^\top Z^\varepsilon(s) + g_x(t,s)^\top] ds + Z^\varepsilon(s) dW(s), & s \in [t, T], \\ Y^\varepsilon(T) = h_x(t)^\top. \end{cases}$$

Then

$$\begin{aligned} & \mathbb{E}_t [h_x(t)\widehat{X}^\varepsilon(T)] = \mathbb{E}_t [\langle Y^\varepsilon(T), \widehat{X}^\varepsilon(T) \rangle] \\ & = \mathbb{E}_t \left[\int_t^T \left(- \langle b_x^\varepsilon(s)^\top Y^\varepsilon(s) + \sigma_x^\varepsilon(s)^\top Z^\varepsilon(s) + g_x(t,s)^\top, \widehat{X}^\varepsilon(s) \rangle \right. \right. \\ & \quad \left. \left. + \langle Y^\varepsilon(s), b_x^\varepsilon(s)\widehat{X}^\varepsilon(s) + \widehat{b}^\varepsilon(s) \rangle + \langle Z^\varepsilon(s), \sigma_x^\varepsilon(s)\widehat{X}^\varepsilon(s) + \widehat{\sigma}^\varepsilon(s) \rangle \right) ds \right] \\ & = \mathbb{E}_t \left[\int_t^T \left(- g_x(t,s)\widehat{X}^\varepsilon(s) + \langle Y^\varepsilon(s), \widehat{b}^\varepsilon(s) \rangle + \langle Z^\varepsilon(s), \widehat{\sigma}^\varepsilon(s) \rangle \right) ds \right]. \end{aligned}$$

Hence, we obtain

$$\begin{aligned}
& \mathbb{E}_t \left(h_x(t) \widehat{X}^\varepsilon(T) + \int_t^T g_x(t,s) \widehat{X}^\varepsilon(s) ds \right) \\
&= \mathbb{E}_t \left[\int_t^T \left(\langle Y^\varepsilon(s), \widehat{b}^\varepsilon(s) \rangle + \langle Z^\varepsilon(s), \widehat{\sigma}^\varepsilon(s) \rangle \right) ds \right] \\
&= \mathbb{E}_t \left[\int_t^{t+\varepsilon} \left(\langle Y^\varepsilon(s), \widehat{b}^\varepsilon(s) \rangle + \langle Z^\varepsilon(s), \widehat{\sigma}^\varepsilon(s) \rangle \right) ds \right].
\end{aligned}$$

On the other hand, we let $(P^\varepsilon(\cdot), \Lambda^\varepsilon(\cdot)) \equiv (P^\varepsilon(\cdot; t), \Lambda^\varepsilon(\cdot; t))$ be the adapted solution of the following BSDE:

$$\begin{cases} dP^\varepsilon(s) = - \left(b_x^\varepsilon(s)^\top P^\varepsilon(s) + P^\varepsilon(s) b_x^\varepsilon(s) + \sigma_x^\varepsilon(s)^\top P^\varepsilon(s) \sigma_x^\varepsilon(s) \right. \\ \quad \left. + \sigma_x^\varepsilon(s)^\top \Lambda^\varepsilon(s) + \Lambda^\varepsilon(s) \sigma_x^\varepsilon(s) + g_{xx}(t) \right) ds + \Lambda^\varepsilon(s) dW(s), \quad s \in [t, T], \\ P^\varepsilon(T) = h_{xx}(t, \bar{X}(T)). \end{cases}$$

One has

$$\begin{aligned}
& \mathbb{E}_t \left\{ \text{tr} \left[P^\varepsilon(T) \mathbb{X}^\varepsilon(T) \right] \right\} = \mathbb{E}_t \left\{ \int_t^T \left[\text{tr} \left(- [b_x^\varepsilon(s)^\top P^\varepsilon(s) + P^\varepsilon(s) b_x^\varepsilon(s) \right. \right. \right. \\
& \quad \left. \left. + \sigma_x^\varepsilon(s)^\top P^\varepsilon(s) \sigma_x^\varepsilon(s) + \sigma_x^\varepsilon(s)^\top \Lambda^\varepsilon(s) + \Lambda^\varepsilon(s) \sigma_x^\varepsilon(s) + g_{xx}(t) \right] \mathbb{X}^\varepsilon(s) \right. \\
& \quad \left. + P^\varepsilon(s) [b_x^\varepsilon(s) \mathbb{X}^\varepsilon(s) + \mathbb{X}^\varepsilon(s) b_x^\varepsilon(s)^\top + \sigma_x^\varepsilon(s) \mathbb{X}^\varepsilon(s) \sigma_x^\varepsilon(s)^\top \right. \\
& \quad \left. + \sigma_x^\varepsilon(s) X_1^\varepsilon(s) \widehat{\sigma}^\varepsilon(s)^\top + \widehat{\sigma}^\varepsilon(s) X_1^\varepsilon(s)^\top \sigma_x^\varepsilon(s)^\top + \widehat{\sigma}^\varepsilon(s) \widehat{\sigma}^\varepsilon(s)^\top] \right. \\
& \quad \left. + \Lambda^\varepsilon(s) [\sigma_x^\varepsilon(s) \mathbb{X}^\varepsilon(s) + \mathbb{X}^\varepsilon(s) \sigma_x^\varepsilon(s)^\top] \right) ds \Big\} \\
&= \mathbb{E}_t \left\{ \int_t^T \left[\text{tr} \left(- g_{xx}(t) \mathbb{X}^\varepsilon(s) + P^\varepsilon(s) [\sigma_x^\varepsilon(s) X_1^\varepsilon(s) \widehat{\sigma}^\varepsilon(s)^\top \right. \right. \right. \\
& \quad \left. \left. + \widehat{\sigma}^\varepsilon(s) X_1^\varepsilon(s)^\top \sigma_x^\varepsilon(s)^\top + \widehat{\sigma}^\varepsilon(s) \widehat{\sigma}^\varepsilon(s)^\top] \right) ds \right\}.
\end{aligned}$$

Note that under our conditions, $P^\varepsilon(\cdot)$ is bounded. Then

$$\begin{aligned}
& \mathbb{E}_t \left| \int_t^T \left[\text{tr} \left(P^\varepsilon(s) [\sigma_x^\varepsilon(s) X_1^\varepsilon(s) \widehat{\sigma}^\varepsilon(s)^\top + \widehat{\sigma}^\varepsilon(s) X_1^\varepsilon(s)^\top \sigma_x^\varepsilon(s)^\top] \right) \right] ds \right| \\
&\leq K \mathbb{E}_t \int_t^{t+\varepsilon} |X_1^\varepsilon(s)| ds \leq K \varepsilon \mathbb{E}_t \left[\sup_{s \in [t, T]} |X_1^\varepsilon(s)| \right] \leq K \varepsilon^{\frac{3}{2}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \mathbb{E}_t \left\{ \text{tr} \left[P^\varepsilon(T) \mathbb{X}^\varepsilon(T) \right] + \int_t^T \text{tr} \left(g_{xx}(t) \mathbb{X}^\varepsilon(s) \right) ds \right\} \\
&= \mathbb{E}_t \int_t^{t+\varepsilon} \text{tr} \left(\widehat{\sigma}^\varepsilon(s) P^\varepsilon(s) \widehat{\sigma}^\varepsilon(s) \right) ds + O(\varepsilon^{\frac{3}{2}}).
\end{aligned}$$

We now obtain

$$\begin{aligned}
 & J(t, \bar{X}(t); u^\varepsilon(\cdot)) - J(t, \bar{X}(t); \bar{u}(\cdot)) \\
 &= \mathbb{E}_t \left[\int_t^{t+\varepsilon} \left(\widehat{g}(t, s) + \langle Y^\varepsilon(s), \widehat{b}(s) \rangle + \langle Z^\varepsilon(s), \widehat{\sigma}(s) \right) \right. \\
 &\qquad \qquad \qquad \left. + \text{tr} [\widehat{\sigma}(s)^\top P^\varepsilon(s) \widehat{\sigma}(s)] \right) ds \Big] + O(\varepsilon^{\frac{3}{2}}).
 \end{aligned}$$

Further,

$$\begin{aligned}
 & J(t, \bar{X}(t); u^\varepsilon(\cdot)) - J(t, \bar{X}(t); \bar{u}(\cdot)) \\
 &= \mathbb{E}_t \left[\int_t^{t+\varepsilon} \left(\langle Y(s), \widehat{b}(s) \rangle + \langle Z(s), \widehat{\sigma}(s) \rangle + \widehat{g}(t, s) \right) \right. \\
 &\qquad \qquad \qquad \left. + \text{tr} [\widehat{\sigma}(s)^\top P(s) \widehat{\sigma}(s)] \right) ds \Big] + o(\varepsilon),
 \end{aligned}$$

with $(P(\cdot), \Lambda(\cdot))$ and $(Y(\cdot), Z(\cdot))$ satisfy (49) and (50). Then our conclusion follows. □

5 Differential Game Approach

In this section, we consider state equation (1). For any initial pair $(t, x) \in [0, T] \times \mathbb{R}^n$ and control $u(\cdot) \in \mathcal{U}^P[t, T]$, the state process is the solution $X(\cdot) \equiv X(\cdot; t, x, u(\cdot))$ of (1). The cost functional is given by the following:

$$(52) \qquad J(t, x; u(\cdot)) = Y(t),$$

with $(Y(\cdot), Z(\cdot, \cdot))$ being the adapted solution to the following BSVIE:

$$\begin{aligned}
 (53) \qquad Y(t) &= h(t, X(t), X(T)) + \int_t^T g(t, s, X(t), X(s), u(s), Y(s), Z(t, s)) ds \\
 &\qquad - \int_t^T Z(t, s) dW(s), \qquad t \in [0, T].
 \end{aligned}$$

We rewrite the optimal control problem as follows.

Problem (N). For given $(t, x) \in [0, T] \times \mathbb{R}^n$, find a $\bar{u}(\cdot) \in \mathcal{U}^P[t, T]$ such that

$$J(t, x; \bar{u}(\cdot)) = \inf_{u(\cdot) \in \mathcal{U}^P[t, T]} J(t, x; u(\cdot)).$$

It is not hard to see that this problem is time-inconsistent. We now use a differential game method to investigate Problem (N). Note that it is new to use differential game approach to find closed-loop equilibrium strategy for the problem with the cost functional being the solution to a BSVIE. Recall that in [112],

Let $\Pi : 0 = t_0 < t_1 < t_2 < \dots < t_{N-1} < t_N = T$ be a partition of $[0, T]$. We denote

$$\|\Pi\| = \max_{1 \leq i \leq N} (t_i - t_{i-1}),$$

which is called the *mesh size* of Π . Associated with the above partition, we have an N -person differential game. Player k takes over the the system (1) at t_{k-1} , controls the system on $[t_{k-1}, t_k)$, and hands over to Player $(k + 1)$ at t_k . The cost functionals of the players are constructed recursively. We now make it precise.

We first look at Player N on $[t_{N-1}, t_N]$. The state equation is (1) on $[t_{N-1}, t_N]$, and the cost functional is given by the following:

$$(54) \quad J^N(t, x; u(\cdot)) = Y^N(t),$$

where $(Y^N(\cdot), Z^N(\cdot))$ is the adapted solution to the following BSDE:

$$(55) \quad \begin{aligned} Y^N(t) &= h(t_{N-1}, x_{N-1}, X(T)) + \int_t^T g(t_{N-1}, s, x_{N-1}, X(s), u(s), Y^N(s), Z^N(s)) ds \\ &\quad - \int_t^T Z^N(s) dW(s), \quad t \in [t_{N-1}, t_N], \end{aligned}$$

with $x_{N-1} \in \mathbb{R}^n$ being a parameter. Thus, minimization of the cost functional $J^N(t, x; \cdot)$ subject to the state equation (1) on $[t_{N-1}, t_N]$ is a standard optimal control problem with a recursive cost functional. For such a problem, the value function $V^N(\cdot, \cdot)$ satisfies the following partial differential equation (PDE, for short), called HJB equation (in a suitable sense):

$$(56) \quad \begin{cases} V_t^N(t, x) + H(t_{N-1}, t, x_{N-1}, x, V^N(t, x), V_x^N(t, x), V_{xx}^N(t, x)) = 0, \\ \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad (t, x) \in [t_{N-1}, t_N] \times \mathbb{R}^n, \\ V^N(t_N, x) = h(t_{N-1}, x_{N-1}, x), \quad x \in \mathbb{R}^n, \end{cases}$$

where

$$\begin{aligned} \mathbb{H}(\tau, \xi, t, x, u, y, \mathbf{p}^\top, \mathbf{P}) &= \frac{1}{2} \text{tr} \left(\mathbf{P} \sigma(t, x, u) \sigma(t, x, u)^\top \right) + \mathbf{p}^\top b(t, x, u) \\ &\quad + g(\tau, t, \xi, x, u, y, \mathbf{p}^\top \sigma(t, x, u)), \end{aligned}$$

and for some $\psi \equiv \psi(\tau, t, \xi, x, y, \mathbf{p}^\top, \mathbf{P})$:

$$(57) \quad H(\tau, t, \xi, x, y, \mathbf{p}^\top, \mathbf{P}) = \inf_{u \in U} \mathbb{H}(\tau, t, \xi, x, u, y, \mathbf{p}^\top, \mathbf{P}).$$

We assume that the map $\psi(\cdot)$ has all needed smoothness (and boundedness of the derivatives). Then under the non-degeneracy condition of the diffusion, the above HJB equation admits a unique classical solution $V^N(\cdot)$, and

$$\begin{aligned} \bar{u}^N(s) &= \psi(t_{N-1}, s, X(t_{N-1}), \bar{X}^N(s), V^N(s, \bar{X}^N(s)), V_x^N(s, \bar{X}^N(s)), V_{xx}^N(s, \bar{X}^N(s))), \\ &\equiv \Psi^N(s, \bar{X}^N(s)), \quad s \in [t_{N-1}, t_N], \end{aligned}$$

is an optimal control of the problem on $[t_{N-1}, t_N]$, where $\bar{X}^N(\cdot)$ is the solution to the following closed-loop system:

$$(58) \quad \begin{cases} d\bar{X}^N(s) = b(s, \bar{X}^N(s), \Psi^N(s, \bar{X}^N(s)))ds + \sigma(s, \bar{X}^N(s), \Psi^N(s, \bar{X}^N(s)))dW(s), \\ s \in [t_{N-1}, t_N], \\ \bar{X}^N(t_{N-1}) = x_{N-1}. \end{cases}$$

We now look at Player $(N - 1)$ on $[t_{N-2}, t_{N-1}]$. The state process, denoted by $X^{N-1}(\cdot) \equiv X(\cdot; t, x, u^{N-1}(\cdot))$, will be the solution to equation (1), with $(t, x) \in [t_{N-2}, t_{N-1}] \times \mathbb{R}^n$, and $u^{N-1}(\cdot) \in \mathcal{U}[t, t_{N-1}]$. For the cost functional, we first solve the following *forward-backward stochastic differential equation* (FBSDE, for short) on $[t_{N-1}, t_N]$:

$$(59) \quad \begin{cases} dX^{N-1}(s) = b(s, X^{N-1}(s), \Psi^N(s, X^{N-1}(s)))ds \\ \quad \quad \quad + \sigma(s, X^{N-1}(s), \Psi^N(s, X^{N-1}(s)))dW(s), \\ dY^{N-1}(s) = -g(t_{N-2}, s, x_{N-2}, X^{N-1}(s), \Psi^N(s, X^{N-1}(s)), Y^{N-1}(s), Z^{N-1}(s))ds \\ \quad \quad \quad + Z^{N-1}(s)dW(s), \quad s \in [t_{N-1}, t_N], \\ X^{N-1}(t_{N-1}) = x_{N-1}, \quad Y^{N-1}(t_N) = h(t_{N-2}, X^{N-1}(t_N)). \end{cases}$$

Following [63], we know that

$$Y^{N-1}(s) = \Theta^{N-1}(s, X^{N-1}(s)), \quad s \in [t_{N-1}, t_N],$$

with $\Theta^{N-1}(\cdot, \cdot)$ being the solution to the following PDE:

$$(60) \quad \begin{cases} \Theta_s^{N-1}(s, x) + \frac{1}{2} \text{tr} \left[\Theta_{xx}^{N-1}(s, x) (\sigma \sigma^\top)(s, x, \Psi^N(s, x)) \right] \\ \quad + \Theta_x^{N-1}(s, x) b(s, x, \Psi^N(s, x)) \\ \quad + g(t_{N-2}, s, x_{N-2}, x, \Theta^{N-1}(s, x), \Theta_x^{N-1}(s, x) \sigma(s, x, \Psi^N(s, x))) = 0, \\ \quad \quad \quad (s, x) \in [t_{N-1}, t_N] \times \mathbb{R}^n, \\ \Theta^{N-1}(t_N, x) = h(t_{N-2}, x), \quad x \in \mathbb{R}^n. \end{cases}$$

Then let $(Y^{N-1}(\cdot), Z^{N-1}(\cdot))$ be the adapted solution to the following BSDE on $[t_{N-2}, t_{N-1}]$:

$$(61) \quad \begin{cases} dY^{N-1}(s) = -g(t_{N-2}, s, x_{N-2}, X^{N-1}(s), Y^{N-1}(s), Z^{N-1}(s))ds \\ \quad \quad \quad + Z^{N-1}(s)dW(s), \quad s \in [t_{N-2}, t_{N-1}], \\ Y^{N-1}(t_{N-1}) = \Theta^{N-1}(t_{N-1}, X^{N-1}(t_{N-1})), \end{cases}$$

and define

$$(62) \quad J^{N-1}(t, x; u(\cdot)) = Y^{N-1}(t), \quad (t, x) \in [t_{N-2}, t_{N-1}] \times \mathbb{R}^n, u(\cdot) \in \mathcal{U}[t, t_{N-1}].$$

The above is referred to as a *sophisticated cost functional* which is defined through an FBSDE (59) over $[t_{N-1}, t_N]$, a *representation PDE* (60) on $[t_{N-1}, t_N] \times \mathbb{R}^n$ and a BSDE (61) over $[t_{N-2}, t_{N-1}]$. It is not hard to see that the problem of minimizing cost functional $J^{N-1}(t, x; u(\cdot))$ subject to the state equation (1) restricted on $[t_{N-2}, t_{N-1}]$ is a standard stochastic optimal control problem with recursive cost functional. For this problem, the value function $V^{N-1}(\cdot, \cdot)$ satisfies the following HJB equation:

$$(63) \quad \begin{cases} V_t^{N-1}(t, x) + H(t_{N-2}, t, x_{N-2}, x, V^{N-1}(t, x), V_x^{N-1}(t, x), V_{xx}^{N-1}(t, x)) = 0, \\ \hspace{15em} (t, x) \in [t_{N-2}, t_{N-1}] \times \mathbb{R}^n, \\ V^{N-1}(t_{N-1}, x) = \Theta^{N-1}(t_{N-1}, x_{N-1}, x), \quad x \in \mathbb{R}^n. \end{cases}$$

We assume that the above admits a classical solution $V^{N-1}(\cdot, \cdot)$. Then

$$\begin{aligned} \bar{u}^{N-1}(s) &= \Psi(t_{N-2}, s, x_{N-2}, \bar{X}^{N-1}(s), V^{N-1}(s, \bar{X}^{N-1}(s)), \\ &\hspace{10em} V_x^{N-1}(s, \bar{X}^{N-1}(s)), V_{xx}^{N-1}(s, \bar{X}^{N-1}(s))), \\ &\equiv \Psi^{N-1}(s, \bar{X}^{N-1}(s)), \quad s \in [t_{N-2}, t_{N-1}], \end{aligned}$$

is an optimal control of the problem on $[t_{N-2}, t_{N-1}]$, where $\bar{X}^{N-1}(\cdot)$ is the solution to the following closed-loop system:

$$(64) \quad \begin{cases} d\bar{X}^{N-1}(s) = b(s, \bar{X}^{N-1}(s), \Psi^{N-1}(s, \bar{X}^{N-1}(s)))ds \\ \hspace{10em} + \sigma(s, \bar{X}^{N-1}(s), \Psi^{N-1}(s, \bar{X}^{N-1}(s)))dW(s), \quad s \in [t_{N-2}, t_{N-1}], \\ \bar{X}^N(t_{N-2}) = x_{N-2}. \end{cases}$$

Note that Player $(N - 1)$ knows that Player N will play optimally through the optimal strategy $\Psi^N(\cdot, \cdot)$ (defined on $[t_{N-1}, t_N] \times \mathbb{R}^n$). On the other hand, Player $(N - 1)$ still “discounts” the future costs in his/her own way despite he/she will not control the system beyond t_{N-1} . That is why t_{N-2} appears in (59), (60), and (61). See [112, 104] for more detailed explanations.

