Risk-Sensitive Markov Decision Problems under Model Uncertainty: Finite Time Horizon Case

Tomasz R. Bielecki, Tao Chen, and Igor Cialenco

Abstract In this paper we study a class of risk-sensitive Markovian control problems in discrete time subject to model uncertainty. We consider a risk-sensitive discounted cost criterion with finite time horizon. The used methodology is the one of adaptive robust control combined with machine learning.

1 Introduction

The main goal of this work is to study finite time horizon *risk-sensitive Markovian control problems* subject to model uncertainty in a discrete time setup, and to develop a methodology to solve such problems efficiently. The proposed approach hinges on the following main building concepts: incorporating model uncertainty through the *adaptive robust* paradigm introduced in [BCC$^+$19] and developing efficient numerical solutions for the obtained Bellman equations by adopting the *machine learning techniques* proposed in [CL19].

There exists a significant body of work on incorporating model uncertainty (or model misspecification) in stochastic control problems, and among some of the well-known and prominent methods we would mention the robust control approach [GS89, HSTW06, HS08], adaptive control [KV15, CG91], and Bayesian adaptive control [KV15]. A comprehensive literature review on this subject is beyond the scope of this paper, and we refer the reader to [BCC$^+$19] and references therein. In [BCC$^+$19] the authors proposed a novel adaptive robust methodology that solves

Tomasz R. Bielecki,
Department of Applied Mathematics, Illinois Institute of Technology,
e-mail: tbielecki@iit.edu Tao Chen,
Department of Mathematics, University of Michigan,
e-mail: chenta@umich.edu and Igor Cialenco
Department of Applied Mathematics, Illinois Institute of Technology,
e-mail: cialenco@iit.edu

G. Yin, T. Zariphopoulou (eds.), *Stochastic Analysis, Filtering, and Stochastic Optimization*,
https://doi.org/10.1007/978-3-030-98519-6_2

time-consistent Markovian control problems in discrete time subject to model uncertainty - the approach that we take in this study too. The core of this methodology was to combine a recursive learning mechanism about the unknown model with the underlying Markovian dynamics, and to demonstrate that the so called adaptive robust Bellman equations produce an optimal adaptive robust control strategy.

In contrast to [BCC$^+$19], where the considered optimization criterion was of the terminal reward type, in the present work, we also allow intermediate rewards and we use the discounted risk sensitive criterion. Accordingly, we derive a new set of adaptive robust Bellman equations, similar to those used in [BCC$^+$19].

Risk sensitive criterion has been broadly used both in the control oriented literature, as well as in the game oriented literature. We refer to, e.g., [BP03, DL14, BR17], and the references therein for insight into risk sensitive control and risk sensitive games both in discrete time and in continuous time.

The paper is organized as follows. In Section 2 we formulate the finite time horizon risk-sensitive Markovian control problem subject to model uncertainty that is studied here. Section 3 is devoted to the formulation and to study of the robust adaptive control problem that is relevant for the problem formulated in Section 2. This section presents the main theoretical developments of the present work. In Section 4 we formulate an illustrative example of our theoretical results that is rooted in the classical linear-quadratic-exponential control problem (see e.g. [HS95]). Next, using machine learning methods, in Section 5 we provide numerical solutions of the example presented in Section 4.

Finally, we want to mention that the important case of an infinite time horizon risk-sensitive Markovian control problem in discrete time subject to model uncertainty will be studies in a follow-up work.

2 Risk-sensitive Markovian discounted control problems with model uncertainty

In this section we state the underlying discounted risk-sensitive stochastic control problems. Let $(\Omega, \mathscr{F})$ be a measurable space, $T \in \mathbb{N}$ be a finite time horizon, and let us denote by $\mathscr{T} := \{0,1,2,\dots,T\}$ and $\mathscr{T}' := \{0,1,2,\dots,T-1\}$. We let $\Theta \subset \mathbb{R}^d$ be a non-empty compact set, which will play the role of the parameter space throughout. We consider a random process $Z = \{Z_t,\ t = 1,2\dots\}$ on $(\Omega, \mathscr{F})$ taking values in $\mathbb{R}^m$, and we denote by $\mathbb{F} = (\mathscr{F}_t,\ t = 0,2\dots)$ its natural filtration, with $\mathscr{F}_0 = \{\emptyset, \Omega\}$. We postulate that this process is observed by the controller, but the true law of Z is unknown to the controller and assumed to be generated by a probability measure belonging to a (known) parameterized family of probability distributions on $(\Omega, \mathscr{F})$, say $\mathbf{P}(\Theta) = \{\mathbb{P}_\theta, \theta \in \Theta\}$. As usually, $\mathbb{E}_\mathbb{P}$ will denote the expectation under a probability measure $\mathbb{P}$ on $(\Omega, \mathscr{F})$, and, for simplicity, we will write $\mathbb{E}_\theta$ instead of $\mathbb{E}_{\mathbb{P}_\theta}$. We denote by $\mathbb{P}_{\theta^*}$ the measure generating the true law of Z, and thus $\theta^* \in \Theta$ is the unknown true parameter. The sets Θ and $\mathbf{P}(\Theta)$ are known to

the observer. Clearly, the model uncertainty may occur if $\Theta \neq \{\theta^*\}$, which we will assume to hold throughout.