By induction, we can continue the above backward procedure. Let us look at the situation of Player k . The state process $X^k(\cdot) \equiv X(\cdot; t, x, u^k(\cdot))$ is the solution to the equation (1) with $(t, x) \in [t_{k-1}, t_k] \times \mathbb{R}^n$ and $u(\cdot) \in \mathcal{U}[t, t_k]$. To define the sophisticated cost functional, we solve the following decoupled FBSDE on $[t_k, t_N]$:

$$(65) \quad \begin{cases} dX^k(s) = b(s, X^k(s), \Psi^\Pi(s, X^k(s)))ds \\ \hspace{10em} + \sigma(s, X^k(s), \Psi^\Pi(s, X^k(s)))dW(s), \quad s \in [t_k, t_N], \\ dY^k(s) = -g(t_{k-1}, s, X^k(s), \Psi^\Pi(s, X^k(s)), Y^k(s), Z^k(s))ds \\ \hspace{10em} + Z^k(s)dW(s), \quad s \in [t_k, t_N], \\ X^k(t_k) = x_k, \quad Y^k(t_N) = h(t_{k-1}, X^k(t_N)), \end{cases}$$

where,

$$(66) \quad \Psi^\Pi(s, x) = \sum_{j=k+1}^N \psi(t_{j-1}, s, x_{j-1}, x, V^j(s, x), V_x^j(s, x), V_{xx}^j(s, x)) \mathbf{1}_{[t_{j-1}, t_j)}(s).$$

Suppose $(X^k(\cdot), Y^k(\cdot), Z^k(\cdot))$ is the adapted solution to the above FBSDE. Then we have the following representation:

$$Y^k(s) = \Theta^k(s, X^k(s)), \quad s \in [t_k, t_N],$$

with $\Theta^k(\cdot, \cdot)$ being the solution to the following PDE:

$$(67) \quad \begin{cases} \Theta_s^k(s, x) + \frac{1}{2} \operatorname{tr} \left[\Theta_{xx}^k(s, x) (\sigma \sigma^\top)(s, x, \Psi^\Pi(s, x)) \right] + \Theta_x^k(s, x) b(s, x, \Psi^\Pi(s, x)) \\ \quad + g(t_{k-1}, s, x_{k-1}, x, \Theta^k(s, x), \Theta_x^k(s, x) \sigma(s, x, \Psi^\Pi(s, x))) = 0, \\ \quad \quad \quad (s, x) \in [t_k, t_N] \times \mathbb{R}^n, \\ \Theta^k(t_N, x) = h(t_{k-1}, x_{k-1}, x), \quad x \in \mathbb{R}^n. \end{cases}$$

Then let $(Y^k(\cdot), Z^k(\cdot))$ be the adapted solution to the following BSDE on $[t_{k-1}, t_k]$:

$$(68) \quad \begin{cases} dY^k(s) = -g(t_{k-1}, s, x_{k-1}, X^k(s), Y^k(s), Z^k(s)) ds + Z^k(s) dW(s), \\ \quad \quad \quad s \in [t_{k-1}, t_k], \\ Y^k(t_k) = \Theta^k(t_k, X^k(t_k)), \end{cases}$$

and define the sophisticated cost functional by

$$(69) \quad J^k(t, x; u(\cdot)) = Y^k(t), \quad (t, x) \in [t_{k-1}, t_k] \times \mathbb{R}^n, u(\cdot) \in \mathcal{U}[t, t_k].$$

Then, we obtain an optimal control problem on $[t_{k-1}, t_k]$ with a recursive cost functional. The value function $V^k(\cdot, \cdot)$ of this problem satisfies the following HJB equation:

$$(70) \quad \begin{cases} V_t^k(t, x) + H(t_{k-1}, t, x_{k-1}, x, V^k(t, x), V_x^k(t, x), V_{xx}^k(t, x)) = 0, \\ \quad \quad \quad (t, x) \in [t_{k-1}, t_k] \times \mathbb{R}^n, \\ V^k(t_k, x) = \Theta^k(t_k, X^k, x), \quad x \in \mathbb{R}^n. \end{cases}$$

Suppose the above admits a classical solution $V^k(\cdot, \cdot)$. Then

$$\begin{aligned} \bar{u}^k(s) &= \psi(t_{k-1}, s, x_{k-1}, \bar{X}^k(s), V^k(s, \bar{X}^k(s)), V_x^k(s, \bar{X}^k(s)), V_{xx}^k(s, \bar{X}^k(s))), \\ &\equiv \Psi^k(s, \bar{X}^k(s)), \quad s \in [t_{k-1}, t_k], \end{aligned}$$

is an optimal control of the problem on $[t_{k-1}, t_k]$, where $\bar{X}^k(\cdot)$ is the solution to the following closed-loop system:

$$(71) \quad \begin{cases} d\bar{X}^k(s) = b(s, \bar{X}^k(s), \Psi^k(s, \bar{X}^k(s)))ds + \sigma(s, \bar{X}^k(s), \Psi^k(s, \bar{X}^k(s)))dW(s), \\ \bar{X}^k(t_{k-1}) = x_{k-1}. \end{cases} \quad s \in [t_{k-1}, t_k],$$

Hence, for any partition Π of $[0, T]$, we could find a $\Psi^\Pi(\cdot, \cdot)$, which is called an *approximate equilibrium strategy*. By letting $\|\Pi\| \rightarrow 0$, we formally obtain the following equation which is called the *equilibrium HJB equation* (comparing with (9) and (19)):

$$(72) \quad \begin{cases} \Theta_s(\tau, s, \xi, x) + \frac{1}{2} \text{tr} \left[\Theta_{xx}(\tau, s, \xi, x) (\sigma \sigma^\top)(s, x, \Psi(s, x)) \right] \\ + \Theta_x(t, s, \xi, x) b(s, x, \Psi(s, x)) \\ + g(\tau, s, \xi, x, \Theta(\tau, s, \xi, x), \Theta_x(\tau, s, \xi, x) \sigma(s, x, \Psi(s, x))) = 0, \\ 0 \leq \tau \leq s \leq T, \quad \xi, x \in \mathbb{R}^n, \\ \Theta(\tau, T, \xi, x) = h(\tau, \xi, x), \quad x \in \mathbb{R}^n. \end{cases}$$

The *equilibrium value function* is given by

$$(73) \quad V(t, x) = \Theta(t, t, x, x), \quad (t, x) \in [0, T] \times \mathbb{R}^n,$$

and an *equilibrium strategy* is given by

$$(74) \quad \Psi(t, x) = \psi(t, t, x, x, V(t, x), \Theta_x(t, t, \xi, x)|_{\xi=x}, \Theta_{xx}(t, t, \xi, x)|_{\xi=x}).$$

In the case that equilibrium HJB equation (72) is well-posed, the above convergence can be proved (following a similar idea from [112, 104, 61]). Moreover, by an argument in [104, 61], we can show that $\Psi(\cdot, \cdot)$ satisfies (46). This means that $\Psi(\cdot, \cdot)$ is an equilibrium strategy in the sense of Definition 1. For the well-posedness of equilibrium HJB equation (72), following the idea from [104], we can establish that when the control does not enter into the diffusion $\sigma(\cdot)$. The general case is still under investigation.

6 Equilibrium Strategy for Mixed Time-Inconsistent Optimal Control Problems

It is expected that the mixed time-inconsistent optimal control problem is much more complicated. To be complete, in this section, we present a special case to exhibit some flavor. The general case is still widely open. The following is a sketch of the results obtained in [113].

Consider the following controlled linear MF-SDE:

$$(75) \quad \begin{cases} dX(s) = [A(s)X(s) + B(s)u(s)]ds + [C(s)X(s) + D(s)u(s)]dW(s), \\ \hspace{20em} s \in [t, T], \\ X(t) = x \in \mathcal{X}_t \equiv L^2_{\mathcal{F}_t}(\Omega; \mathbb{R}^n), \end{cases}$$

Note that for any $(t, x) \in \mathcal{D} \equiv \{(t, x) \mid t \in [0, T], x \in \mathcal{X}_t\}$ and $u(\cdot) \in \mathcal{U}[t, T]$, the corresponding state process $X(\cdot) = X(\cdot; t, x, u(\cdot))$ depends on $(t, x, u(\cdot))$. The cost functional is as follows:

$$(76) \quad J(t, x; u(\cdot)) = \mathbb{E}_t \left\{ \int_t^T \left[\langle Q(s, t)X(s), X(s) \rangle + \langle \bar{Q}(s, t)\mathbb{E}_t[X(s)], \mathbb{E}_t[X(s)] \rangle \right. \right. \\ \left. \left. + \langle R(s, t)u(s), u(s) \rangle + \langle \bar{R}(s, t)\mathbb{E}_t[u(s)], \mathbb{E}_t[u(s)] \rangle \right] ds \right. \\ \left. + \langle G(t)X(T), X(T) \rangle + \langle \bar{G}(t)\mathbb{E}_t[X(T)], \mathbb{E}_t[X(T)] \rangle \right\}.$$

Let us introduce the following hypotheses:

(H4) The following hold:

$$(77) \quad A(\cdot), C(\cdot) \in C([0, T]; \mathbb{R}^{n \times n}), \quad B(\cdot), D(\cdot) \in C([0, T]; \mathbb{R}^{n \times m}).$$

(H5) The following hold:

$$(78) \quad \begin{cases} Q(\cdot, \cdot), \bar{Q}(\cdot, \cdot) \in C([0, T]^2; \mathbb{S}^n), & R(\cdot, \cdot), \bar{R}(\cdot, \cdot) \in C([0, T]^2; \mathbb{S}^m), \\ G(\cdot), \bar{G}(\cdot) \in C([0, T]; \mathbb{S}^n), \end{cases}$$

with \mathbb{S}^n being the set of all $(n \times n)$ symmetric matrices, and for some $\delta > 0$,

$$(79) \quad \begin{cases} Q(s, t), Q(s, t) + \bar{Q}(s, t) \geq 0, & 0 \leq t \leq s \leq T, \\ R(s, t), R(s, t) + \bar{R}(s, t) \geq \delta I, \\ G(t), G(t) + \bar{G}(t) \geq 0, & 0 \leq t \leq T. \end{cases}$$

(H6) The following monotonicity conditions are satisfied:

$$(80) \quad \begin{cases} Q(s, t) \leq Q(s, \tau), & Q(s, t) + \bar{Q}(s, t) \leq Q(s, \tau) + \bar{Q}(s, \tau), \\ R(s, t) \leq R(s, \tau), & R(s, t) + \bar{R}(s, t) \leq R(s, \tau) + \bar{R}(s, \tau), \\ G(t) \leq G(\tau), & G(t) + \bar{G}(t) \leq G(\tau) + \bar{G}(\tau), \\ & 0 \leq t \leq \tau \leq s \leq T. \end{cases}$$

It is clear that under (H4)–(H5), for any $(t, x) \in [0, T] \times \mathbb{R}^n$ and $u(\cdot) \in \mathcal{U}[t, T]$, state equation (75) admits a unique solution $X(\cdot) \equiv X(\cdot; t, x, u(\cdot))$, and the cost functional $J(t, x; u(\cdot))$ is well-defined. Then we can state the following problem.

Problem (MF-LQ). For any $(t, x) \in \mathcal{D}$, find a $u^*(\cdot) \in \mathcal{U}[t, T]$ such that

$$(81) \quad J(t, x; u^*(\cdot)) = \inf_{u(\cdot) \in \mathcal{U}[t, T]} J(t, x; u(\cdot)) \equiv V(t, x).$$

In what follows, we will denote

$$(82) \quad \begin{cases} \widehat{Q}(s, t) = Q(s, t) + \bar{Q}(s, t), & \widehat{R}(s, t) = R(s, t) + \bar{R}(s, t), \\ \widehat{G}(t) = G(t) + \bar{G}(t). \end{cases} \quad 0 \leq t \leq s \leq T.$$

We introduce the following: For any $t \in [0, T)$,

$$(83) \quad \begin{aligned} & \widetilde{J}(t; X(\cdot), u(\cdot)) \\ &= \mathbb{E}_t \left\{ \int_t^T \left[\langle Q(s, t)X(s), X(s) \rangle + \langle \bar{Q}(s, t)\mathbb{E}_t[X(s)], \mathbb{E}_t[X(s)] \rangle \right. \right. \\ & \quad \left. \left. + \langle R(s, t)u(s), u(s) \rangle + \langle \bar{R}(s, t)\mathbb{E}_t[u(s)], \mathbb{E}_t[u(s)] \rangle \right] ds \right. \\ & \quad \left. + \langle G(t)X(T), X(T) \rangle + \langle \bar{G}(t)\mathbb{E}_t[X(T)], \mathbb{E}_t[X(T)] \rangle \right\}, \end{aligned}$$

for any $(X(\cdot), u(\cdot)) \in \mathcal{X}[t, T] \times \mathcal{U}[t, T]$, where $\mathcal{X}[t, T] = L^2_{\mathbb{R}}(\Omega; C([t, T]; \mathbb{R}^n))$. We point out that in the above $(X(\cdot), u(\cdot))$ does not have to be a state-control pair of the original control system. Thus, $\widetilde{J}(t; X(\cdot), u(\cdot))$ is an extension of the cost functional $J(t, x; u(\cdot))$, and

$$\widetilde{J}(t; X(\cdot; t, x, u(\cdot)), u(\cdot)) = J(t, x; u(\cdot)), \quad \forall (t, x) \in \mathcal{D}, u(\cdot) \in \mathcal{U}[t, T].$$

Next, and hereafter, we denote any partition of $[0, T]$ by Π :

$$\Pi = \{t_k \mid 0 \leq k \leq N\} \equiv \{0 = t_0 < t_1 < t_2 < \dots < t_{N-1} < t_N = T\},$$

with N being some natural number, and define its *mesh size* by the following:

$$\|\Pi\| = \max_{0 \leq k \leq N-1} (t_{k+1} - t_k).$$

For the above Π , we define

$$(84) \quad J_k^\Pi(X(\cdot), u(\cdot)) = \widetilde{J}(t_k, X(\cdot), u(\cdot)),$$

for any $(X(\cdot), u(\cdot)) \in \mathcal{X}[t_k, T] \times \mathcal{U}[t_k, T]$, $k = 0, 1, 2, \dots, N - 1$.

Now, we introduce some notions.

Definition 2. Let $\Pi = \{0 = t_0 < t_1 < \dots < t_{N-1} < t_N = T\}$ be a partition of $[0, T]$, and let $\Theta^\Pi, \widehat{\Theta}^\Pi : [0, T] \rightarrow \mathbb{R}^{m \times n}$ be two given maps, possibly depending on Π .

(i) For any $x \in \mathbb{R}^n$ fixed, let $X^\Pi(\cdot) \equiv X^\Pi(\cdot; x)$ be the solution to the following linear MF-SDE:

$$\left\{ \begin{aligned} dX^\Pi(s) &= \left\{ [A(s) - B(s)\Theta^\Pi(s)]X^\Pi(s) \right. \\ &\quad + [\bar{A}(s) + B(s)[\Theta^\Pi(s) - \widehat{\Theta}^\Pi(s)]] \mathbb{E}_{\rho^\Pi(s)}[X^\Pi(s)] \Big\} ds \\ &\quad + \left\{ [C(s) - D(s)\Theta^\Pi(s)]X^\Pi(s) \right. \\ &\quad + [\bar{C}(s) + D(s)[\Theta^\Pi(s) - \widehat{\Theta}^\Pi(s)]] \mathbb{E}_{\rho^\Pi(s)}[X^\Pi(s)] \Big\} dW(s), \quad s \in [0, T], \\ X^\Pi(0) &= x, \end{aligned} \right.$$

where

$$\rho^\Pi(s) = \sum_{k=0}^{N-1} t_k I_{[t_k, t_{k+1})}(s), \quad s \in [0, T],$$

and let $u^\Pi(\cdot) \equiv u^\Pi(\cdot; x)$ be defined by

$$(85) \quad u^\Pi(s) = -\Theta^\Pi(s)X^\Pi(s) + [\Theta^\Pi(s) - \widehat{\Theta}^\Pi(\cdot)] \mathbb{E}_{\rho^\Pi(s)}[X^\Pi(s)], \quad s \in [0, T].$$

The pair $(X^\Pi(\cdot), u^\Pi(\cdot))$ is called the *closed-loop pair* associated with Π and $(\Theta^\Pi(\cdot), \widehat{\Theta}^\Pi(\cdot))$, starting from x .

(ii) For each $t_k \in \Pi$ and any $u_k(\cdot) \in \mathcal{U}[t_k, t_{k+1}]$, let $X_k(\cdot)$ be the solution to the following system:

$$\left\{ \begin{aligned} dX_k(s) &= [A(s)X_k(s) + B(s)u_k(s)] ds + [C(s)X_k(s) + D(s)u_k(s)] dW(s), \\ &\hspace{20em} s \in [t_k, t_{k+1}], \\ X_k(t_k) &= X^\Pi(t_k), \end{aligned} \right.$$

and $X_{k+1}^\Pi(\cdot)$ be the solution to the following:

$$\left\{ \begin{aligned} dX_{k+1}^\Pi(s) &= \left\{ [A(s) - B(s)\Theta^\Pi(s)]X_{k+1}^\Pi(s) \right. \\ &\quad + [B(s)[\Theta^\Pi(s) - \widehat{\Theta}^\Pi(s)]] \mathbb{E}_{\rho^\Pi(s)}[X_{k+1}^\Pi(s)] \Big\} ds \\ &\quad + \left\{ [C(s) - D(s)\Theta^\Pi(s)]X_{k+1}^\Pi(s) \right. \\ &\quad + [D(s)[\Theta^\Pi(s) - \widehat{\Theta}^\Pi(s)]] \mathbb{E}_{\rho^\Pi(s)}[X_{k+1}^\Pi(s)] \Big\} dW(s), \quad s \in [t_{k+1}, T], \\ X_{k+1}^\Pi(t_{k+1}) &= X_k(t_{k+1}). \end{aligned} \right.$$

Denote

$$\left\{ \begin{aligned} X_k(\cdot) \oplus X^\Pi(\cdot) &\equiv X_k(\cdot) I_{[t_k, t_{k+1})}(\cdot) + X_{k+1}^\Pi(\cdot) I_{[t_{k+1}, T]}(\cdot), \\ u_k(\cdot) \oplus u^\Pi(\cdot) &= u_k(\cdot) I_{[t_k, t_{k+1})}(\cdot) \\ &\quad - \{ \Theta^\Pi(\cdot) X_{k+1}^\Pi(\cdot) + [\Theta^\Pi(\cdot) - \widehat{\Theta}^\Pi(\cdot)] \mathbb{E}_{\rho^\Pi(\cdot)}[X_{k+1}^\Pi(\cdot)] \} I_{[t_{k+1}, T]}(\cdot). \end{aligned} \right.$$

We call $(X_k(\cdot) \oplus X^\Pi(\cdot), u_k(\cdot) \oplus u^\Pi(\cdot))$ a *local variation* of $(X^\Pi(\cdot), u^\Pi(\cdot))$ on $[t_k, t_{k+1}]$. Suppose the following *local optimality condition* holds:

$$J_k^\Pi(X_k^\Pi(\cdot), u_k^\Pi(\cdot)) \leq J_k^\Pi(X_k(\cdot) \oplus X^\Pi(\cdot), u_k(\cdot) \oplus u^\Pi(\cdot)), \quad \forall u_k(\cdot) \in \mathcal{U}[t_k, t_{k+1}].$$

Then we call $(\Theta^\Pi(\cdot), \widehat{\Theta}^\Pi(\cdot))$ a *closed-loop Π -equilibrium strategy* of Problem (MF-LQ), and call $(X^\Pi(\cdot; x), u^\Pi(\cdot; x))$ a *closed-loop Π -equilibrium pair* of Problem (MF-LQ) for the initial state x .

(iii) If the following holds:

$$(86) \quad \lim_{\|\Pi\| \rightarrow 0} \left[\|\Theta^\Pi(\cdot) - \Theta(\cdot)\| + \|\widehat{\Theta}^\Pi(\cdot) - \widehat{\Theta}(\cdot)\| \right] = 0,$$

for some $\Theta, \widehat{\Theta} \in C([0, T]; \mathbb{R}^{m \times n})$, then $(\Theta(\cdot), \widehat{\Theta}(\cdot))$ is called a *closed-loop equilibrium strategy* of Problem (MF-LQ). For any $(t, x) \in \mathcal{D}$, let $\widehat{X}^*(\cdot) \equiv \widehat{X}^*(\cdot; t, x)$ be the solution to the following system:

$$(87) \quad \begin{cases} d\widehat{X}^*(s) = [A(s) - B(s)\widehat{\Theta}(s)]\widehat{X}^*(s)ds + [C(s) - D(s)\widehat{\Theta}(s)]\widehat{X}^*(s)dW(s), \\ \widehat{X}^*(t) = x, \end{cases} \quad s \in [t, T],$$

and define $\widehat{u}^*(\cdot) \equiv \widehat{u}^*(\cdot; t, x)$ as follows:

$$(88) \quad \widehat{u}^*(s) = -\widehat{\Theta}(s)\widehat{X}^*(s), \quad s \in [t, T].$$

Then $(t, x) \mapsto (\widehat{X}^*(\cdot; t, x), \widehat{u}^*(\cdot; t, x))$ is called a *closed-loop equilibrium pair flow* of Problem (MF-LQ). Further,

$$(89) \quad \widehat{V}(t, x) = \widetilde{J}(t, x; \widehat{X}^*(\cdot; t, x), \widehat{u}^*(\cdot; t, x)), \quad (t, x) \in \mathcal{D}$$

is called a *closed-loop equilibrium value function* of Problem (MF-LQ).

We point out that $(\Theta^\Pi(\cdot), \widehat{\Theta}^\Pi(\cdot))$ and $(\Theta(\cdot), \widehat{\Theta}(\cdot))$ are independent of the initial state $x \in \mathbb{R}^n$. To state the main result of this paper, we need one more assumption.

(H7) There exists a $\widetilde{C}(\cdot) \in C([0, T]; \mathbb{R}^{n \times n})$ such that

$$(90) \quad C(s) = D(s)\widetilde{C}(s), \quad s \in [0, T].$$

Theorem 2. *Let (H4)–(H7) hold. Then there exists a unique pair $(\Gamma(\cdot, \cdot), \widehat{\Gamma}(\cdot, \cdot))$ of \mathbb{S}^n -valued functions solving the following system of equations:*

$$(91) \quad \left\{ \begin{array}{l} \Gamma_s(s,t) + \Gamma(s,t)[A(s) - B(s)\widehat{\Theta}(s)] + [A(s) - B(s)\widehat{\Theta}(s)]^\top \Gamma(s,t) + Q(s,t) \\ \quad + [C(s) - D(s)\widehat{\Theta}(s)]^\top \Gamma(s,t)[C(s) - D(s)\widehat{\Theta}(s)] + \widehat{\Theta}(s)^\top R(s,t)\widehat{\Theta}(s) = 0, \\ \widehat{\Gamma}_s(s,t) + \widehat{\Gamma}(s,t)[A(s) - B(s)\widehat{\Theta}(s)] + [A(s) - B(s)\widehat{\Theta}(s)]^\top \widehat{\Gamma}(s,t) + \widehat{Q}(s,t) \\ \quad + [C(s) - D(s)\widehat{\Theta}(s)]^\top \Gamma(s,t)[C(s) - D(s)\widehat{\Theta}(s)] + \widehat{\Theta}(s)^\top \widehat{R}(s,t)\widehat{\Theta}(s) = 0, \\ \hspace{15em} 0 \leq t \leq s \leq T, \\ \Gamma(T,t) = G(t), \quad \widehat{\Gamma}(T,t) = \widehat{G}(t), \quad 0 \leq t \leq T, \end{array} \right.$$

where $\widehat{\Theta}(\cdot)$ is given by the following:

$$(92) \quad \widehat{\Theta}(s) = [\widehat{R}(s,s) + D(s)^\top \Gamma(s,s)D(s)]^{-1} [B(s)^\top \widehat{\Gamma}(s,s) + D(s)^\top \Gamma(s,s)C(s)], \\ s \in [0, T].$$

The closed-loop equilibrium state process $X^*(\cdot)$ is the solution to the following system:

$$(93) \quad \left\{ \begin{array}{l} dX^*(s) = [A(s) - B(s)\widehat{\Theta}(s)]X^*(s)ds + [C(s) - D(s)\widehat{\Theta}(s)]X^*(s)dW(s), \\ \hspace{15em} s \in [0, T], \\ X^*(0) = x, \end{array} \right.$$

the closed-loop equilibrium control admits the following representation:

$$(94) \quad u^*(s) = -\widehat{\Theta}(s)X^*(s), \quad s \in [0, T],$$

and the closed-loop equilibrium value function is given by the following:

$$(95) \quad \widehat{V}(t,x) = \langle \widehat{\Gamma}(t,t)x, x \rangle, \quad \forall (t,x) \in \mathcal{D}.$$

7 Concluding Remarks

Time-inconsistent problems appear frequently in the real world. Two main reasons: time-preferences and risk-preferences. We have formulated various types of problems, introduced the notions of open-loop and closed-loop equilibrium control/strategy. For open-loop equilibrium control of general discounting problem, we presented a Pontryagin's type maximum principle by standard variational method modified from those found in [80, 116]. For closed-loop equilibrium strategy of general discounting problem, we follow the idea of [112] (which is an adaptation of that in [86]) by introducing a family of multi-person differential games, derived an equilibrium HJB equation. The procedure has been significantly simplified and the idea has been much clearly illustrated. Via the equilibrium HJB equation, an closed-loop equilibrium strategy has been constructed. Further, we considered the mixed

case of general discounting and nonlinear appearance of conditional expectations. This can be regarded as the situation that both time-preferences and risk-preferences appear. For linear quadratic problem, such a situation was studied in [113]. For general nonlinear situation, some relevant results can be found in [8, 10, 9]. Our result presented has quite a different looking, which seems to be more natural, from our point of view.

We admit that there are many problems left open. Here is a partial list just for the well-posedness of the equilibrium HJB equation:

(i) When the diffusion coefficient $\sigma(\cdot)$ depends on the control;

(ii) The case that the map $\psi(\cdot)$ appearing in (57) is not regular;

(iii) The case that $\sigma(\cdot)$ is possibly degenerate (recall that it was assumed to be non-degenerate). Is it possible to introduce a proper notion relevant to viscosity solutions?

It is not hard to list more. We will report our further results in the near future.

Although it is not possible to include all relevant references, we have tried our best to collect all found ones below.

References

1. A. Adams, L. Cherchye, B. De Rock, and E. Verriest, *Consume now or later? Time-inconsistency, collective choice and revealed preference*, *Amer. Econ. Review*, 104 (2012), 4147–4183.
2. J. Ai, L. Zhao, and W. Zhu, *Equilibrium and welfare in insurance markets with time-inconsistent consumers*, preprint.
3. Z. Akin, *The role of time-inconsistent preferences in intertemporal investment decisions and bargaining*, preprint.
4. Z. Akin, *Time inconsistency and learning in bargaining game*, *Int. J. Game Theory*, 36 (2007), 275–299.
5. M. Allais, *Le comportement de l'homme rationnel devant de risque, critique des postulats et axiomes de l'école Américaine*, *Econometrica*, 21 (1953), 503–546.
6. J. Andreoni and C. Sprenger, *Risk preferences are not time preferences*, *Amer. Econ. Review*, 102 (2012), 3357–3376.
7. E. Bayraktar, A. Cosso and H. Pham, *Randomized dynamic programming principle and Feynman-Kac representation for optimal control of McKean-Vlasov dynamics*, *Trans AMS*, 370 (2017), 2115–2160.
8. T. Björk and A. Murgoci, *A theory of Markovian time-inconsistent stochastic control in discrete time*, *Finance Stoch.*, 18 (2014), 545–592.
9. T. Björk, M. Khapko, and A. Murgoci, *On time-inconsistent stochastic control in continuous time*, *Finance Stoch.*, 21 (2017), 331–360.
10. T. Björk, A. Murgoci, and X. Y. Zhou, *Mean-variance portfolio optimization with state dependent risk aversion*, *Math. Finance*, 24 (2014), 1–24.
11. M. K. Brunnermeier, F. Papakonstantinou, and J. A. Parker, *Optimal time-inconsistent beliefs: misplanning, procrastination, and commitment*, *Management Sci.*, 63 (2017), 1318–1340.
12. F. N. Caliendo, *Time-inconsistent preferences and social security: Revisited in continuous time*, *J. Econ. Dynamic & Control*, 35 (2011), 668–675.

13. R. Carmona and F. Delarue, *Proabilistic Theory of Mean Field Games with Applications I, Mean Field FBSDEs, Control, and Games*, Springer, 2018.
14. A. Caplin and J. Leahy, *The recursive approach to time inconsistency*, *J. Economic Theory*, 131 (2006), 134–156.
15. F. Cherbonnier, *Optimal insurance for time-inconsistent agents*, preprint.
16. G. Choquet, *Theory of capacities*, *Ann. Inst. Fourier*, 5 (1954), 131–296.
17. B. de Finetti, *Theory of Probability: A Critical Introductory Treatment*, Vol. 1, John Wiley & Sons, New York.
18. M. Dodd, *Obesity and time-inconsistent preferences*, *Obesity Research & Clinical Practice*, 2 (2008), 83–89.
19. A. Dominiak and J.-P. Lefort, *Unambiguous events and dynamic Choquet preferences*, *Econ. Theory*, 46 (2011), 401–425.
20. Y. Dong and R. Sircar, *Time-inconsistent portfolio investment problems*, *Springer Proc. Math. Stat.*, 100, *Stoch. Anal. Appl.*, 239–281.
21. D. Duffie, L. Epstein, *Stochastic differential utility*, *Econometrica*, 60 (1992), 353–394.
22. I. Ekeland and A. Lazrak, *The golden rule when preferences are time-inconsistent*, *Math. Finance Econ.*, 4 (2010), 29–55.
23. I. Ekeland, Y. Long, and Q. Zhou, *A new class of problems in the calculus of variations, Regular and Chaotic Dynamics*, 18 (2013), 553–584.
24. I. Ekeland, O. Mbodji, and T. A. Pirvu, *Time-consistent portfolio management*, *SIAM J. Fin. Math.*, 3 (2012), 1–32.
25. I. Ekeland and T. A. Pirvu, *Investment and consumption without commitment*, *Math. Finance Econ.*, 2 (2007), 57–68.
26. J. D. El Baghdady, *Equilibrium strategies for time-inconsistent stochastic optimal control of asset allocation*, PhD Dissertation, KTH Royal Institute of Technology, Stockholm, Sweden, 2017.
27. N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, *Math. Finance*, 7 (1997), 1–71.
28. N. El Karoui, S. Peng, and M. C. Quenez, *A dynamic maximum principle for the optimization of recursive utilities under constraints*, *Ann. Appl. Probab.*, 11 (2001), 664–693.
29. D. Ellsberg, *Risk, ambiguity and the Savage axioms*, *Quarterly Journal of Economics*, 75 (1961), 643–649.
30. T. S. Findley and F. N. Caliendo, *Short horizons, time inconsistency and optimal social security*, *Int. Tax & Public Finance*, 16 (2009), 487–513.
31. T. S. Findley and J. A. Feigenbaum, J. A. *Quasi-hyperbolic discounting and the existence of time-inconsistent retirement*, *Theoretical Econ. Lett.*, 3 (2013), 119–123.
32. S. Frederick, G. Loewenstein, and T. O'Donoghue, *Time discounting and time preference: A critical review* *J. Economic Literature*, 40 (2002), 351–401.
33. S. M. Goldman, *Consistent plans*, *Review of Economic Studies*, 47 (1980), 533–537.
34. S. L. Green, *Time inconsistency, self-control, and remembrance*, *Faith & Economics*, 42 (2003), 51–60.
35. S. R. Grenadier and N. Wang, *Investment under uncertainty and time-inconsistent preferences*, *J. Financial Economics*, 84 (2007), 2–39.
36. L. Guo and F. N. Caliendo, *Time-inconsistent preferences and time-inconsistent policies*, *J. Math. Econ.*, 51 (2011), 102–108.
37. Y. Halevy, *Time consistency: stationarity and time invariance*, *Econometrica*, 83 (2015), 335–352.
38. X. He and X. Y. Zhou, *Portfolio choice via quantiles*, *Mathematical Finance*, 21 (2011), 203–231.
39. P. J. Herings and K. I. M. Rohde, *Time-inconsistent preferences in a general equilibrium model*, *Economic Theory*, 29 (2006), 591–619.
40. G. Heutel, *Optimal policy instruments for externality-producing durable goods under time inconsistency*, preprint.
41. M. Hinnosaar, *Time inconsistency and alcohol sales restrictions*, *European Economic Review*, 87 (2012), 108–131.

42. S. J. Hoch and G. F. Loewenstein, *Time-inconsistent preferences and consumer self-control*, *J. Consumer Research*, 17 (1991), 492–507.
43. Y. Hu, H. Jin, and X. Y. Zhou, *Time-inconsistent stochastic linear-quadratic control*, *SIAM J. Control Optim.*, 50 (2012), 1548–1572.
44. Y. Hu, H. Jin, and X. Y. Zhou, *Time-inconsistent stochastic linear-quadratic control: characterization and uniqueness of equilibrium*, *SIAM J. Control Optim.*, 55 (2017), 1261–1279.
45. D. Hume, *A Treatise of Human Nature*, First Edition, New York, Oxford Univ. Press 1978.
46. P. Ireland, *Does the time-inconsistency problem explain the behavior of inflation in the United States?* *J. Monetary Economics*, 44 (1999), 279–291.
47. M. O. Jackson and L. Yariv, *Collective dynamic choice: The necessity of time inconsistency*, *Amer. Econ. J. Microeconomics*, 7 (2015), 159–178.
48. H. Jin and X. Y. Zhou, *Behavioral portfolio selection in continuous time*, *Mathematical Finance*, (2008), 385–426.
49. D. Jones and A. Mahajan, *Time-inconsistency and savings: Experimental evidence from low-income tax filers*, Center for Financial Security, University of Wisconsin-Madison, 2011.
50. D. Kahneman and A. Tversky, *Prospective theory: an analysis of decision under risk*, *Econometrica*, 47 (1979), 263–291.
51. C. Karnam, J. Ma, and J. Zhang, *Dynamic approaches for some time-inconsistent optimization problems*, *Ann. Appl. Probab.*, 27 (2017), 3435–3477.
52. J. Klenberg and S. Oren, *Time-inconsistent planning: a computational problem in behavioral economics*, In *Proc. 15th ACM Conference on Economics and Computation*. ACM, 547–564.
53. L. Karp, *Global warming and hyperbolic discounting*, *J. Public Econ.*, 8 (2005), 261–282.
54. L. Karp, *Non-constant discounting in continuous time*, *J. Economic Theory*, 132 (2007), 557–568.
55. L. Karp and I. H. Lee, *Time-consistent policies*, *J. Economic Theory*, 112 (2003), 353–364.
56. T. Komatsubara, *Illiquid securities and time-inconsistent preferences*, preprint.
57. F. Kydland and E. Prescott, *Rules rather than discretion: the inconsistency of optimal plans*, *J. Political Economy*, 85 (1977), 473–491.
58. D. Laibson, *Golden eggs and hyperbolic discounting*, *Quartely J. Economics*, 112 (1987), 443–477.
59. K. Lindensjö, *Time-inconsistent stochastic control: solving the extended HJB system is a necessary condition for regular equilibria*, Research Report, No.20, Math. Stat., Stockholm Univ.
60. L. L. Lopes, *Between hope and fear: The psychology of risk*, *Advances in Experimental Social Psychology*, 20 (1987), 286–313.
61. H. Mei and J. Yong, *Equilibrium strategies for time-inconsistent stochastic switching systems*, *COCV*, to appear.
62. E. G. J. Luttmer and T. Mariotti, *Efficiency and equilibrium when preferences are time-inconsistent*, *J. Economic Theory*, 132 (2007), 493–506.
63. J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Math., Vol.1702, Springer-Verlag.
64. A. Mahajan and A. Tarozzi, *Time inconsistency, expectations and technology adoption: The case of insecticide treated nets*, preprint.
65. J. Marin-Solano, *Time-consistent equilibria in a differential game model with time inconsistent preferences and partial cooperation*, *Dynamic games in economics, Dynamic Modeling and Econometrics in Economics and Finance*, 16 (2014), 219–238.
66. J. Marin-Solano and J. Navas, *Non-constant discounting in finite horizon: the free terminal time case*, *J. Economic Dynamics and Control*, 33 (2009), 666–675.
67. J. Marin-Solano and J. Navas, *Consumption and portfolio rules for time-inconsistent investors*, *European J. Operational Research*, 201 (2010), 860–872.
68. J. Marin-Solano and E. V. Shevkoplyas, *Non-constant discounting and differential games with random time horizon*, *Automatica J. IFAC*, 47 (2011), 2626–2638.
69. R. Mehra and E. C. Prescott, *The equality premium: A puzzle*, *J. Monetary Econ.*, 15 (1985), 145–161.

70. T. Michielson, *Environmental catastrophes under time-inconsistent preferences*, preprint.
71. C. W. Miller, *Methods for optimal stochastic control and optimal stopping problems featuring time-inconsistency*, PhD dissertation, UC Berkeley.
72. Y. Narukawa and T. Murofushi, *Decision modeling using the Choquet integral*, *Lecture Notes in Computer Science*, 3131 (2004), 183–193.
73. N. Netzer, *Evolution of time preferences and attitudes towards risk*, *Amer. Econ. Review*, 99 (2009), 937–955.
74. L. Mou and J. Yong, *A variational formula for stochastic controls and some applications*, *Pure & Appl. Math. Quarterly*, 3 (2007), 539–567.
75. T. O'Donoghue and M. Rabin, *Doing it now or later*, *Amer. Economic Review*, 103–124.
76. T. O'Donoghue and M. Rabin, *Procrastination on long-term projects*, *J. Economic Behavior & Organization*, 66 (2008), 161–175.
77. I. Palacios-Huerta, *Time-inconsistent preferences in Adam Smith and Davis Hume*, *History of Political Economy*, 35 (2003), 241–268.
78. E. Pardoux and S. Peng, *Adapted solution of a backward stochastic differential equation*, *Syst. Control Lett.*, 14 (1990), 55–61.
79. B. Peleg and M. E. Yaari, *On the existence of a consistent course of action when tastes are changing*, *Review of Economic Studies* 40 (1973), 391–401.
80. S. Peng, *A general stochastic maximum principle for optimal control problems*, *SIAM J. Control Optim.*, 28 (1990), 966–979.
81. S. Peng, *Backward SDE and related g-expectation*. *Backward stochastic differential equations (Paris, 1995–1996)*, 141–159, *Pitman Res. Notes Math. Ser.*, 364, Longman, Harlow, 1997.
82. S. Peng, *Nonlinear expectations, nonlinear evaluations and risk measures*, *Stochastic methods in finance*, 165–253, *Lecture Notes in Math.*, 1856, Springer, Berlin.
83. S. Peng, *Backward stochastic differential equations, nonlinear expectation and their applications*, *Proc. ICM*, Hyderabad, India.
84. H. Pham, *Linear quadratic optimal control of conditional McKean-Vlasov equation with random coefficients and applications*, *Prob. Uncertainty & Quantitative Risk*, 1 (2016): 7
85. H. Pham and X. Wei, *Dynamic programming for optimal control of stochastic McKean-Vlasov dynamics*, *SIAM J. Control Optim.*, 55 (2017), 1069–1101.
86. R. A. Pollak, *Consistent planning*, *Review of Economic Studies*, 35 (1968), 201–208.
87. D. Prelec and G. Loewenstein, *Beyond time discounting*, *Marketing letters*, 8 (1997), 97–108.
88. F. P. Ramsey, *The foundation of probability*, *Frank Plumpton Ramsey Papers, 1920–1930*, ASP, 1983.01, *Archives of Scientific Philosophy*, Special Collections Department, Univ. of Pittsburgh.
89. K. I. Rohde, *The hyperbolic factor: A measure of time inconsistency*, *J. Risk Uncertain*, 41 (2010), 125–140.
90. L. J. Savage, *The Foundations of Statistics*, Wiley, New York, 1954.
91. P. Schreiber and M. Weber, *Time inconsistent preferences and the annuitization decision*, *J. Econ. Behavior & Organization*, 129 (2016), 37–55.
92. D. Schmeidler, *Subjective probability and expected utility without additivity*, *Econometrica*, 60 (1989), 1255–1272.
93. H. Shefrin and M. Statman, *Behavioral portfolio theory*, *J. Financial & Quantitative Anal.*, 35 (2000), 127–151.
94. Y. Shi, T. Wang, and J. Yong, *Mean-field backward stochastic Volterra integral equations*, *Discrete and Continuous Dynamic Systems*, 18 (2013), 1929–1967.
95. H. Shui, *Time inconsistency in the credit card market*, preprint.
96. A. Smith, *The Theory of Moral Sentiments*, First Edition, Oxford Univ. Press, 1976.
97. R. H. Strotz, *Myopia and inconsistency in dynamic utility maximization*, *Review of Economic Studies*, 23 (1955–1956), 165–180.
98. A. Tversky and D. Kahneman, *The framing of decisions and the psychology of choice*, *Science, (New Series)*, 211 (1981), No. 4481, 453–458.
99. A. Tversky and D. Kahneman, *Rational choice and the framing of decisions*, *J. Business*, 59 (1986), S251–S278.

100. A. Tversky and D. Kahneman, *Advances in prospect theory: cumulative representation of uncertainty*, *J. Risk Uncertainty*, 5 (1992), 297–323.
101. H. Wang and Z. Wu, *Time-inconsistent optimal control problems with random coefficients and stochastic equilibrium HJB equation*, *Math. Control Rel. Fields*, 5 (2015), 651–678.
102. S. Wang, V. R. Young, and H. H. Panjer, *Axiomatic characterization of insurance prices*, *Insurance: Mathematics and Economics*, 21 (1997), 173–183.
103. J. Wei, *Time-inconsistent optimal control problems with regime switching*, *Math. Control Rel. Fields*, 7 (2017), 585–622.
104. Q. Wei, J. Yong, and Z. Yu, *Time-inconsistent recursive stochastic optimal control problems*, *SIAM J. Control Optim.*, 55 (2017), 4156–4201.
105. M. E. Yaari, *The dual theory of choice under risk*, *Econometrica*, 55 (1987), 95–115.
106. M. Yilmaz, *Contracting with a naive time-inconsistent agent: To exploit or not to exploit?* *Math. Soc. Sci.*, 77 (2013), 46–71.
107. J. Yong, *Backward stochastic Volterra integral equations and some related problems*, *Stoch. Proc. Appl.*, 116 (2006), 779–795.
108. J. Yong, *Well-posedness and regularity of backward stochastic Volterra integral equations*, *Prob. Theory Rel. Fields*, 142 (2008), 21–77.
109. J. Yong, *Optimality variational principle for optimal controls of forward-backward stochastic differential equations*, *SIAM J. Control & Optim.*, 48 (2010), 4119–4156.
110. J. Yong, *A deterministic linear quadratic time-inconsistent optimal control problem*, *Math. Control & Related Fields*, 1 (2011), 83–118.
111. J. Yong, *Deterministic time-inconsistent optimal control problems — An essentially cooperative approach*, *Acta Math. Appl. Sinica*, 28 (2012), 1–30.
112. J. Yong, *Time-inconsistent optimal control problems and the equilibrium HJB equation*, *Math. Control & Related Fields*, 2 (2012), 271–329.
113. J. Yong, *Linear-quadratic optimal control problems for mean-field stochastic differential equations — time-consistent solutions*, *Trans. AMS*, 369 (2017), 5467–5523.
114. J. Yong, *A linear-quadratic optimal control problem for mean-field stochastic differential equations*, *SIAM J. Control & Optim.*, 51 (2013), 2809–2838.
115. J. Yong, *Time-inconsistent optimal control problems*, *Proc. ICM 2014, Section 16. Control Theory and Optimization*, 947–969.
116. J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, 1999.
117. P. C. Yu, *Optimal retirement policies with time-inconsistent agents*, preprint.
118. Q. Zhao, *On time-inconsistent investment and dividend problems*, PhD Dissertation, East China Normal University, 2015.
119. X. Y. Zhou, *Mathematicalising behavioural finance*, *Proc. ICM, Hyderabad, India*, 2010.



Regime-Switching Jump Diffusions with Non-Lipschitz Coefficients and Countably Many Switching States: Existence and Uniqueness, Feller, and Strong Feller Properties^{*}

Fubao Xi, George Yin, and Chao Zhu

Abstract This work focuses on a class of regime-switching jump diffusion processes, which is a two component Markov processes $(X(t), \Lambda(t))$, where $\Lambda(t)$ is a component representing discrete events taking values in a countably infinite set. Considering the corresponding stochastic differential equations, our main focus is on treating those with non-Lipschitz coefficients. We first show that there exists a unique strong solution to the corresponding stochastic differential equation. Then Feller and strong Feller properties are investigated.

Keywords. Regime-switching jump diffusion, non-Lipschitz condition, Feller property, strong Feller property.

Mathematics Subject Classification. 60J27, 60J60, 60J75, 60G51.

F. Xi

School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081, China, e-mail: xifb@bit.edu.cn

G. Yin

Department of Mathematics, Wayne State University, Detroit, MI 48202, e-mail: gyin@math.wayne.edu

C. Zhu

Department of Mathematical Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, e-mail: zhu@uwm.edu

^{*} The study was initiated during a workshop held at IMA, University of Minnesota. The support of IMA with funding provided by the National Science Foundation is acknowledged. The research was also supported in part by the National Natural Science Foundation of China under Grant No. 11671034, the US Army Research Office, and the Simons Foundation Collaboration Grant (No. 523736).

1 Introduction

In the past decade, much attention has been devoted to a class of hybrid systems, namely, regime-switching diffusions. Roughly, such processes can be considered as a two component process $(X(t), \Lambda(t))$, an analog (or continuous state) component $X(t)$ and a switching (or discrete event) process $\Lambda(t)$. Some of the representative works can be found in [14] and [27]. The former dealt with regime-switching diffusions in which the switching process is a continuous-time Markov chain independent of the Brownian motion, whereas the latter treated processes in which the switching component depends on the continuous-state component. It has been found that the discrete event process, taking values in a finite or countable set, can be used to delineate, for example, random environment or other random factors that are not represented in the usual diffusion formulation. Seemingly similar to the diffusion processes, in fact, regime-switching diffusions have very different behavior compared to the usual diffusion processes. For example, it has been demonstrated in [11, 26] that two stable (resp., unstable) ordinary differential equations can be coupled to produce an unstable (resp., stable) regime-switching system. The consideration of regime-switching diffusions has substantially enlarged the applicability of stochastic processes for a wide variety of problems ranging from network systems, multi-agent systems, ecological and biological applications, financial engineering, risk management, etc.

Continuing on the effort of studying regime-switching diffusions, [3] obtained maximum principle and Harnack inequalities for switching jump diffusions using mainly probabilistic arguments, and [4] proceeded further to obtain recurrence and ergodicity of switching jump diffusions. In another direction, [22] dealt with regime-switching jump diffusions with countable number of switching values. [17] considered switching diffusions in which the switching process depends on the past information of the continuous state and takes values in a countable state space; the corresponding recurrence and ergodicity was considered in [16].

A standing assumption in the aforementioned references is that the coefficients of the associated stochastic differential equations are (locally) Lipschitz. While it is a convenient assumption, it is rather restrictive in many applications. For example, the diffusion coefficients in the Feller branching diffusion and the Cox-Ingersoll-Ross model are only Hölder continuous. We refer to Chapters 12 and 13 of [8] for an introduction to these models. Motivated by these considerations, there has been much efforts devoted to the study of stochastic differential equations with non-Lipschitz coefficients. An incomplete list includes [1, 6, 12, 13, 25], among many others.

While there are many works on diffusions and jump diffusions with non-Lipschitz coefficients, the related research on regime-switching jump diffusions is relatively scarce. This work aims to investigate regime-switching jump diffusion processes with non-Lipschitz coefficients. More precisely, the purpose of this paper is two-fold: (i) to establish the strong existence and uniqueness result for stochastic differential equations associated with regime-switching jump diffusions, in which the coefficients are non-Lipschitz and the switching component has countably many

states; and (ii) to derive sufficient conditions for Feller and strong Feller properties. Our focus is devoted to establishing non-Lipschitz sufficient conditions for the aforementioned properties.

The rest of the paper is arranged as follows. Examining the associated stochastic differential equations, we begin to obtain the existence and uniqueness of the solution of the stochastic differential equations in Section 2. Then Section 3 proceeds with the study of Feller properties. Section 4 further extends the study to treat strong Feller properties.

2 Strong Solution: Existence and Uniqueness

We work with (U, \mathcal{U}) a measurable space, ν a σ -finite measure on U , and $\mathbb{S} := \{1, 2, \dots\}$. Assume that $d \geq 1$ is a positive integer, $b : \mathbb{R}^d \times \mathbb{S} \mapsto \mathbb{R}^d$, $\sigma : \mathbb{R}^d \times \mathbb{S} \mapsto \mathbb{R}^{d \times d}$, and $c : \mathbb{R}^d \times \mathbb{S} \times U \mapsto \mathbb{R}^d$ be Borel measurable functions. Let (X, Λ) be a right continuous, strong Markov process with left-hand limits on $\mathbb{R}^d \times \mathbb{S}$. The first component X satisfies the following stochastic differential-integral equation

$$dX(t) = b(X(t), \Lambda(t))dt + \sigma(X(t), \Lambda(t))dW(t) + \int_U c(X(t-), \Lambda(t-), u)\tilde{N}(dt, du), \tag{1}$$

where W is a standard d -dimensional Brownian motion, N is a Poisson random measure on $[0, \infty) \times U$ with intensity $dt \nu(du)$, and \tilde{N} is the associated compensated Poisson random measure. The second component Λ is a continuous-time random process taking values in the countably infinite set \mathbb{S} such that

$$\mathbb{P}\{\Lambda(t + \Delta) = l | \Lambda(t) = k, X(t) = x\} = \begin{cases} q_{kl}(x)\Delta + o(\Delta), & \text{if } k \neq l, \\ 1 + q_{kk}(x)\Delta + o(\Delta), & \text{if } k = l, \end{cases} \tag{2}$$

uniformly in \mathbb{R}^d , provided $\Delta \downarrow 0$.

To proceed, we construct a family of disjoint intervals $\{\Delta_{ij}(x) : i, j \in \mathbb{S}\}$ on the positive half real line as follows

$$\begin{aligned} \Delta_{12}(x) &= [0, q_{12}(x)), \\ \Delta_{13}(x) &= [q_{12}(x), q_{12}(x) + q_{13}(x)), \\ &\vdots \\ \Delta_{21}(x) &= [q_1(x), q_1(x) + q_{21}(x)), \\ \Delta_{23}(x) &= [q_1(x) + q_{21}(x), q_1(x) + q_{21}(x) + q_{23}(x)), \\ &\vdots \\ \Delta_{31}(x) &= [q_1(x) + q_2(x), q_1(x) + q_2(x) + q_{31}(x)), \\ &\vdots \end{aligned}$$

where for convenience, we set $\Delta_{ij}(x) = \emptyset$ if $q_{ij}(x) = 0, i \neq j$. Note that for each $x \in \mathbb{R}^d, \{\Delta_{ij}(x) : i, j \in \mathbb{S}\}$ are disjoint intervals, and the length of the interval $\Delta_{ij}(x)$ is equal to $q_{ij}(x)$. We then define a function $h: \mathbb{R}^d \times \mathbb{S} \times \mathbb{R}_+ \rightarrow \mathbb{R}$ by

$$h(x, k, r) = \sum_{l \in \mathbb{S}} (l - k) \mathbf{1}_{\Delta_{kl}(x)}(r). \tag{3}$$

That is, for each $x \in \mathbb{R}^d$ and $k \in \mathbb{S}$, we set $h(x, k, r) = l - k$ if $r \in \Delta_{kl}(x)$ for some $l \neq k$; otherwise $h(x, k, r) = 0$. Consequently, we can describe the evolution of Λ using the following stochastic differential equation

$$\Lambda(t) = \Lambda(0) + \int_0^t \int_{\mathbb{R}_+} h(X(s-), \Lambda(s-), r) N_1(ds, dr), \tag{4}$$

where N_1 is a Poisson random measure on $[0, \infty) \times [0, \infty)$ with characteristic measure $\mathbf{m}(dz)$, the Lebesgue measure.

For convenience in the subsequent discussion, let us give the infinitesimal generator \mathcal{A} of the regime-switching jump diffusion (X, Λ)

$$\mathcal{A}f(x, k) := \mathcal{L}_k f(x, k) + Q(x)f(x, k), \tag{5}$$

with $a(x, k) := \sigma \sigma^T(x, k)$ and

$$\mathcal{L}_k f(x, k) := \frac{1}{2} \text{tr}(a(x, k) \nabla^2 f(x, k)) + \langle b(x, k), \nabla f(x, k) \rangle \tag{6}$$

$$+ \int_U (f(x + c(x, k, u), k) - f(x, k) - \langle \nabla f(x, k), c(x, k, u) \rangle) \nu(du),$$

$$Q(x)f(x, k) := \sum_{j \in \mathbb{S}} q_{kj}(x) [f(x, j) - f(x, k)] \tag{7}$$

$$= \int_{[0, \infty)} [f(x, k + h(x, k, z)) - f(x, k)] \mathbf{m}(dz).$$

Define a metric $\lambda(\cdot, \cdot)$ on $\mathbb{R}^d \times \mathbb{S}$ as $\lambda((x, m), (y, n)) = |x - y| + d(m, n)$, where $d(m, n) = \mathbf{1}_{\{m \neq n\}}$ is the discrete metric on \mathbb{S} . Let $\mathfrak{B}(\mathbb{R}^d \times \mathbb{S})$ be the Borel σ -algebra on $\mathbb{R}^d \times \mathbb{S}$. Then $(\mathbb{R}^d \times \mathbb{S}, \lambda(\cdot, \cdot), \mathfrak{B}(\mathbb{R}^d \times \mathbb{S}))$ is a locally compact and separable metric space. For the existence and uniqueness of the strong Markov process (X, Λ) satisfying system (1) and (4), we make the following assumptions.

Assumption 2.1 There exists a nondecreasing function $\zeta : [0, \infty) \mapsto [1, \infty)$ that is continuously differentiable and that satisfies

$$\int_0^\infty \frac{dr}{r\zeta(r) + 1} = \infty, \tag{8}$$

such that for all $x \in \mathbb{R}^d$ and $k \in \mathbb{S}$,

$$2\langle x, b(x, k) \rangle + |\sigma(x, k)|^2 + \int_U |c(x, k, u)|^2 \nu(du) \leq H[|x|^2 \zeta(|x|^2) + 1], \tag{9}$$

$$q_k(x) := -q_{kk}(x) = \sum_{l \in \mathbb{S} \setminus \{k\}} q_{kl}(x) \leq Hk, \tag{10}$$

$$\sum_{l \in \mathbb{S} \setminus \{k\}} (f(l) - f(k))q_{kl}(x) \leq H(1 + \Phi(x) + f(k)), \tag{11}$$

where H is a positive constant,

$$\Phi(x) := \exp \left\{ \int_0^{|x|^2} \frac{dr}{r\zeta(r) + 1} \right\}, \quad x \in \mathbb{R}^d, \tag{12}$$

and the function $f : \mathbb{S} \mapsto \mathbb{R}_+$ is nondecreasing satisfying $f(m) \rightarrow \infty$ as $m \rightarrow \infty$. In addition, assume there exists some $\delta \in (0, 1]$ such that

$$\sum_{l \in \mathbb{S} \setminus \{k\}} |q_{kl}(x) - q_{kl}(y)| \leq H|x - y|^\delta \tag{13}$$

for all $k \in \mathbb{S}$ and $x, y \in \mathbb{R}^d$.

Assumption 2.2 Assume the following conditions hold.

- If $d = 1$, then there exist a positive number δ_0 and a nondecreasing and concave function $\rho : [0, \infty) \mapsto [0, \infty)$ satisfying

$$\int_{0^+} \frac{dr}{\rho(r)} = \infty \tag{14}$$

such that for all $k \in \mathbb{S}$, $R > 0$, and $x, z \in \mathbb{R}$ with $|x| \vee |z| \leq R$ and $|x - z| \leq \delta_0$,

$$\text{sgn}(x - z)(b(x, k) - b(z, k)) \leq \kappa_R \rho(|x - z|), \tag{15}$$

$$|\sigma(x, k) - \sigma(z, k)|^2 + \int_U |c(x, k, u) - c(z, k, u)|^2 \nu(du) \leq \kappa_R |x - z|, \tag{16}$$

where κ_R is a positive constant and $\text{sgn}(a) = 1$ if $a > 0$ and -1 if $a \leq 0$. In addition, for each $k \in \mathbb{S}$, the function c satisfies that

$$\text{the function } x \mapsto x + c(x, k, u) \text{ is nondecreasing for all } u \in U; \tag{17}$$

or, there exists some $\beta > 0$ such that

$$|x - z + \theta(c(x, k, u) - c(z, k, u))| \geq \beta|x - z|, \quad \forall (x, z, u, \theta) \in \mathbb{R} \times \mathbb{R} \times U \times [0, 1]. \tag{18}$$

- If $d \geq 2$, then there exist a positive number δ_0 , and a nondecreasing and concave function $\rho : [0, \infty) \mapsto [0, \infty)$ satisfying

$$0 < \rho(r) \leq (1 + r)^2 \rho(r/(1 + r)) \text{ for all } r > 0, \text{ and } \int_{0^+} \frac{dr}{\rho(r)} = \infty \tag{19}$$

such that for all $k \in \mathbb{S}$, $R > 0$, and $x, z \in \mathbb{R}^d$ with $|x| \vee |z| \leq R$ and $|x - z| \leq \delta_0$,

$$2\langle x - z, b(x, k) - b(z, k) \rangle + |\sigma(x, k) - \sigma(z, k)|^2 + \int_U |c(x, k, u) - c(z, k, u)|^2 \nu(du) \leq \kappa_R \rho(|x - z|^2), \tag{20}$$

where κ_R is a positive constant.

Remark 2.3 We make some comments concerning Assumptions 2.1 and 2.2. Examples of functions satisfying (8) include $\zeta(r) = 1$, $\zeta(r) = \log r$, and $\zeta(r) = \log r \log(\log r)$ for r large. When $\zeta(r) = 1$, (9) reduces to the usual linear growth condition. With other choices of ζ , (8) allows super-linear condition for the coefficients of (1) with respect to the variable x for each $k \in \mathbb{S}$. This is motivated by applications such as Lotka-Volterra models, in which the coefficients have superlinear growth conditions. Conditions (10) and (11) are imposed so that the Λ component will not explode in finite time with probability 1; see the proof of Theorem 2.5 for details.

Examples of functions satisfying (14) or (19) include $\rho(r) = r$ and concave and increasing functions such as $\rho(r) = r \log(1/r)$, $\rho(r) = r \log(\log(1/r))$, and $\rho(r) = r \log(1/r) \log(\log(1/r))$ for $r \in (0, \delta)$ with $\delta > 0$ small enough. When $\rho(r) = r$, Assumption 2.2 is just the usual local Lipschitz condition. With other choices of continuity modularity, Assumption 2.2 allows the drift, diffusion, and jump coefficients of (1) to be non-Lipschitz with respect to the variable x . This, in turn, presents more opportunities for building realistic and flexible mathematical models for a wide range of applications. Indeed, non-Lipschitz coefficients are present in areas such as branching diffusion in biology, the Cox-Ingersoll-Ross model in math finance, etc.

It is also worth pointing out that (15), (16), and (20) of Assumption 2.2 only require the modulus of continuity to hold in a small neighborhood of the diagonal line $x = z$ in $\mathbb{R}^d \otimes \mathbb{R}^d$ with $|x| \vee |z| \leq R$ for each $R > 0$. This is in contrast to those in [13] and adds some subtlety in the proof of pathwise uniqueness for (21).

When $d = 1$, Assumption 2.2 allows the diffusion coefficient $\sigma(\cdot, k)$ to be locally Hölder continuous with exponent $\alpha \in [\frac{1}{2}, 1]$. This is the celebrated result in [25]. Such a result was extended to stochastic differential equations with jumps; see, for example, [7, 12] and [13], among others. In particular, [13] shows that if (17) holds, the function $x \mapsto \int_U c(x, k, u) \nu(du)$ can be locally Hölder continuous with exponent $\alpha \in [\frac{1}{2}, 1]$ as well. The continuity assumption (15) on the drift coefficient $b(\cdot, k)$ is slightly more general than that in [13]. In particular, (15) will be satisfied as long as $b(\cdot, k)$ is decreasing.

Lemma 2.4 *Suppose Assumption 2.2 and (9) hold. Then for each $k \in \mathbb{S}$, the stochastic differential equation*

$$X(t) = x + \int_0^t b(X(s), k) ds + \int_0^t \sigma(X(s), k) dW(s) + \int_0^t \int_U c(X(s-), k, u) \tilde{N}(ds, du) \tag{21}$$

has a unique non-explosive strong solution.

Proof. Condition (9) guarantees that the solution to (21) will not explode in finite time with probability 1; see, for example, Theorem 2.1 in [24]. When $d \geq 2$, the existence and uniqueness of a strong solution to (21) under Assumption 2.2 follows from Theorem 2.8 of [24].

When $d = 1$, we follow the arguments in the proof of Theorem 3.2 of [13] to show that pathwise uniqueness holds for (21). First, let $\{a_n\}$ be a strictly decreasing sequence of real numbers satisfying $a_0 = 1$, $\lim_{n \rightarrow \infty} a_n = 0$, and $\int_{a_n}^{a_{n-1}} \frac{dr}{r} = n$ for each $n \geq 1$. For each $n \geq 1$, let ρ_n be a nonnegative continuous function with support on (a_n, a_{n-1}) so that

$$\int_{a_n}^{a_{n-1}} \rho_n(r) dr = 1 \text{ and } \rho_n(r) \leq 2(kr)^{-1} \text{ for all } r > 0.$$

For $x \in \mathbb{R}$, define

$$\psi_n(x) = \int_0^{|x|} \int_0^y \rho_n(z) dz dy. \tag{22}$$

We can immediately verify that ψ_n is even and twice continuously differentiable, with

$$\psi'_n(r) = \operatorname{sgn}(r) \int_0^{|r|} \rho_n(z) dz = \operatorname{sgn}(r) |\psi'_n(r)|, \tag{23}$$

and

$$|\psi'_n(r)| \leq 1, \quad 0 \leq |r| \psi''_n(r) = |r| \rho_n(|r|) \leq \frac{2}{n}, \quad \text{and} \quad \lim_{n \rightarrow \infty} \psi_n(r) = |r| \tag{24}$$

for $r \in \mathbb{R}$. Furthermore, for each $r > 0$, the sequence $\{\psi_n(r)\}_{n \geq 1}$ is nondecreasing. Note also that for each $n \in \mathbb{N}$, ψ_n , ψ'_n , and ψ''_n all vanish on the interval $(-a_n, a_n)$. Moreover the classical arguments reveal that

$$\begin{aligned} & \frac{1}{2} \psi''_n(x-z) |\sigma(x, k) - \sigma(z, k)|^2 + \int_U [\psi_n(x-z + c(x, k, u)) - \psi_n(x-z) \\ & \quad - \psi'_n(x-z)(c(x, k, u) - c(z, k, u))] \nu(du) \\ & \leq \frac{1}{2} \cdot \frac{2}{n} \kappa_R + \frac{\kappa_R}{n} \left(\frac{1}{\beta} \vee 2 \right) \leq K \frac{\kappa_R}{n}, \end{aligned} \tag{25}$$

for all x, z with $|x| \vee |z| \leq R$ and $0 < |x-z| \leq \delta_0$, where K is a positive constant independent of R and n . On the other hand, for any $x, z \in \mathbb{R}$ with $|x| \vee |z| \leq R$ and $|x-z| \leq \delta_0$, it follows from (15) and (23) that

$$\begin{aligned} \psi'_n(x-z)(b(x, k) - b(z, k)) &= \operatorname{sgn}(x-z) |\psi'_n(x-z)| (b(x, k) - b(z, k)) \\ &\leq \kappa_R \rho(|x-z|). \end{aligned} \tag{26}$$

Let \tilde{X} and X be two solutions to (21). Denote $\Delta_t := \tilde{X}(t) - X(t)$ for $t \geq 0$. Assume $|\Delta_0| = |\tilde{x} - x| < \delta_0$ and define

$$S_{\delta_0} := \inf\{t \geq 0 : |\Delta_t| \geq \delta_0\} = \inf\{t \geq 0 : |\tilde{X}(t) - X(t)| \geq \delta_0\}.$$

For $R > 0$, let $\tau_R := \inf\{t \geq 0 : |\tilde{X}(t)| \vee |X(t)| > R\}$. Then $\tau_R \rightarrow \infty$ a.s. as $R \rightarrow \infty$. Moreover, by Itô’s formula, we have

$$\begin{aligned} \mathbb{E}[\psi_n(\Delta_{t \wedge S_{\delta_0} \wedge \tau_R})] &= \psi_n(|\Delta_0|) + \mathbb{E} \left[\int_0^{t \wedge \tau_R \wedge S_{\delta_0}} \left\{ \psi'_n(\Delta_s) [b(\tilde{X}(s), k) - b(X(s), k)] \right. \right. \\ &\quad + \frac{1}{2} \psi''_n(\Delta_s) [\sigma(\tilde{X}(s), k) - \sigma(X(s), k)]^2 \\ &\quad + \int_U [\psi_n(\Delta_s + c(\tilde{X}(s), k, u) - c(X(s), k, u)) - \psi_n(\Delta_s) \\ &\quad \left. \left. - \psi'_n(\Delta_s)(c(\tilde{X}(s), k, u) - c(X(s), k, u))] \nu(du) \right\} ds \right]. \end{aligned}$$

Furthermore, using (25) and (26), we obtain

$$\begin{aligned} \mathbb{E}[\psi_n(\Delta_{t \wedge S_{\delta_0} \wedge \tau_R})] &\leq \psi_n(|\Delta_0|) + \mathbb{E} \left[\int_0^{t \wedge \tau_R \wedge S_{\delta_0}} \left(\kappa_R \rho(|\Delta_s|) + K \frac{\kappa_R}{n} \right) ds \right] \\ &\leq \psi_n(|\Delta_0|) + K \frac{\kappa_R}{n} t + \int_0^t \kappa_R \rho(\mathbb{E}[|\Delta_{s \wedge \tau_R \wedge S_{\delta_0}}|]) ds, \end{aligned}$$

where the second inequality follows from the concavity of ρ and Jensen’s inequality. Upon passing to the limit as $n \rightarrow \infty$, we obtain from the third equation in (24) and the monotone convergence theorem that

$$\mathbb{E}[|\Delta_{t \wedge S_{\delta_0} \wedge \tau_R}|] \leq |\Delta_0| + \kappa_R \int_0^t \rho(\mathbb{E}[|\Delta_{s \wedge \tau_R \wedge S_{\delta_0}}|]) ds.$$

When $\Delta_0 = 0$, Bihari’s inequality then implies that $\mathbb{E}[|\Delta_{t \wedge \tau_R \wedge S_{\delta_0}}|] = 0$. Hence by Fatou’s lemma, we have $\mathbb{E}[|\Delta_{t \wedge S_{\delta_0}}|] = 0$. This implies that $\Delta_{t \wedge S_{\delta_0}} = 0$ a.s.

On the set $\{S_{\delta_0} \leq t\}$, we have $|\Delta_{t \wedge S_{\delta_0}}| \geq \delta_0$. Thus it follows that $0 = \mathbb{E}[|\Delta_{t \wedge S_{\delta_0}}|] \geq \delta_0 \mathbb{P}\{S_{\delta_0} \leq t\}$. Then, we have $\mathbb{P}\{S_{\delta_0} \leq t\} = 0$ and hence $\Delta_t = 0$ a.s. The desired pathwise uniqueness for (21) then follows from the fact that \tilde{X} and X have right continuous sample paths. Next similar to the proof of Theorem 5.1 of [13], (21) has a weak solution, which further yields that the existence and uniqueness of a non-explosive strong solution to (21). □

Theorem 2.5 *Under Assumptions 2.1 and 2.2, for any $(x, k) \in \mathbb{R}^d \times \mathbb{S}$, the system given by (1) and (4) has a unique non-explosive strong solution (X, Λ) with initial condition $(X(0), \Lambda(0)) = (x, k)$.*

Proof. The proof is divided into two steps. First, we show that (1) and (4) has a non-explosive solution. The second step then derives the pathwise uniqueness for (1) and (4). While the proof of the existence of a solution to (1) and (4) use the same line of arguments as in the proof of Theorem 2.1 of [22], some care are required here since the assumptions in [22] have been relaxed. Moreover, an error in the proof of [22] is corrected here. The proof for pathwise uniqueness is more delicate than that

in [22] since the global Lipschitz conditions with respect to the variable x in [22] are no longer true in this paper.

Step 1. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$ be a complete filtered probability space, on which are defined a d -dimensional standard Brownian motion B , and a Poisson random measure $N(\cdot, \cdot)$ on $[0, \infty) \times U$ with a σ -finite characteristic measure ν on U . In addition, let $\{\xi_n\}$ be a sequence of independent exponential random variables with mean 1 on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$ that is independent of B and N .

Let $k \in \mathbb{S}$ and consider the stochastic differential equation

$$\begin{aligned}
 X^{(k)}(t) = x + \int_0^t b(X^{(k)}(s), k) ds + \int_0^t \sigma(X^{(k)}(s), k) dW(s) \\
 + \int_0^t \int_U c(X^{(k)}(s-), k, u) \tilde{N}(ds, du).
 \end{aligned}
 \tag{27}$$

Lemma 2.4 guarantees that SDE (27) has a unique non-explosive strong solution $X^{(k)}$. As in the proof of Theorem 2.1 of [22], we define

$$\tau_1 = \theta_1 := \inf \left\{ t \geq 0 : \int_0^t q_k(X^{(k)}(s)) ds > \xi_1 \right\}.
 \tag{28}$$

Thanks to (10), we have $\mathbb{P}(\tau_1 > 0) = 1$. We define a process (X, Λ) on $[0, \tau_1]$ as

$$X(t) = X^{(k)}(t) \text{ for all } t \in [0, \tau_1], \text{ and } \Lambda(t) = k \text{ for all } t \in [0, \tau_1].$$

Moreover, we define $\Lambda(\tau_1) \in \mathbb{S}$ according to the probability distribution

$$\mathbb{P} \{ \Lambda(\tau_1) = l | \mathcal{F}_{\tau_1-} \} = \frac{q_{kl}(X(\tau_1-))}{q_k(X(\tau_1-))} (1 - \delta_{kl}) \mathbf{1}_{\{q_k(X(\tau_1-)) > 0\}} + \delta_{kl} \mathbf{1}_{\{q_k(X(\tau_1-)) = 0\}},
 \tag{29}$$

for $l \in \mathbb{S}$. In general, having determined (X, Λ) on $[0, \tau_n]$, we let

$$\theta_{n+1} := \inf \left\{ t \geq 0 : \int_0^t q_{\Lambda(\tau_n)}(X^{\Lambda(\tau_n)}(s)) ds > \xi_{n+1} \right\},
 \tag{30}$$

where

$$\begin{aligned}
 X^{\Lambda(\tau_n)}(t) := X(\tau_n) + \int_0^t \sigma(X^{\Lambda(\tau_n)}(s), \Lambda(\tau_n)) dB(s) + \int_0^t b(X^{\Lambda(\tau_n)}(s), \Lambda(\tau_n)) ds \\
 + \int_0^t \int_U c(X^{\Lambda(\tau_n)}(s-), \Lambda(\tau_n), u) \tilde{N}(ds, du).
 \end{aligned}$$

As before, (10) implies that $\mathbb{P}\{\theta_{n+1} > 0\} = 1$. Then we let

$$\tau_{n+1} := \tau_n + \theta_{n+1}
 \tag{31}$$

and define (X, Λ) on $[\tau_n, \tau_{n+1}]$ by

$$X(t) = X^{\Lambda(\tau_n)}(t - \tau_n) \text{ for } t \in [\tau_n, \tau_{n+1}], \Lambda(t) = \Lambda(\tau_n) \text{ for } t \in [\tau_n, \tau_{n+1}),
 \tag{32}$$

and

$$\begin{aligned} \mathbb{P}\{\Lambda(\tau_{n+1}) = l | \mathcal{F}_{\tau_{n+1}-}\} &= \delta_{\Lambda(\tau_n), l} \mathbf{1}_{\{q_{\Lambda(\tau_n)}(X(\tau_{n+1}-))=0\}} \\ &+ \frac{q_{\Lambda(\tau_n), l}(X(\tau_{n+1}-))}{q_{\Lambda(\tau_n)}(X(\tau_{n+1}-))} (1 - \delta_{\Lambda(\tau_n), l}) \mathbf{1}_{\{q_{\Lambda(\tau_n)}(X(\tau_{n+1}-))>0\}}. \end{aligned} \tag{33}$$

As argued in [22], this ‘‘interlacing procedure’’ uniquely determines a solution $(X, \Lambda) \in \mathbb{R}^d \times \mathbb{S}$ to (1) and (4) for all $t \in [0, \tau_\infty)$, where

$$\tau_\infty = \lim_{n \rightarrow \infty} \tau_n. \tag{34}$$

Since the sequence τ_n is strictly increasing, the limit $\tau_\infty \leq \infty$ exists.

Next we show that $\tau_\infty = \infty$ a.s. To this end, fix $(X(0), \Lambda(0)) = (x, k) \in \mathbb{R}^d \times \mathbb{S}$ as in Step 1 and for any $m \geq k + 1$, denote by $\tilde{\tau}_m := \inf\{t \geq 0 : \Lambda(t) \geq m\}$ the first exit time for the Λ component from the finite set $\{0, 1, \dots, m - 1\}$. Let $A^c := \{\omega \in \Omega : \tau_\infty > \tilde{\tau}_m \text{ for all } m \geq k + 1\}$ and $A := \{\omega \in \Omega : \tau_\infty \leq \tilde{\tau}_{m_0} \text{ for some } m_0 \geq k + 1\}$. Then we have

$$\mathbb{P}\{\tau_\infty = \infty\} = \mathbb{P}\{\tau_\infty = \infty | A^c\} \mathbb{P}(A^c) + \mathbb{P}\{\tau_\infty = \infty | A\} \mathbb{P}(A). \tag{35}$$

Let $A_m := \{\omega \in \Omega : \tau_\infty \leq \tilde{\tau}_m\}$ for $m \geq k + 1$. Then $A = \bigcup_{m=k+1}^\infty A_m$, and $A^c = \bigcap_{m=k+1}^\infty A_m^c$. Also denote $B_{k+1} := A_{k+1}$ and let

$$B_m := A_m \setminus A_{m-1} = \{\omega \in \Omega : \tilde{\tau}_{m-1} < \tau_\infty \leq \tilde{\tau}_m\}$$

for $m \geq k + 2$. Clearly, $\{B_m\}_{m=k+1}^\infty$ is a sequence of disjoint sets and we have

$$A := \bigcup_{m=k+1}^\infty B_m. \tag{36}$$

We proceed to show that $\mathbb{P}\{\tau_\infty = \infty | B_m\} = 1$ for each m . Note that on the set B_m , $\Lambda(\tau_n) \leq m$ for all $n = 1, 2, \dots$. Consequently, using (10) in Assumption 2.1, we have $q_{\Lambda(\tau_n)}(X^{(\Lambda(\tau_n))}(s)) \leq Hm$ for all n and $s \geq 0$. On the other hand, thanks to the definition of θ_1 in (28), for any $\varepsilon > 0$, we have

$$\xi_1 < \int_0^{\theta_1 + \varepsilon} q_k(X^{(k)}(s)) ds.$$

Consequently, it follows that

$$\mathbf{1}_{B_m} \xi_1 \leq \mathbf{1}_{B_m} \int_0^{\theta_1 + \varepsilon} q_k(X^{(k)}(s)) ds \leq \mathbf{1}_{B_m} Hm(\theta_1 + \varepsilon).$$

In the same manner, we have from (30) that

$$\xi_n < \int_0^{\theta_n + \varepsilon/2^n} q_{\Lambda(\tau_{n-1})}(X^{(\Lambda(\tau_{n-1}))}(s)) ds,$$

and hence

$$\mathbf{1}_{B_m} \xi_n \leq \mathbf{1}_{B_m} \int_0^{\theta_n + \varepsilon/2^n} q_{\Lambda(\tau_{n-1})}(X^{\Lambda(\tau_{n-1})}(s)) ds \leq \mathbf{1}_{B_m} Hm(\theta_n + \varepsilon/2^n), \quad \forall n = 1, 2, \dots$$

Summing over these inequalities and noting that $\tau_\infty = \sum_{n=1}^\infty \theta_n$, we arrive at

$$\mathbf{1}_{B_m} \sum_{n=1}^\infty \xi_n \leq \mathbf{1}_{B_m} Hm(\tau_\infty + 2\varepsilon). \tag{37}$$

By virtue of Theorem 2.3.2 of [18], we have $\sum_{n=1}^\infty \xi_n = \infty$ a.s. Therefore it follows that $\mathbb{P}(\sum_{n=1}^\infty \xi_n = \infty | B_m) = 1$. Then (37) implies that

$$\mathbb{P}\{\tau_\infty = \infty | B_m\} \geq \mathbb{P}\left\{\sum_{n=1}^\infty \xi_n = \infty | B_m\right\} = 1,$$

as desired. Consequently, we can use (36) to compute

$$\begin{aligned} \mathbb{P}\{\tau_\infty = \infty | A\} &= \frac{\mathbb{P}\{\tau_\infty = \infty, A\}}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}\{\tau_\infty = \infty, \bigcup_{m=k+1}^\infty B_m\}}{\mathbb{P}(A)} \\ &= \frac{\sum_{m=k+1}^\infty \mathbb{P}\{\tau_\infty = \infty, B_m\}}{\mathbb{P}(A)} \\ &= \frac{\sum_{m=k+1}^\infty \mathbb{P}\{\tau_\infty = \infty | B_m\} \mathbb{P}(B_m)}{\mathbb{P}(A)} \\ &= \frac{\sum_{m=k+1}^\infty \mathbb{P}(B_m)}{\mathbb{P}(A)} = 1. \end{aligned} \tag{38}$$

If $\mathbb{P}(A) = 1$ or $\mathbb{P}(A^c) = 0$, then (35) and (38) imply that $\mathbb{P}\{\tau_\infty = \infty\} = 1$ and the proof is complete. Therefore, it remains to consider the case when $\mathbb{P}(A^c) > 0$. Denote $\tilde{\tau}_\infty := \lim_{m \rightarrow \infty} \tilde{\tau}_m$. Note that $A^c = \{\tau_\infty \geq \tilde{\tau}_\infty\}$. Thus $\mathbb{P}\{\tau_\infty = \infty | A^c\} \geq \mathbb{P}\{\tilde{\tau}_\infty = \infty | A^c\}$ and hence (35) holds if we can show that

$$\mathbb{P}\{\tilde{\tau}_\infty = \infty | A^c\} = 1. \tag{39}$$

Assume on the contrary that (39) were false, then there would exist a $T > 0$ such that

$$\delta := \mathbb{P}\{\tilde{\tau}_\infty \leq T, A^c\} > 0.$$

Let $f : \mathbb{S} \mapsto \mathbb{R}_+$ be as in Assumption 2.1. Then we have for any $m \geq k + 1$,

$$\begin{aligned} f(k) &= \mathbb{E}[e^{-H(T \wedge \tau_\infty \wedge \tilde{\tau}_m)} f(\Lambda(T \wedge \tau_\infty \wedge \tilde{\tau}_m))] \\ &+ \mathbb{E}\left[\int_0^{T \wedge \tau_\infty \wedge \tilde{\tau}_m} e^{-Hs} \left(Hf(\Lambda(s)) - \sum_{l \in \mathbb{S}} q_{\Lambda(s), l}(X(s)) [f(l) - f(\Lambda(s))]\right) ds\right] \end{aligned}$$

$$\begin{aligned} &\geq \mathbb{E}[e^{-H(T \wedge \tau_\infty \wedge \tilde{\tau}_m)} f(\Lambda(T \wedge \tau_\infty \wedge \tilde{\tau}_m))] \\ &\quad + \mathbb{E}\left[\int_0^{T \wedge \tau_\infty \wedge \tilde{\tau}_m} e^{-Hs} [Hf(\Lambda(s)) - H(1 + \Phi(X(s)) + f(\Lambda(s)))] ds\right] \\ &\geq \mathbb{E}[e^{-H(T \wedge \tau_\infty \wedge \tilde{\tau}_m)} f(\Lambda(T \wedge \tau_\infty \wedge \tilde{\tau}_m))], \end{aligned}$$

where the first inequality above follows from (11) in Assumption 2.1. Consequently, we have

$$\begin{aligned} e^{HT} f(k) &\geq \mathbb{E}[f(\Lambda(T \wedge \tau_\infty \wedge \tilde{\tau}_m))] \geq \mathbb{E}[f(\Lambda(\tilde{\tau}_m)) \mathbf{1}_{\{\tilde{\tau}_m \leq T \wedge \tau_\infty\}}] \\ &\geq f(m) \mathbb{P}\{\tilde{\tau}_m \leq T \wedge \tau_\infty\} \geq f(m) \mathbb{P}\{\tilde{\tau}_m \leq T \wedge \tau_\infty, A^c\} \tag{40} \\ &\geq f(m) \mathbb{P}\{\tilde{\tau}_\infty \leq T \wedge \tau_\infty, A^c\}, \end{aligned}$$

where the third inequality follows from the facts that $\Lambda(\tilde{\tau}_m) \geq m$ and that f is non-decreasing, and the last inequality follows from the fact that $\tilde{\tau}_m \uparrow \tilde{\tau}_\infty$. Recall that $A^c = \{\tau_\infty \geq \tilde{\tau}_\infty\}$. Thus

$$\begin{aligned} \mathbb{P}\{\tilde{\tau}_\infty \leq T \wedge \tau_\infty, A^c\} &= \mathbb{P}\{\tilde{\tau}_\infty \leq T \wedge \tau_\infty, \tilde{\tau}_\infty \leq \tau_\infty\} \\ &\geq \mathbb{P}\{\tilde{\tau}_\infty \leq T, \tilde{\tau}_\infty \leq \tau_\infty\} = \mathbb{P}\{\tilde{\tau}_\infty \leq T, A^c\} = \delta > 0. \end{aligned}$$

Using this observation in (40) yields $\infty > e^{HT} f(k) \geq f(m) \delta \rightarrow \infty$ as $m \rightarrow \infty$, thanks to the fact that $f(m) \rightarrow \infty$ as $m \rightarrow \infty$, which is a contradiction. This establishes (39) and hence $\mathbb{P}(\tau_\infty = \infty) = 1$. In other words, the interlacing procedure uniquely determines a solution $(X, \Lambda) = (X^{(x,k)}, \Lambda^{(x,k)})$ for all $t \in [0, \infty)$.

Next we show that the solution (X, Λ) to the system (1) and (4) is non-explosive a.s. Consider the function $V(x, k) := 1 + \Phi(x) + f(k)$, where the functions $\Phi : \mathbb{R}^d \mapsto \mathbb{R}_+$ of (12) and $f : \mathbb{S} \mapsto \mathbb{R}_+$ are defined in Assumption 2.1. Note that $V(x, k) \rightarrow \infty$ as $|x| \vee k \rightarrow \infty$ thanks to Assumption 2.1. Using the definition of \mathcal{A} of (5), we have

$$\mathcal{A}V(x, k) = \mathcal{L}_k \Phi(x) + Q(x)f(k).$$

Moreover, detailed computations using (8) and (9) reveal that $\mathcal{L}_k \Phi(x) \leq H\Phi(x)$ for all $x \in \mathbb{R}^d$ and $k \in \mathbb{S}$. On the other hand, (11) implies that $Q(x)f(k) \leq H(1 + \Phi(x) + f(k))$. Combining these estimates, we obtain $\mathcal{A}V(x, k) \leq 2HV(x, k)$. This, together with Itô’s formula, shows that the process $\{e^{-2Ht}V(X(t), \Lambda(t)), t \geq 0\}$ is a nonnegative local supermartingale. Then we can apply the optional sampling theorem to the process $\{e^{-2Ht}V(X(t), \Lambda(t)), t \geq 0\}$ to argue that $\mathbb{P}\{\lim_{n \rightarrow \infty} T_n = \infty\} = 1$, where $T_n := \inf\{t \geq 0 : |X(t)| \vee \Lambda(t) \geq n\}$. This shows that the solution (X, Λ) has no finite explosion time a.s.

Step 2. Suppose (X, Λ) and $(\tilde{X}, \tilde{\Lambda})$ are two solutions to (1) and (4) starting from the same initial condition $(x, k) \in \mathbb{R}^d \times \mathbb{S}$. Then we have

$$\begin{aligned} \tilde{X}(t) - X(t) &= \int_0^t [b(\tilde{X}(s), \tilde{\Lambda}(s)) - b(X(s), \Lambda(s))] ds \\ &\quad + \int_0^t [\sigma(\tilde{X}(s), \tilde{\Lambda}(s)) - \sigma(X(s), \Lambda(s))] dW(s) \end{aligned}$$

$$+ \int_U [c(\tilde{X}(s-), \tilde{\Lambda}(s-), z) - c(X(s-), \Lambda(s-), z)] \tilde{N}(ds, du),$$

and

$$\tilde{\Lambda}(t) - \Lambda(t) = \int_0^t \int_{\mathbb{R}_+} [h(\tilde{X}(s-), \tilde{\Lambda}(s-), z) - h(X(s-), \Lambda(s-), z)] N_1(ds, dz).$$

Let $\zeta := \inf\{t \geq 0 : \Lambda(t) \neq \tilde{\Lambda}(t)\}$ be the first time when the discrete components differ from each other. Let us also define $T_R := \inf\{t \geq 0 : |\tilde{X}(t)| \vee |X(t)| \vee \tilde{\Lambda}(t) \vee \Lambda(t) \geq R\}$ for $R > 0$ and $S_{\delta_0} := \inf\{t \geq 0 : |\tilde{X}(t) - X(t)| \geq \delta_0\}$. We have $\Lambda(t) = \tilde{\Lambda}(t)$ for $t \in [0, \zeta)$. To simplify notation, let us define $\Delta_t := \tilde{X}(t) - X(t)$. Then from the proof of Theorem 2.6 of [24], we have $\mathbb{E}[H(|\Delta_{t \wedge \zeta \wedge S_{\delta_0}}|)] = 0$ for $d \geq 2$, where $H(r) := \frac{r^2}{1+r^2}$, $r \geq 0$. When $d = 1$, the proof of Lemma 2.4 reveals that $\mathbb{E}[|\Delta_{t \wedge \zeta \wedge S_{\delta_0}}|] = 0$ and hence $\mathbb{E}[H(|\Delta_{t \wedge \zeta \wedge S_{\delta_0}}|)] = 0$. Note that on the set $\{S_{\delta_0} \leq t \wedge \zeta\}$, $|\Delta_{t \wedge \zeta \wedge S_{\delta_0}}| \geq \delta_0$. Also we can readily check that H is an increasing function. Thus, it follows that

$$0 = \mathbb{E}[H(|\Delta_{t \wedge \zeta \wedge S_{\delta_0}}|)] \geq \mathbb{E}[H(|\Delta_{t \wedge \zeta \wedge S_{\delta_0}}|) \mathbf{1}_{\{S_{\delta_0} \leq t \wedge \zeta\}}] \geq H(\delta_0) \mathbb{P}\{S_{\delta_0} \leq t \wedge \zeta\}.$$

This implies that $\mathbb{P}\{S_{\delta_0} \leq t \wedge \zeta\} = 0$. Consequently, we have

$$\begin{aligned} \mathbb{E}[H(|\Delta_{t \wedge \zeta}|)] &= \mathbb{E}[H(|\Delta_{t \wedge \zeta}|) \mathbf{1}_{\{S_{\delta_0} \leq t \wedge \zeta\}}] + \mathbb{E}[H(|\Delta_{t \wedge \zeta}|) \mathbf{1}_{\{S_{\delta_0} > t \wedge \zeta\}}] \\ &\leq \mathbb{P}\{S_{\delta_0} \leq t \wedge \zeta\} + \mathbb{E}[H(|\Delta_{t \wedge \zeta \wedge S_{\delta_0}}|) \mathbf{1}_{\{S_{\delta_0} > t \wedge \zeta\}}] \\ &\leq 0. \end{aligned}$$

It follows that $\mathbb{E}[|\Delta_{t \wedge \zeta}|] = \mathbb{E}[|\tilde{X}(t \wedge \zeta) - X(t \wedge \zeta)|] = 0$. Then we have

$$\mathbb{E}[|\tilde{X}(t \wedge \zeta) - X(t \wedge \zeta)|^\delta] = 0, \tag{41}$$

where $\delta \in (0, 1]$ is the Hölder constant in (13).

Note that $\zeta \leq t$ if and only if $\tilde{\Lambda}(t \wedge \zeta) - \Lambda(t \wedge \zeta) \neq 0$. Therefore, it follows that

$$\begin{aligned} \mathbb{P}\{\zeta \leq t\} &= \mathbb{E}[\mathbf{1}_{\{\tilde{\Lambda}(t \wedge \zeta) - \Lambda(t \wedge \zeta) \neq 0\}}] \\ &= \mathbb{E} \left[\int_0^{t \wedge \zeta} \int_{\mathbb{R}_+} (\mathbf{1}_{\{\tilde{\Lambda}(s-) - \Lambda(s-) + h(\tilde{X}(s-), \Lambda(s-), z) - h(X(s-), \Lambda(s-), z) \neq 0\}} \right. \\ &\quad \left. - \mathbf{1}_{\{\tilde{\Lambda}(s-) - \Lambda(s-) \neq 0\}}) \mathbf{m}(dz) ds \right] \\ &= \mathbb{E} \left[\int_0^{t \wedge \zeta} \int_{\mathbb{R}_+} \mathbf{1}_{\{h(\tilde{X}(s-), \Lambda(s-), z) - h(X(s-), \Lambda(s-), z) \neq 0\}} \mathbf{m}(dz) ds \right] \\ &\leq \mathbb{E} \left[\int_0^{t \wedge \zeta} \sum_{l \in \mathbb{S}, l \neq \Lambda(s-)} |q_{\Lambda(s-), l}(\tilde{X}(s-)) - q_{\Lambda(s-), l}(X(s-))| ds \right] \end{aligned}$$

$$\begin{aligned} &\leq H\mathbb{E}\left[\int_0^{t\wedge\zeta} |\tilde{X}(s-) - X(s-)|^\delta ds\right] \\ &\leq H\int_0^t \mathbb{E}[|\tilde{X}(s\wedge\zeta) - X(s\wedge\zeta)|^\delta] ds = 0, \end{aligned}$$

where the second inequality follows from (13). In particular, we have

$$\mathbb{E}[\mathbf{1}_{\{\tilde{\Lambda}(t)\neq\Lambda(t)\}}] \leq \mathbb{P}\{\zeta \leq t\} = 0. \tag{42}$$

Now we can compute

$$\begin{aligned} \mathbb{E}[H(|\tilde{X}(t) - X(t)|)] &= \mathbb{E}[H(|\tilde{X}(t) - X(t)|)\mathbf{1}_{\{\zeta>t\}}] + \mathbb{E}[H(|\tilde{X}(t) - X(t)|)\mathbf{1}_{\{\zeta\leq t\}}] \\ &= \mathbb{E}[H(|\tilde{X}(t\wedge\zeta) - X(t\wedge\zeta)|)\mathbf{1}_{\{\zeta>t\}}] + \mathbb{E}[1\cdot\mathbf{1}_{\{\zeta\leq t\}}] \\ &\leq \mathbb{E}[H(|\tilde{X}(t\wedge\zeta) - X(t\wedge\zeta)|)] + 0 \\ &= 0. \end{aligned}$$

Thus $\mathbb{P}\{\tilde{X}(t) = X(t)\} = 1$. This, together with (42), implies that $\mathbb{P}\{(\tilde{X}(t), \tilde{\Lambda}(t)) = (X(t), \Lambda(t))\} = 1$ for all $t \geq 0$. Since the sample paths of (X, Λ) are right continuous, we obtain the desired pathwise uniqueness result. \square

Example 2.6 Let us consider the following SDE

$$\begin{aligned} dX(t) &= b(X(t), \Lambda(t))dt + \sigma(X(t), \Lambda(t))dW(t) \\ &\quad + \int_U c(X(t-), \Lambda(t-), u)\tilde{N}(dt, du), \quad X(0) = x \in \mathbb{R}^3, \end{aligned} \tag{43}$$

where W is a 3-dimensional standard Brownian motion, $\tilde{N}(dt, du)$ is a compensated Poisson random measure with compensator $dt \nu(du)$ on $[0, \infty) \times U$, in which $U = \{u \in \mathbb{R}^3 : 0 < |u| < 1\}$ and $\nu(du) := \frac{du}{|u|^{3+\alpha}}$ for some $\alpha \in (0, 2)$. The Λ component in (43) takes value in $\mathbb{S} = \{1, 2, \dots\}$ and is generated by $Q(x) = (q_{kl}(x))$, with $q_{kl}(x) = \frac{k}{2l} \cdot \frac{|x|^2}{1+|x|^2}$ for $x \in \mathbb{R}^3$ and $k \neq l \in \mathbb{S}$. Let $q_k(x) = -q_{kk}(x) = \sum_{l \neq k} q_{kl}(x)$. The coefficients of (43) are given by

$$b(x, k) = \begin{pmatrix} -x_1^{1/3} - kx_1^3 \\ -x_2^{1/3} - kx_2^3 \\ -x_3^{1/3} - kx_3^3 \end{pmatrix}, \quad c(x, k, u) = c(x, u) = \begin{pmatrix} \gamma x_1^{2/3} |u| \\ \gamma x_2^{2/3} |u| \\ \gamma x_3^{2/3} |u| \end{pmatrix},$$

and

$$\sigma(x, k) = \begin{pmatrix} \frac{x_1^{2/3}}{\sqrt{2}} + 1 & \frac{\sqrt{k}x_2^2}{3} & \frac{\sqrt{k}x_3^2}{3} \\ \frac{\sqrt{k}x_1^2}{3} & \frac{x_2^{2/3}}{\sqrt{2}} + 1 & \frac{\sqrt{k}x_3^2}{3} \\ \frac{\sqrt{k}x_1^2}{3} & \frac{\sqrt{k}x_2^2}{3} & \frac{x_3^{2/3}}{\sqrt{2}} + 1 \end{pmatrix},$$

in which γ is a positive constant so that $\gamma^2 \int_U |u|^2 \nu(du) = \frac{1}{2}$.

Note that σ and b grow very fast in the neighborhood of ∞ and they are Hölder continuous with orders $\frac{2}{3}$ and $\frac{1}{3}$, respectively. Nevertheless, the coefficients of (43) still satisfy Assumptions 2.2 and 2.1 and hence a unique non-exploding strong solution of (43) exists. The verifications of these assumptions are as follows.

$$\begin{aligned} & 2\langle x, b(x, k) \rangle + |\sigma(x, k)|^2 + \int_U |c(x, k, u)|^2 \nu(du) \\ &= 2 \sum_{j=1}^3 x_j (-x_j^{1/3} - kx_j^3) + \sum_{j=1}^3 \left(\frac{1}{2} x_j^{4/3} + \frac{2k}{9} x_j^4 + \sqrt{2} x_j^{2/3} + 1 \right) \\ &\quad + \int_U \gamma^2 |u|^2 \sum_{j=1}^3 x_j^{4/3} \nu(du) \\ &= -\frac{16k}{9} \sum_{j=1}^3 x_j^4 - \sum_{j=1}^3 x_j^{4/3} + \sqrt{2} \sum_{j=1}^3 x_j^{2/3} + 3. \end{aligned}$$

Thus (9) of Assumption 2.1 hold. Furthermore, (10) is trivially satisfied. Consider the function $f(l) = l, l \in \mathbb{S}$. We have

$$\sum_{l \neq k} (f(l) - f(k)) q_{kl}(x) = \sum_{l \neq k} (l - k) \frac{k}{2^l} \frac{|x|^2}{1 + |x|^2} \leq \sum_{l \neq k} l \frac{k}{2^l} \frac{|x|^2}{1 + |x|^2} \leq k \sum_{l \in \mathbb{S}} \frac{l}{2^l} = 2k,$$

which yields (11). If $x, y \in \mathbb{R}^3$, we obtain

$$\begin{aligned} \sum_{l \neq k} |q_{kl}(x) - q_{kl}(y)| &= \sum_{l \neq k} \frac{l}{2^k} \left| \frac{|x|^2}{1 + |y|^2} - \frac{|x|^2}{1 + |y|^2} \right| = \sum_{l \neq k} \frac{l}{2^k} \frac{||x| - |y|| (|x| + |y|)}{(1 + |x|^2)(1 + |y|^2)} \\ &\leq \sum_{l \neq k} \frac{l}{2^k} |x - y| \left(\frac{|x|}{1 + |x|^2} + \frac{|y|}{1 + |y|^2} \right) \leq 2|x - y|. \end{aligned}$$

This establishes (13) and therefore verifies Assumption 2.1.

For the verification of Assumption 2.2, we compute

$$\begin{aligned} & 2\langle x - y, b(x, k) - b(y, k) \rangle + |\sigma(x, k) - \sigma(y, k)|^2 + \int_U |c(x, k, u) - c(y, k, u)|^2 \nu(du) \\ &= -2 \sum_{j=1}^3 (x_j - y_j) (x_j^{1/3} - y_j^{1/3} + kx_j^3 - ky_j^3) + \frac{1}{2} \sum_{j=1}^3 (x_j^{2/3} - y_j^{2/3})^2 \\ &\quad + \frac{2k}{9} \sum_{j=1}^3 (x_j^2 - y_j^2)^2 + \int_U \sum_{j=1}^3 \gamma^2 (x_j^{2/3} - y_j^{2/3})^2 |u|^2 \nu(du) \\ &= -\frac{16k}{9} \sum_{j=1}^3 (x_j - y_j)^2 \left[\left(x_j + \frac{7}{16} y_j \right)^2 + \frac{207}{256} y_j^2 \right] - \sum_{j=1}^3 (x_j^{1/3} - y_j^{1/3})^2 (x_j^{2/3} + y_j^{2/3}). \end{aligned}$$

Obviously this implies (20) and thus verifies Assumption 2.2.

3 Feller Property

In Section 2, we established the existence and uniqueness of a solution in the strong sense to system (1) and (4) under Assumptions 2.1 and 2.2. The solution (X, Λ) is a two-component càdlàg strong Markov process. In this section, we study the Feller property for such processes. For any $f \in C_b(\mathbb{R}^d \times \mathbb{S})$, by the continuity of f and the right continuity of the sample paths of (X, Λ) , we can use the bounded convergence theorem to obtain $\lim_{t \downarrow 0} \mathbb{E}_{x,k}[f(X(t), \Lambda(t))] = f(x, k)$. Therefore the process (X, Λ) satisfies the Feller property if the semigroup $P_t f(x, k) := \mathbb{E}_{x,k}[f(X(t), \Lambda(t))]$, $f \in \mathfrak{B}_b(\mathbb{R}^d \times \mathbb{S})$ maps $C_b(\mathbb{R}^d \times \mathbb{S})$ into itself. Obviously, to establish the Feller property, we only need the distributional properties of the process (X, Λ) . Thus in lieu of the strong formulation used in Section 2, we will assume the following “weak formulation” throughout the section.

Assumption 3.1 For any initial data $(x, k) \in \mathbb{R}^d \times \mathbb{S}$, the system of stochastic differential equations (1) and (4) has a non-exploding weak solution $(X^{(x,k)}, \Lambda^{(x,k)})$ and the solution is unique in the sense of probability law.

Assumption 3.2 There exist a positive constant δ_0 and an increasing and concave function $\rho : [0, \infty) \mapsto [0, \infty)$ satisfying (19) such that for all $R > 0$, there exists a constant $\kappa_R > 0$ such that

$$\sum_{l \in \mathbb{S} \setminus \{k\}} |q_{kl}(x) - q_{kl}(z)| \leq \kappa_R \rho(F(|x - z|)), \text{ for all } k \in \mathbb{S} \text{ and } |x| \vee |z| \leq R \quad (44)$$

where $F(r) := \frac{r}{1+r}$ for $r \geq 0$, and either (i) or (ii) below holds:

- (i) $d = 1$. Then (15) and (17) hold.
- (ii) $d \geq 2$. Then

$$\begin{aligned} & \int_U [|c(x, k, u) - c(z, k, u)|^2 \wedge (4|x - z| \cdot |c(x, k, u) - c(z, k, u)|)] \nu(du) \\ & + 2\langle x - z, b(x, k) - b(z, k) \rangle + |\sigma(x, k) - \sigma(z, k)|^2 \leq 2\kappa_R |x - z| \rho(|x - z|), \end{aligned} \quad (45)$$

for all $k \in \mathbb{S}$, $x, z \in \mathbb{R}^d$ with $|x| \vee |z| \leq R$ and $|x - z| \leq \delta_0$.

Theorem 3.3 Under Assumptions 3.1 and 3.2, the process (X, Λ) possesses the Feller property.

Remark 3.4 Feller and strong Feller properties for regime-switching (jump) diffusions have been investigated in [20, 22, 27], among others. A standard assumption in these references is that the coefficients satisfy the Lipschitz condition. In contrast, Theorem 3.3 establishes Feller property for system (1) and (4) under local non-Lipschitz conditions. When $d = 1$, the result is even more remarkable. Indeed, Feller property is derived with only very mild conditions on $b(\cdot, k)$, $c(\cdot, k, u)$, and $Q(x)$, and with virtually no condition imposed on $\sigma(\cdot, k)$.

We will use the coupling method to prove Theorem 3.3. To this end, let us first construct a coupling operator $\tilde{\mathcal{A}}$ for \mathcal{A} : For $f(x, i, z, j) \in C_c^2(\mathbb{R}^d \times \mathbb{S} \times \mathbb{R}^d \times \mathbb{S})$, we define

$$\tilde{\mathcal{A}}f(x, i, z, j) := [\tilde{\Omega}_d + \tilde{\Omega}_j + \tilde{\Omega}_s]f(x, i, z, j), \tag{46}$$

where $\tilde{\Omega}_d$, $\tilde{\Omega}_j$, and $\tilde{\Omega}_s$ are defined as follows. For $x, z \in \mathbb{R}^d$ and $i, j \in \mathbb{S}$, we set $a(x, i) = \sigma(x, i)\sigma(x, i)'$ and

$$a(x, i, z, j) = \begin{pmatrix} a(x, i) & \sigma(x, i)\sigma(z, j)' \\ \sigma(z, j)\sigma(x, i)' & a(z, j) \end{pmatrix}, \quad b(x, i, z, j) = \begin{pmatrix} b(x, i) \\ b(z, j) \end{pmatrix}.$$

Then we define

$$\tilde{\Omega}_d f(x, i, z, j) := \frac{1}{2} \text{tr}(a(x, i, z, j)D^2 f(x, i, z, j)) + \langle b(x, i, z, j), Df(x, i, z, j) \rangle, \tag{47}$$

$$\begin{aligned} \tilde{\Omega}_j f(x, i, z, j) := & \int_U [f(x + c(x, i, u), i, z + c(z, j, u), j) - f(x, i, z, j) \\ & - \langle D_x f(x, i, z, j), c(x, i, u) \rangle - \langle D_z f(x, i, z, j), c(z, j, u) \rangle] \nu(du), \end{aligned} \tag{48}$$

where $Df(x, i, z, j) = (D_x f(x, i, z, j), D_z f(x, i, z, j))'$ is the gradient and $D^2 f(x, i, z, j)$ the Hessian matrix of f with respect to the x, z variables, and

$$\begin{aligned} \tilde{\Omega}_s f(x, i, z, j) := & \sum_{l \in \mathbb{S}} [q_{il}(x) - q_{jl}(z)]^+ (f(x, l, z, j) - f(x, i, z, j)) \\ & + \sum_{l \in \mathbb{S}} [q_{jl}(z) - q_{il}(x)]^+ (f(x, i, z, l) - f(x, i, z, j)) \\ & + \sum_{l \in \mathbb{S}} [q_{il}(x) \wedge q_{jl}(z)] (f(x, l, z, l) - f(x, i, z, j)). \end{aligned} \tag{49}$$

For convenience of later presentation, for any function $f : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$, let $\tilde{f} : \mathbb{R}^d \times \mathbb{S} \times \mathbb{R}^d \times \mathbb{S} \mapsto \mathbb{R}$ be defined by $\tilde{f}(x, i, z, j) := f(x, z)$. Now we denote

$$\tilde{\mathcal{L}}_k f(x, z) = (\tilde{\Omega}_d^{(k)} + \tilde{\Omega}_j^{(k)})f(x, z) := (\tilde{\Omega}_d + \tilde{\Omega}_j)\tilde{f}(x, k, z, k), \forall f \in C_c^2(\mathbb{R}^d \times \mathbb{R}^d)$$

for each $k \in \mathbb{S}$. We proceed to establish the following lemma.

Lemma 3.5 *Suppose Assumption 3.2 holds. Consider the functions*

$$g(x, k, z, l) := \mathbf{1}_{\{k \neq l\}}, \text{ and } f(x, k, z, l) := F(|x - z|) + \mathbf{1}_{\{k \neq l\}}, \tag{50}$$

for $(x, k, z, l) \in \mathbb{R}^d \times \mathbb{S} \times \mathbb{R}^d \times \mathbb{S}$. Then we have

$$\tilde{\mathcal{A}}g(x, k, z, l) \leq \kappa_R \rho(F(|x - y|)), \text{ for all } k, l \in \mathbb{S} \text{ and } x, z \in \mathbb{R}^d \text{ with } |x| \vee |z| \leq R, \tag{51}$$

and

$$\tilde{\mathcal{A}}f(x, k, z, k) \leq 2\kappa_R \rho(F(|x - z|)), \tag{52}$$

for all $k \in \mathbb{S}$ and $x, z \in \mathbb{R}^d$ with $|x| \vee |z| \leq R$ and $0 < |x - z| \leq \delta_0$; in which κ_R is the same positive constant as in Assumption 3.2.

Proof. Consider the function $g(x, k, z, l) := \mathbf{1}_{\{k \neq l\}}$. It follows directly from the definition that $\widetilde{\mathcal{A}}g(x, k, z, l) = \widetilde{\Omega}_s g(x, k, z, l) \leq 0$ when $k \neq l$. When $k = l$, we have from (44) that

$$\begin{aligned} \widetilde{\mathcal{A}}g(x, k, z, l) &= \widetilde{\Omega}_s g(x, k, z, k) \\ &= \sum_{i \in \mathbb{S}} [q_{ki}(x) - q_{ki}(z)]^+ (\mathbf{1}_{\{i \neq k\}} - \mathbf{1}_{\{k \neq k\}}) \\ &\quad + \sum_{i \in \mathbb{S}} [q_{ki}(z) - q_{ki}(x)]^+ (\mathbf{1}_{\{i \neq k\}} - \mathbf{1}_{\{k \neq k\}}) + 0 \\ &\leq \sum_{i \in \mathbb{S}, i \neq k} |q_{ki}(x) - q_{ki}(z)| \leq \kappa_R \rho(F(|x - y|)). \end{aligned} \tag{53}$$

Hence (51) holds for all $k, l \in \mathbb{S}$ and $x, z \in \mathbb{R}^d$ with $|x| \vee |z| \leq R$. On the other hand, when $d \geq 2$, (45) and Lemma 4.5 of [24] reveals that

$$\widetilde{\mathcal{L}}_k F(|x - z|) = (\widetilde{\Omega}_d^{(k)} + \widetilde{\Omega}_j^{(k)}) F(|x - z|) \leq \kappa_R \rho(F(|x - z|))$$

and hence

$$\widetilde{\mathcal{A}}f(x, k, z, k) = \widetilde{\mathcal{L}}_k F(|x - z|) + \widetilde{\Omega}_s g(x, k, z, k) \leq 2\kappa_R \rho(F(|x - z|)), \tag{54}$$

for all $k \in \mathbb{S}$, $x, z \in \mathbb{R}^d$ with $|x| \vee |z| \leq R$ and $0 < |x - z| \leq \delta_0$, where $\widetilde{\mathcal{L}}_k$ is the basic coupling operator for \mathcal{L}_k of (6).

We next show that (52) holds when $d = 1$. Indeed, taking advantage of the fact that $d = 1$, we see that

$$\begin{aligned} &\widetilde{\Omega}_d^{(k)} F(|x - z|) \\ &= F'(|x - z|) \operatorname{sgn}(x - z) (b(x, k) - b(z, k)) + F''(|x - z|) (\sigma(x, k) - \sigma(z, k))^2. \end{aligned}$$

But since $F'(r) = \frac{1}{(1+r)^2}$ and $F''(r) = -\frac{2}{(1+r)^3} < 0$ for $r \geq 0$, we have from (15) that

$$\begin{aligned} \widetilde{\Omega}_d^{(k)} F(|x - z|) &\leq \frac{1}{(1 + |x - z|)^2} \operatorname{sgn}(x - z) (b(x, k) - b(z, k)) \\ &\leq \frac{\kappa_R \rho(|x - z|)}{(1 + |x - z|)^2} \leq \kappa_R \rho(F(|x - z|)), \end{aligned} \tag{55}$$

for all $x, z \in \mathbb{R}$ with $|x| \vee |z| \leq R$ and $0 < |x - z| \leq \delta_0$, where we used the first equation in (19) to derive the last inequality.

On the other hand, since the function F is concave on $[0, \infty)$, we have $F(r) - F(r_0) \leq F'(r_0)(r - r_0)$ for all $r, r_0 \in [0, \infty)$. Applying this inequality with $r_0 = |x - z|$ and $r = |x - z + c(x, k, u) - c(z, k, u)|$ yields

$$\begin{aligned}
 & F(|x - z + c(x, k, u) - c(z, k, u)|) - F(|x - z|) \\
 & \leq F'(|x - z|)(|x - z + c(x, k, u) - c(z, k, u)| - |x - z|).
 \end{aligned}$$

Furthermore, since by (17), the function $x \mapsto x + c(x, k, u)$ is increasing, it follows that for $x > z$

$$\begin{aligned}
 & F(|x - z + c(x, k, u) - c(z, k, u)|) - F(|x - z|) \\
 & \leq F'(|x - z|)(x - z + c(x, k, u) - c(z, k, u) - (x - z)) \\
 & = F'(|x - z|)(c(x, k, u) - c(z, k, u)).
 \end{aligned}$$

As a result, we can compute

$$\begin{aligned}
 & \tilde{\Omega}_j^{(k)} F(|x - z|) \\
 & = \int_U [F(|x - z + c(x, k, u) - c(z, k, u)|) - F(|x - z|) \\
 & \quad - F'(|x - z|)\text{sgn}(x - z)(c(x, k, u) - c(z, k, u))] \nu(\mathbf{d}u) \\
 & \leq \int_U [F'(|x - z|)(c(x, k, u) - c(z, k, u)) - F'(|x - z|)(c(x, k, u) - c(z, k, u))] \nu(\mathbf{d}u) \\
 & = 0,
 \end{aligned}$$

for all $x > z$. By symmetry, we also have $\tilde{\Omega}_j^{(k)} F(|x - z|) \leq 0$ for $x < z$. These observations, together with (53) and (55), imply that

$$\tilde{\mathcal{A}}f(x, k, z, k) = (\tilde{\Omega}_d^{(k)} + \tilde{\Omega}_j^{(k)})F(|x - z|) + \tilde{\Omega}_s g(x, k, z, k) \leq 2\kappa_R \rho(F(|x - z|)),$$

for all $k \in \mathbb{S}$, $x, z \in \mathbb{R}$ with $|x| \vee |z| \leq R$ and $0 < |x - z| \leq \delta_0$. This completes the proof. \square

Proof (Proof of Theorem 3.3). It is straightforward to verify that the function f of (50) defines a bounded metric on $\mathbb{R}^d \times \mathbb{S}$. Let $(\tilde{X}(\cdot), \tilde{\Lambda}(\cdot), \tilde{Z}(\cdot), \tilde{\Xi}(\cdot))$ denote the coupling process corresponding to the coupling operator $\tilde{\mathcal{A}}$ with initial condition (x, k, z, k) , in which $\delta_0 > |x - z| > 0$. Define $\zeta := \inf\{t \geq 0 : \tilde{\Lambda}(t) \neq \tilde{\Xi}(t)\}$. Note that $\mathbb{P}\{\zeta > 0\} = 1$. Suppose $|x - z| > \frac{1}{n_0}$ for some $n_0 \in \mathbb{N}$. For $n \geq n_0$ and $R > |x| \vee |z|$, define

$$\begin{aligned}
 T_n & := \inf\left\{t \geq 0 : |\tilde{X}(t) - \tilde{Z}(t)| < \frac{1}{n}\right\}, \\
 \tau_R & := \inf\{t \geq 0 : |\tilde{X}(t)| \vee |\tilde{Z}(t)| \vee \tilde{\Lambda}(t) \vee \tilde{\Xi}(t) > R\},
 \end{aligned}$$

and

$$S_{\delta_0} := \inf\{t \geq 0 : |\tilde{X}(t) - \tilde{Z}(t)| > \delta_0\}.$$

We have $\tau_R \rightarrow \infty$ and $T_n \rightarrow T$ a.s. as $R \rightarrow \infty$ and $n \rightarrow \infty$, respectively, in which T denotes the first time when $\tilde{X}(t)$ and $\tilde{Z}(t)$ coalesce. To simplify notation, denote

$\tilde{\Delta}(s) := \tilde{X}(s) - \tilde{Z}(s)$. By Itô's formula and (54), we have

$$\begin{aligned} & \mathbb{E}[F(|\tilde{\Delta}(t \wedge T_n \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta)|)] \\ & \leq \mathbb{E}[f(\tilde{X}(t \wedge T_n \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta), \tilde{\Lambda}(t \wedge T_n \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta), \\ & \quad \tilde{Z}(t \wedge T_n \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta), \tilde{\Xi}(t \wedge T_n \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta))] \\ & = F(|x - z|) + \mathbb{E}\left[\int_0^{t \wedge T_n \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta} \tilde{\mathcal{A}}f(\tilde{X}(s), \Delta(s), \tilde{Z}(s), \tilde{\Xi}(s)) ds\right] \\ & \leq F(|x - z|) + 2\kappa_R \mathbb{E}\left[\int_0^{t \wedge T_n \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta} \rho(F(|\tilde{\Delta}(s)|)) ds\right]. \end{aligned}$$

Now passing to the limit as $n \rightarrow \infty$, it follows from the bounded and monotone convergence theorems that

$$\begin{aligned} & \mathbb{E}[F(|\tilde{\Delta}(t \wedge T \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta)|)] \\ & \leq F(|x - z|) + 2\kappa_R \mathbb{E}\left[\int_0^{t \wedge T \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta} \rho(F(|\tilde{\Delta}(s)|)) ds\right] \\ & \leq F(|x - z|) + 2\kappa_R \mathbb{E}\left[\int_0^t \rho(F(|\tilde{\Delta}(s \wedge T \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta)|)) ds\right] \\ & \leq F(|x - z|) + 2\kappa_R \int_0^t \rho(\mathbb{E}[F(|\tilde{\Delta}(s \wedge T \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta)|)]) ds, \end{aligned}$$

where we used the concavity of ρ and Jensen's inequality to obtain the last inequality. Then using Bihari's inequality, we have

$$\mathbb{E}[F(|\tilde{\Delta}(t \wedge T \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta)|)] \leq G^{-1}(G \circ F(|x - z|) + 2\kappa_R t),$$

where the function $G(r) := \int_1^r \frac{ds}{\rho(s)}$ is strictly increasing and satisfies $G(r) \rightarrow -\infty$ as $r \downarrow 0$. In addition, since the function F is strictly increasing, we have

$$\begin{aligned} F(\delta_0) \mathbb{P}\{S_{\delta_0} < t \wedge T \wedge \tau_R \wedge \zeta\} & \leq \mathbb{E}[F(|\tilde{\Delta}(t \wedge T \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta)|) \mathbf{1}_{\{S_{\delta_0} < t \wedge T \wedge \tau_R \wedge \zeta\}}] \\ & \leq \mathbb{E}[F(|\tilde{\Delta}(t \wedge T \wedge S_{\delta_0} \wedge \tau_R \wedge \zeta)|)] \\ & \leq G^{-1}(G \circ F(|x - z|) + 2\kappa_R t). \end{aligned}$$

This implies that

$$\begin{aligned} \mathbb{P}\{S_{\delta_0} < t \wedge T \wedge \tau_R \wedge \zeta\} & \leq \frac{G^{-1}(G \circ F(|x - z|) + 2\kappa_R t)}{F(\delta_0)} \\ & = \frac{1 + \delta_0}{\delta_0} G^{-1}(G \circ F(|x - z|) + 2\kappa_R t). \end{aligned}$$

For any $t \geq 0$ and $\varepsilon > 0$, since $\lim_{R \rightarrow \infty} \tau_R = \infty$ a.s., we can choose $R > 0$ sufficiently large so that

$$\mathbb{P}(t \wedge \zeta > \tau_R) \leq \mathbb{P}(t > \tau_R) < \varepsilon. \tag{56}$$

Then it follows that

$$\begin{aligned}
& \mathbb{E}[F(|\tilde{\Delta}(t \wedge \zeta)|)] \\
&= \mathbb{E}[F(|\tilde{\Delta}(t \wedge \zeta \wedge \tau_R)|)\mathbf{1}_{\{t \wedge \zeta \leq \tau_R\}}] + \mathbb{E}[F(|\tilde{\Delta}(t \wedge \zeta)|)\mathbf{1}_{\{t \wedge \zeta > \tau_R\}}] \\
&\leq \mathbb{E}[F(|\tilde{\Delta}(t \wedge \zeta \wedge T \wedge \tau_R)|)] + \varepsilon \\
&= \mathbb{E}[F(|\tilde{\Delta}(t \wedge \zeta \wedge T \wedge \tau_R)|)\mathbf{1}_{\{S_{\delta_0} < t \wedge T \wedge \tau_R \wedge \zeta\}}] \\
&\quad + \mathbb{E}[F(|\tilde{\Delta}(t \wedge \zeta \wedge T \wedge \tau_R)|)\mathbf{1}_{\{S_{\delta_0} \geq t \wedge T \wedge \tau_R \wedge \zeta\}}] + \varepsilon \\
&\leq \mathbb{P}\{S_{\delta_0} < t \wedge T \wedge \tau_R \wedge \zeta\} + \mathbb{E}[F(|\tilde{\Delta}(t \wedge T \wedge \tau_R \wedge S_{\delta_0} \wedge \zeta)|)] + \varepsilon \\
&\leq \frac{1 + 2\delta_0}{\delta_0} G^{-1}(G \circ F(|x - z|) + 2\kappa_R t) + \varepsilon. \tag{57}
\end{aligned}$$

Passing to the limit, we obtain

$$\lim_{x-z \rightarrow 0} \mathbb{E}[F(|\tilde{\Delta}(t \wedge \zeta)|)] \leq 0 + \varepsilon = \varepsilon. \tag{58}$$

Since $\varepsilon > 0$ is arbitrary, it follows that $\lim_{x-z \rightarrow 0} \mathbb{E}[F(|\tilde{X}(t \wedge \zeta) - \tilde{Z}(t \wedge \zeta)|)] = 0$.

Choose $R > 0$ as in (56). Then we use (51) and (57) to compute

$$\begin{aligned}
\mathbb{P}\{\zeta \leq t\} &= \mathbb{P}\{\zeta \leq t, \tau_R < t\} + \mathbb{P}\{\zeta \leq t, \tau_R \geq t\} \\
&\leq \mathbb{P}\{\tau_R < t\} + \mathbb{E}[\mathbf{1}_{\{\tilde{\Lambda}(t \wedge \zeta \wedge \tau_R) \neq \tilde{\Xi}(t \wedge \zeta \wedge \tau_R)\}}] \\
&< \varepsilon + \mathbb{E}[g(\tilde{X}(t \wedge \zeta \wedge \tau_R), \tilde{\Lambda}(t \wedge \zeta \wedge \tau_R), \tilde{Z}(t \wedge \zeta \wedge \tau_R), \tilde{\Xi}(t \wedge \zeta \wedge \tau_R))] \\
&= \varepsilon + \mathbb{E}\left[\int_0^{t \wedge \zeta \wedge \tau_R} \mathcal{A}g(\tilde{X}(s), \tilde{\Lambda}(s), \tilde{Z}(s), \tilde{\Xi}(s)) ds\right] \\
&\leq \varepsilon + \mathbb{E}\left[\int_0^{t \wedge \zeta \wedge \tau_R} \kappa_R \rho(F(|\tilde{\Delta}(s)|)) ds\right] \\
&\leq \varepsilon + \mathbb{E}\left[\int_0^{t \wedge \tau_R} \kappa_R \rho(F(|\tilde{\Delta}(s \wedge \zeta)|)) ds\right] \\
&\leq \varepsilon + \mathbb{E}\left[\int_0^t \kappa_R \rho(F(|\tilde{\Delta}(s \wedge \zeta)|)) ds\right] \\
&\leq \varepsilon + \kappa_R \int_0^t \rho(\mathbb{E}[F(|\tilde{\Delta}(s \wedge \zeta)|)]) ds \\
&\leq \varepsilon + \kappa_R \int_0^t \rho\left(\frac{1 + 2\delta_0}{\delta_0} G^{-1}(G \circ F(|x - z|) + 2\kappa_R s) + \varepsilon\right) ds \\
&\leq \varepsilon + \kappa_R t \rho\left(\frac{1 + 2\delta_0}{\delta_0} G^{-1}(G \circ F(|x - z|) + 2\kappa_R t) + \varepsilon\right).
\end{aligned}$$

Passing to the limit as $x - z \rightarrow 0$, we obtain

$$\limsup_{x-z \rightarrow 0} \mathbb{P}\{\zeta \leq t\} \leq \varepsilon + \kappa_R t \rho(\varepsilon). \tag{59}$$

Finally, we combine (58) and (59) to obtain

$$\begin{aligned} & \mathbb{E}[f(\tilde{X}(t), \tilde{\Lambda}(t), \tilde{Z}(t), \tilde{\Xi}(t))] \\ &= \mathbb{E}[F(|\tilde{X}(t) - \tilde{Z}(t)|) + \mathbf{1}_{\{\tilde{\Lambda}(t) \neq \tilde{\Xi}(t)\}}] \\ &= \mathbb{E}[F(|\tilde{X}(t) - \tilde{Z}(t)|)\mathbf{1}_{\{\zeta > t\}} + F(|\tilde{X}(t) - \tilde{Z}(t)|)\mathbf{1}_{\{\zeta \leq t\}} + \mathbf{1}_{\{\tilde{\Lambda}(t) \neq \tilde{\Xi}(t)\}}] \\ &\leq \mathbb{E}[F(|\tilde{X}(t \wedge \zeta) - \tilde{Z}(t \wedge \zeta)|)] + 2\mathbb{P}\{\zeta \leq t\} \\ &\rightarrow \varepsilon + 2(\varepsilon + \kappa_R t \rho(\varepsilon)), \text{ as } |x - z| \rightarrow 0. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary and $\lim_{r \downarrow 0} \rho(r) = 0$, it follows that

$$\lim_{x \rightarrow z} \mathbb{E}[f(\tilde{X}(t), \tilde{\Lambda}(t), \tilde{Z}(t), \tilde{\Xi}(t))] = 0.$$

Recall that f is a bounded metric on $\mathbb{R}^d \times \mathbb{S}$. Hence it follows that

$$W_f(P(t, x, k, \cdot), P(t, z, k, \cdot)) \leq \mathbb{E}[f(\tilde{X}(t), \tilde{\Lambda}(t), \tilde{Z}(t), \tilde{\Xi}(t))] \rightarrow 0 \text{ as } x \rightarrow z,$$

where for two probability measures μ and ν on $\mathbb{R}^d \times \mathbb{S}$, the Wasserstein distance $W_f(\mu, \nu)$ is defined as

$$W_f(\mu, \nu) := \inf \left\{ \sum_{i, j \in \mathbb{S}} \int f(x, i, y, j) \pi(\mathrm{d}x, i, \mathrm{d}y, j), \pi \in \mathcal{C}(\mu, \nu) \right\},$$

here $\mathcal{C}(\mu, \nu)$ is the collection of coupling measures for μ and ν . Therefore the desired Feller property follows from Theorem 5.6 of [2]. □

4 Strong Feller Property

Assumption 4.1 For each $k \in \mathbb{S}$ and $x \in \mathbb{R}^d$, the stochastic differential equation (27) has a non-exploding weak solution $X^{(k)}$ with initial condition x and the solution is unique in the sense of probability law.

Assumption 4.2 The process $X^{(k)}$ is strong Feller.

Assumption 4.3 Assume that

$$H := \sup\{q_k(x) : x \in \mathbb{R}^d, k \in \mathbb{S}\} < \infty, \tag{60}$$

and that there exists a positive constant κ such that

$$0 \leq q_{kl}(x) \leq \kappa l 3^{-l} \text{ for all } x \in \mathbb{R}^d \text{ and } k \neq l \in \mathbb{S}. \tag{61}$$

Let us briefly comment on the above assumptions. The existence and uniqueness of weak solution to (27) is related to the study of martingale problem for Lévy

type operators; see, for example, [9] and [21]. Condition (60) in Assumption 4.2 is stronger than (10) in Assumption 2.1. We need such a uniform bound in (60) so that we can establish the series representation for the resolvent of the regime-switching jump diffusion (X, Λ) in Lemma 4.7, which, in turn, helps to establish the strong Feller property for (X, Λ) . In general one can obtain the strong Feller property for $X^{(k)}$ under suitable non-degenerate conditions ([10]) and certain regularity conditions such as (local) Lipschitz conditions of the coefficients. The following non-Lipschitz sufficient condition for strong Feller property was established in [24].

Lemma 4.4 *Suppose that Assumptions 4.1 holds. In addition, for any given $k \in \mathbb{S}$, suppose that for each $R > 0$, there exist positive constants λ_R and κ_R such that for all $x, z \in \mathbb{R}^d$ with $|x| \vee |z| \leq R$, we have*

$$\langle \xi, a(x, k)\xi \rangle \geq \lambda_R |\xi|^2, \quad \forall \xi \in \mathbb{R},$$

and

$$\int_U [|c(x, k, u) - c(z, k, u)|^2 \wedge (4|x - z| \cdot |c(x, k, u) - c(z, k, u)|)] \nu(du) + 2\langle x - z, b(x, k) - b(z, k) \rangle + |\sigma_{\lambda_R}(x, k) - \sigma_{\lambda_R}(z, k)|^2 \leq 2\kappa_R |x - z| \vartheta(|x - z|)$$

whenever $|x - z| \leq \delta_0$, where δ_0 is a positive constant, ϑ is a nonnegative function defined on $[0, \delta_0]$ satisfying $\lim_{r \rightarrow 0} \vartheta(r) = 0$, and $\sigma_{\lambda_R}(x, k)$ is the unique symmetric nonnegative definite matrix-valued function such that $\sigma_{\lambda_R}(x, k)^2 = a(x, k) - \lambda_R I$. Then the process $X^{(k)}$ of (27) is strong Feller continuous.

Next for each $(x, k) \in \mathbb{R}^d \times \mathbb{S}$, as in [19, Section 8.2], we kill the process $X^{(k)}$ at rate $(-q_{kk})$:

$$\begin{aligned} \mathbb{E}_k[f(\tilde{X}_x^{(k)}(t))] &= \mathbb{E}_k \left[f(X_x^{(k)}(t)) \exp \left\{ \int_0^t q_{kk}(X_x^{(k)}(s)) ds \right\} \right] \\ &= \mathbb{E}^{(x,k)}[t < \tau; f(X^{(k)}(t))], \quad f \in \mathfrak{B}_b(\mathbb{R}^d), \end{aligned} \tag{62}$$

to get a subprocess $\tilde{X}^{(k)}$, where $\tau := \inf\{t \geq 0 : \Lambda(t) \neq \Lambda(0)\}$. Equivalently, $\tilde{X}^{(k)}$ can be defined as $\tilde{X}^{(k)}(t) = X^{(k)}(t)$ if $t < \tau$ and $\tilde{X}^{(k)}(t) = \partial$ if $t \geq \tau$, where ∂ is a cemetery point or a coffin state added to \mathbb{R}^d as in [19, p. 145]. Note that in the above, to get the killed process $\tilde{X}^{(k)}$ from the original process $X^{(k)}$, the killing rate is just the jumping rate of Λ from state k . Namely, the killing time is just the first switching time τ . To proceed, we denote the transition probability families of the process $X^{(k)}$ and the killed process $\tilde{X}^{(k)}$ by $\{P^{(k)}(t, x, A) : t \geq 0, x \in \mathbb{R}^d, A \in \mathfrak{B}(\mathbb{R}^d)\}$ and $\{\tilde{P}^{(k)}(t, x, A) : t \geq 0, x \in \mathbb{R}^d, A \in \mathfrak{B}(\mathbb{R}^d)\}$, respectively.

Lemma 4.5 *Under Assumptions 4.1, 4.2, and 4.3, for each $k \in \mathbb{S}$, the killed process $\tilde{X}^{(k)}$ has strong Feller property.*

Proof. Let $\{P_t^{(k)}\}$ and $\{\tilde{P}_t^{(k)}\}$ denote the transition semigroups of $X^{(k)}$ and $\tilde{X}^{(k)}$, respectively. To prove the strong Feller property $\tilde{X}^{(k)}$, we need only prove that for

any given bounded measurable function f on \mathbb{R}^d , $\tilde{P}_t^{(k)} f(z)$ is continuous with respect to z for all $t > 0$. To this end, for fixed $t > 0$ and $0 < s < t$, set $g_s(z) := \tilde{P}_{t-s}^{(k)} f(z)$. Clearly, the function $g_s(\cdot)$ is bounded and measurable, see the Corollary to Theorem 1.1 in [5]. By the strong Feller property of $X^{(k)}$, $P_s^{(k)} g_s \in C_b(\mathbb{R}^d)$.

To proceed, by the Markov property, we have that

$$\begin{aligned} \tilde{P}_t^{(k)} f(x) &= \mathbb{E}_k^{(x)} \left[f(X^{(k)}(t)) \exp \left\{ \int_0^t q_{kk}(X^{(k)}(u)) du \right\} \right] \\ &= \mathbb{E}_k^{(x)} \left[\exp \left\{ \int_0^s q_{kk}(X^{(k)}(u)) du \right\} \right. \\ &\quad \left. \times \mathbb{E}_k^{(X^{(k)}(s))} \left[f(X^{(k)}(t-s)) \exp \left\{ \int_0^{t-s} q_{kk}(X^{(k)}(u)) du \right\} \right] \right]. \end{aligned} \tag{63}$$

Meanwhile, we also have that

$$\begin{aligned} P_s^{(k)} g_s(x) &= P_s^{(k)} \tilde{P}_{t-s}^{(k)} f(x) = \mathbb{E}_k^{(x)} \left[\tilde{P}_{t-s}^{(k)} f(X^{(k)}(s)) \right] \\ &= \mathbb{E}_k^{(x)} \left[\mathbb{E}_k^{(X^{(k)}(s))} \left[f(X^{(k)}(t-s)) \exp \left\{ \int_0^{t-s} q_{kk}(X^{(k)}(u)) du \right\} \right] \right]. \end{aligned} \tag{64}$$

Recall from Assumption 4.2 that $+\infty > H \geq -\inf\{q_{kk}(x) : (x, k) \in \mathbb{R}^{2d} \times \mathbb{S}\}$ and $q_{kk}(x) \leq 0$, and so

$$0 \leq 1 - \exp \left\{ \int_0^s q_{kk}(X^{(k)}(u)) du \right\} \leq (1 - e^{-Hs}). \tag{65}$$

Thus, it follows from (63), (64) and (65) that

$$|P_s^{(k)} g_s(x) - \tilde{P}_t^{(k)} f(x)| \leq (1 - e^{-Hs}) \|f\| \rightarrow 0 \text{ uniformly as } s \rightarrow 0, \tag{66}$$

where $\|\cdot\|$ denotes the uniform (or supremum) norm. Combining this with the fact that $P_s^{(k)} g_s \in C_b(\mathbb{R}^d)$ implies that $\tilde{P}_t^{(k)} f \in C_b(\mathbb{R}^d)$, and so the desired strong Feller property follows. \square

The following lemma was proved in [23].

Lemma 4.6 *Let Ξ be a right continuous strong Markov process and $q : \mathbb{R}^d \mapsto \mathbb{R}$ a nonnegative bounded measurable function. Denote by $\tilde{\Xi}$ the subprocess of Ξ killed at rate q with lifetime ζ :*

$$\mathbb{E}[f(\tilde{\Xi}^{(z)}(t))] := \mathbb{E}[t < \zeta; f(\Xi^{(z)}(t))] = \mathbb{E} \left[f(\Xi^{(z)}(t)) \exp \left\{ - \int_0^t q(\Xi^{(z)}(s)) ds \right\} \right]. \tag{67}$$

Then for any constant $\alpha > 0$ and nonnegative function ϕ on \mathbb{R}^d , we have

$$\mathbb{E}[e^{-\alpha\zeta} \phi(\tilde{\Xi}^{(z)}(\zeta-))] = G_{\alpha}^{\tilde{\Xi}}(q\phi)(z), \tag{68}$$

where $\{\tilde{G}_\alpha^{\tilde{\Xi}}, \alpha > 0\}$ denotes the resolvent for the killed process $\tilde{\Xi}$.

For each $k \in \mathbb{S}$, let $\{\tilde{G}_\alpha^{(k)}, \alpha > 0\}$ be the resolvent for the generator $\mathcal{L}_k + q_{kk}$. Denote by $\{G_\alpha, \alpha > 0\}$ the resolvent for the generator \mathcal{A} defined in (5). Let

$$\tilde{G}_\alpha = \begin{pmatrix} \tilde{G}_\alpha^{(1)} & 0 & 0 & \dots \\ 0 & \tilde{G}_\alpha^{(2)} & 0 & \dots \\ 0 & 0 & \tilde{G}_\alpha^{(3)} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \text{ and } Q^0(x) = Q(x) - \begin{pmatrix} q_{11}(x) & 0 & 0 & \dots \\ 0 & q_{22}(x) & 0 & \dots \\ 0 & 0 & q_{33}(x) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Lemma 4.7 *Suppose that Assumptions 4.1, 4.2, and 4.3 hold. Then there exists a constant $\alpha_1 > 0$ such that for any $\alpha \geq \alpha_1$ and any $f(\cdot, k) \in \mathfrak{B}_b(\mathbb{R}^d)$ with $k \in \mathbb{S}$,*

$$G_\alpha f = \tilde{G}_\alpha f + \sum_{m=1}^\infty \tilde{G}_\alpha (Q^0 \tilde{G}_\alpha)^m f. \tag{69}$$

Proof. Let $f(z, k) \geq 0$ on $\mathbb{R}^{2d} \times \mathbb{S}$. Applying the strong Markov property at the first switching time τ and recalling the construction of (Z, Λ) , we obtain

$$\begin{aligned} G_\alpha f(z, k) &= \mathbb{E}_{z,k} \left[\int_0^\infty e^{-\alpha t} f(Z(t), \Lambda(t)) dt \right] \\ &= \mathbb{E}_{z,k} \left[\int_0^\tau e^{-\alpha t} f(Z(t), k) dt \right] + \mathbb{E}_{z,k} \left[\int_\tau^\infty e^{-\alpha t} f(Z(t), \Lambda(t)) dt \right] \\ &= \tilde{G}_\alpha^{(k)} f(z, k) + \mathbb{E}_{z,k} \left[e^{-\alpha \tau} G_\alpha f(Z(\tau), \Lambda(\tau)) \right] \\ &= \tilde{G}_\alpha^{(k)} f(z, k) + \sum_{l \in \mathbb{S} \setminus \{k\}} \mathbb{E}_{z,k} \left[e^{-\alpha \tau} \left(-\frac{q_{kl}}{q_{kk}} \right) (Z(\tau-)) G_\alpha f(Z(\tau-), l) \right] \\ &= \tilde{G}_\alpha^{(k)} f(z, k) + \sum_{l \in \mathbb{S} \setminus \{k\}} \tilde{G}_\alpha^{(k)} (q_{kl} G_\alpha f(\cdot, l))(z), \end{aligned}$$

where the last equality follows from (68) in Lemma 4.6. Hence we have

$$G_\alpha f(z, k) = \tilde{G}_\alpha^{(k)} f(\cdot, k)(z) + \tilde{G}_\alpha^{(k)} \left(\sum_{l \in \mathbb{S} \setminus \{k\}} q_{kl} G_\alpha f(\cdot, l) \right) (z). \tag{70}$$

Repeating the above argument, the second term on the right-hand side of (70) equals

$$\tilde{G}_\alpha^{(k)} \left(\sum_{l \in \mathbb{S} \setminus \{k\}} q_{kl} \tilde{G}_\alpha^{(l)} f(\cdot, l) \right) (z) + \tilde{G}_\alpha^{(k)} \left(\sum_{l \in \mathbb{S} \setminus \{k\}} q_{kl} \tilde{G}_\alpha^{(l)} \left(\sum_{l_1 \in \mathbb{S} \setminus \{l\}} q_{ll_1} G_\alpha f(\cdot, l_1) \right) \right) (z).$$

Hence, we further obtain that for any fixed $k \in \mathbb{S}$ and any integer $m \geq 1$,

$$G_\alpha f(z, k) = \sum_{i=0}^m \psi_i^{(k)}(z) + R_m^{(k)}(z), \tag{71}$$

where

$$\begin{aligned} \psi_0^{(k)} &= \tilde{G}_\alpha^{(k)} f(\cdot, k), \\ \psi_1^{(k)} &= \tilde{G}_\alpha^{(k)} \left(\sum_{l \in \mathbb{S} \setminus \{k\}} q_{kl} \tilde{G}_\alpha^{(l)} f(\cdot, l) \right) = \tilde{G}_\alpha^{(k)} \left(\sum_{l \in \mathbb{S} \setminus \{k\}} q_{kl} \psi_0^{(l)} \right), \\ \psi_i^{(k)} &= \tilde{G}_\alpha^{(k)} \left(\sum_{l \in \mathbb{S} \setminus \{k\}} q_{kl} \psi_{i-1}^{(l)} \right) \quad \text{for } i \geq 1, \end{aligned}$$

and

$$R_m^{(k)} = \tilde{G}_\alpha^{(k)} \left(\sum_{l_1 \in \mathbb{S} \setminus \{k\}} q_{k,l_1} \tilde{G}_\alpha^{(l_1)} \left(\sum_{l_2 \in \mathbb{S} \setminus \{l_1\}} q_{l_1,l_2} \tilde{G}_\alpha^{(l_2)} \left(\dots \left(\sum_{l_{m-1} \in \mathbb{S} \setminus \{l_{m-2}\}} q_{l_{m-2},l_{m-1}} \tilde{G}_\alpha^{(l_{m-1})} \left(\sum_{l_m \in \mathbb{S} \setminus \{l_{m-1}\}} q_{l_{m-1},l_m} G_\alpha f(\cdot, l_m) \right) \right) \right) \right) \right).$$

We have

$$\|\psi_0^{(k)}\| = \left\| \mathbb{E}_{\cdot, k} \left[\int_0^\tau e^{-\alpha t} f(Z(t), k) dt \right] \right\| \leq \frac{\|f\|}{\alpha}. \tag{72}$$

Note that the same calculation reveals that (72) in fact holds for all $l \in \mathbb{S}$, $\|\psi_0^{(l)}\| \leq \frac{\|f\|}{\alpha}$. Thanks to Assumption 4.3, $q_{kl}(z) \leq \frac{\kappa l}{3^l}$ for all $l \neq k$ and $x \in \mathbb{R}^d$. Consequently, we can compute

$$\begin{aligned} \|\psi_1^{(k)}\| &\leq \sum_{l \in \mathbb{S} \setminus \{k\}} \|\tilde{G}_\alpha^{(k)}(q_{kl} \psi_0^{(l)})\| \leq \sum_{l \in \mathbb{S} \setminus \{k\}} \frac{\kappa l}{3^l} \cdot \frac{\|\psi_0^{(l)}\|}{\alpha} \\ &\leq \sum_{l \in \mathbb{S} \setminus \{k\}} \frac{\kappa l}{3^l} \cdot \frac{\|f\|}{\alpha^2} = \frac{3\kappa}{4\alpha} \cdot \frac{\|f\|}{\alpha}. \end{aligned} \tag{73}$$

As before, we observe that (73) actually holds for all $l \in \mathbb{S}$. Similarly, we can use induction to show that

$$\|\psi_i^{(k)}\| \leq \left(\frac{3\kappa}{4\alpha} \right)^i \cdot \frac{\|f\|}{\alpha}, \quad \text{for } i \geq 2, \tag{74}$$

and

$$\|R_m^{(k)}\| \leq \left(\frac{3\kappa}{4\alpha} \right)^{m+1} \cdot \frac{\|f\|}{\alpha}. \tag{75}$$

Now let $\alpha_1 := \frac{3\kappa+1}{4}$ and $\alpha \geq \alpha_1$. Then we have for each $k \in \mathbb{S}$, $G_\alpha f(\cdot, k) = \sum_{i=0}^\infty \psi_i^{(k)}$, which clearly implies (69). The lemma is proved. \square

Lemma 4.7 establishes an explicit relationship of the resolvents for (Z, Λ) and the killed processes $\tilde{Z}^{(k)}$, $k \in \mathbb{S}$. This, together with the strong Feller property for the killed processes $\tilde{Z}^{(k)}$, $k \in \mathbb{S}$ (Lemma 4.5), enables us to derive the strong Feller property for (Z, Λ) in the following theorem.

Theorem 4.8 *Suppose that Assumptions 3.1, 4.1, 4.2, and 4.3 hold. Then the process (X, Λ) has the strong Feller property.*

Proof. The proof is almost identical to that of Theorem 5.4 in [23] and for brevity, we shall only give a sketch here. Denote the transition probability family of Markov process (X, Λ) by $\{P(t, (x, k), A) : t \geq 0, (x, k) \in \mathbb{R}^d \times \mathbb{S}, A \in \mathcal{B}(\mathbb{R}^d \times \mathbb{S})\}$. Then it follows from Lemma 4.7 that

$$\begin{aligned}
 P(t, (x, k), A \times \{l\}) &= \delta_{kl} \tilde{P}^{(k)}(t, x, A) \\
 &+ \sum_{m=1}^{+\infty} \int_{0 < t_1 < \dots < t_m < t} \int_{l_1 \in \mathbb{S} \setminus \{l_0\}, l_2 \in \mathbb{S} \setminus \{l_1\}, \dots, l_m \in \mathbb{S} \setminus \{l_{m-1}\}, l_0 = k, l_m = l} \int_{\mathbb{R}^d} \dots \int_{\mathbb{R}^d} \tilde{P}^{(l_0)}(t_1, x, dx_1) \\
 &\times q_{l_0 l_1}(x_1) \tilde{P}^{(l_1)}(t_2 - t_1, x_1, dx_2) \dots q_{l_{m-1} l_m}(x_m) \tilde{P}^{(l_m)}(t - t_m, x_m, A) dt_1 dt_2 \dots dt_m,
 \end{aligned}
 \tag{76}$$

where δ_{kl} is the Kronecker symbol in k, l , which equals 1 if $k = l$ and 0 if $k \neq l$. By Lemma 4.5, we know that for every $k \in \mathbb{S}$, $\tilde{X}^{(k)}$ has the strong Feller property. Therefore, in view of Proposition 6.1.1 in [15], we derive that $\tilde{P}^{(k)}(t, x, A)$ and every term in the series on the right-hand side of (76) are lower semicontinuous with respect to x whenever A is an open set in $\mathfrak{B}(\mathbb{R}^d)$. Note that \mathbb{S} is a countably infinite set and has discrete metric. Therefore it follows that the left-hand side of (76) is lower semicontinuous with respect to (x, k) for every $l \in \mathbb{S}$ whenever A is an open set in $\mathfrak{B}(\mathbb{R}^d)$. Consequently, (X, Λ) has the strong Feller property (see Proposition 6.1.1 in [15] again). The theorem is proved. \square

Remark 4.9 [20] proves that for a state-independent regime-switching diffusion processes, the strong Feller property for each subdiffusion implies the strong Feller property for regime-switching diffusion processes. This work further proves this implication for state-dependent regime-switching jump diffusion processes.

Remark 4.10 The strong Feller property for regime-switching jump diffusions was also studied in [22], where it is assumed that $\nu(U) < \infty$ is a finite measure, i.e., the jump part is modeled by a compound Poisson process. In addition, a finite-range condition for the switching component is placed in that paper and is key to the analyses there. Here these two restrictions are removed.

References

1. Richard F. Bass. Stochastic differential equations driven by symmetric stable processes. In *Séminaire de Probabilités, XXXVI*, volume 1801 of *Lecture Notes in Math.*, pages 302–313. Springer, Berlin, 2003.
2. Mu-Fa Chen. *From Markov chains to non-equilibrium particle systems*. World Scientific Publishing Co. Inc., River Edge, NJ, second edition, 2004.
3. Xiaoshan Chen, Zhen-Qing Chen, Ky Tran, and George Yin. Properties of switching jump diffusions: Maximum principles and harnack inequalities. *Bernoulli*, to appear, 2018.
4. Xiaoshan Chen, Zhen-Qing Chen, Ky Tran, and George Yin. Recurrence and ergodicity for a class of regime-switching jump diffusions. *Applied Mathematics & Optimization*, 2018. <https://doi.org/10.1007/s00245-017-9470-9>
5. Kai Lai Chung and Zhong Xin Zhao. *From Brownian motion to Schrödinger's equation*, volume 312 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1995.
6. Shizan Fang and Tusheng Zhang. A study of a class of stochastic differential equations with non-Lipschitzian coefficients. *Probab. Theory Related Fields*, 132(3):356–390, 2005.
7. Zongfei Fu and Zenghu Li. Stochastic equations of non-negative processes with jumps. *Stochastic Process. Appl.*, 120(3):306–330, 2010.
8. Fima C. Klebaner. *Introduction to stochastic calculus with applications*. Imperial College Press, London, second edition, 2005.
9. Takashi Komatsu. Markov processes associated with certain integro-differential operators. *Osaka J. Math.*, 10:271–303, 1973.
10. Hiroshi Kunita. Nondegenerate SDE's with jumps and their hypoelliptic properties. *J. Math. Soc. Japan*, 65(3):993–1035, 2013.
11. Sean D. Lawley, Jonathan C. Mattingly, and Michael C. Reed. Sensitivity to switching rates in stochastically switched ODEs. *Commun. Math. Sci.*, 12(7):1343–1352, 2014.
12. Zenghu Li and Leonid Mytnik. Strong solutions for stochastic differential equations with jumps. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(4):1055–1067, 2011.
13. Zenghu Li and Fei Pu. Strong solutions of jump-type stochastic equations. *Electron. Commun. Probab.*, 17(33):1–13, 2012.
14. Xuerong Mao and Chenggui Yuan. *Stochastic differential equations with Markovian switching*. Imperial College Press, London, 2006.
15. S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993.
16. Dang H. Nguyen and George Yin. Recurrence and ergodicity of switching diffusions with past-dependent switching having a countable state space. *Potential Anal.*, 48(4):405–435, 2018.
17. Dang Hai Nguyen and George Yin. Modeling and analysis of switching diffusion systems: past-dependent switching with a countable state space. *SIAM J. Control Optim.*, 54(5):2450–2477, 2016.
18. J. R. Norris. *Markov chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. Reprint of 1997 original.
19. Bernt Øksendal. *Stochastic differential equations, An introduction with applications*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003.
20. Jinghai Shao. Strong solutions and strong Feller properties for regime-switching diffusion processes in an infinite state space. *SIAM J. Control Optim.*, 53(4):2462–2479, 2015.
21. Daniel W. Stroock. Diffusion processes associated with Lévy generators. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 32(3):209–244, 1975.
22. Fubao Xi and Chao Zhu. On Feller and strong Feller properties and exponential ergodicity of regime-switching jump diffusion processes with countable regimes. *SIAM J. Control Optim.*, 55(3):1789–1818, 2017.
23. Fubao Xi and Chao Zhu. On the martingale problem and Feller and strong Feller properties for weakly coupled Lévy type operators. *Stochastic Process. Appl.*, 12(12):4277–4308, 2018.

24. Fubao Xi and Chao Zhu. Jump type stochastic differential equations with non-lipschitz coefficients: Non confluence, feller and strong feller properties, and exponential ergodicity. *J. Differential Equations*, 266(8):4668–4711, 2019.
25. Toshio Yamada and Shinzo Watanabe. On the uniqueness of solutions of stochastic differential equations. *J. Math. Kyoto Univ.*, 11:155–167, 1971.
26. G. Yin, Guangliang Zhao, and Fuke Wu. Regularization and stabilization of randomly switching dynamic systems. *SIAM J. Appl. Math.*, 72(5):1361–1382, 2012.
27. George Yin and Chao Zhu. *Hybrid Switching Diffusions: Properties and Applications*, volume 63 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2010.