We let $A \subset \mathbb{R}^k$ be a finite set,[1] and $S : \mathbb{R}^n \times A \times \mathbb{R}^m \to \mathbb{R}^d$ be a measurable mapping. An admissible control process φ is an $\mathbb{F}$-adapted process, taking values in A, and we will denote by $\mathscr{A}$ the set of all admissible control processes.

We consider an underlying discrete time controlled dynamical system with the state process X taking values in $\mathbb{R}^n$ and control process φ taking values in A. Specifically, we let

$$X_{t+1} = S(X_t, \varphi_t, Z_{t+1}), \quad t \in \mathscr{T}', \quad X_0 = x_0 \in \mathbb{R}^n. \tag{1}$$

At each time $t = 0, \ldots, T-1$, the running reward $r_t(X_t, \varphi_t)$ is delivered, where, for every $a \in A$, the function $r_t(\cdot, a) : \mathbb{R}^n \to \mathbb{R}_+$ is bounded and continuous. Similarly, at the terminal time $t = T$ the terminal reward $r_T(X_T)$ is delivered, where $r_T : \mathbb{R}^n \to \mathbb{R}_+$ is a bounded and continuous function.

Let $\beta \in (0,1)$ be a discount factor, and let $\gamma \neq 0$ be the risk sensitivity factor. The underlying discounted, risk-sensitive control problem is:

$$\sup_{\varphi \in \mathscr{A}} \frac{1}{\gamma} \ln \left(\mathbb{E}_{\theta^*} e^{\gamma \left(\sum_{t=0}^{T-1} \beta^t r_t(X_t, \varphi_t) + \beta^T r_T(X_T) \right)} \right) \tag{2}$$

subject to (1). Clearly, since θ^* is not known to the controller, the above problem can not be solved as it is stated. The main goal of this paper is formulate and solve the adaptive robust control problem corresponding to (2).

Remark 1 (i) The risk-sensitive criterion in (2) is in fact an example of application of the entropic risk measure, say $\rho_{\theta^*,\gamma}$, which is defined as

$$\rho_{\theta^*,\gamma}(\xi) := \frac{1}{\gamma} \ln \mathbb{E}_{\theta^*} e^{\gamma \xi},$$

where ξ is a random variable on $(\Omega, \mathscr{F}, P_{\theta^*})$ that admits finite moments of all orders.
(ii) It can be verified that

$$\rho_{\theta^*,\gamma}(\xi) = \mathbb{E}_{\theta^*}(\xi) + \frac{\gamma}{2} \mathbb{VAR}_{\theta^*}(\xi) + O(\gamma^2).$$

Thus, in case when $\gamma < 0$ the term $\frac{\gamma}{2} \mathbb{VAR}_{\theta^*}(\xi)$ can be interpreted as the risk-penalizing term. On the contrary, when $\gamma > 0$, the term $\frac{\gamma}{2} \mathbb{VAR}_{\theta^*}(\xi)$ can be viewed as the risk-favoring term.
(iii) In the rest of the paper we focus on the case $\gamma > 0$. The case $\gamma < 0$ can be treated in an analogous way.

[1] A will represent the set of control values, and we assume it is finite for simplicity, in order to avoid technical issues regarding the existence of measurable selectors.

3 The adaptive robust risk sensitive discounted control problem

We follow here the developments presented in [BCC+19]. The key difference is that in this work we deal with running and terminal costs.

In what follows, we will be making use of a recursive construction of confidence regions for the unknown parameter θ^* in our model. We refer to [BCC17] for a general study of recursive constructions of (approximate) confidence regions for time homogeneous Markov chains. Section 4 provides details of a specific such recursive construction corresponding to the example presented in that section. Here, we just postulate that the recursive algorithm for building confidence regions uses a Θ-valued and observed process, say $C = (C_t,\ t \in \mathbb{N}_0)$, satisfying the following abstract dynamics

$$C_{t+1} = R(t, C_t, Z_{t+1}), \quad t \in \mathbb{N}_0,\ C_0 = c_0 \in \Theta, \tag{3}$$

where $R : \mathbb{N}_0 \times \mathbb{R}^d \times \mathbb{R}^m \to \Theta$ is a deterministic measurable function. Note that, given our assumptions about process Z, the process C is $\mathbb{F}$-adapted. This is one of the key features of our model. Usually C_t is taken to be a consistent estimator of θ^*.

Now, we fix a confidence level $\alpha \in (0,1)$, and for each time $t \in \mathbb{N}_0$, we assume that an $(1-\alpha)$-confidence region, say $\Theta_t \subset \mathbb{R}^d$, for θ^*, can be represented as

$$\Theta_t = \tau(t, C_t), \tag{4}$$

where, for each $t \in \mathbb{N}_0$, $\tau(t,\cdot) : \mathbb{R}^d \to 2^\Theta$ is a deterministic set valued function, where, as usual, 2^Θ denotes the set of all subsets of Θ. Note that in view of (3) the construction of confidence regions given in (4) is indeed recursive. In our construction of confidence regions, the mapping $\tau(t,\cdot)$ will be a measurable set valued function, with compact values. It needs to be noted that we will only need to compute Θ_t until time $T-1$. In addition, we assume that for any $t \in \mathscr{T}'$, the mapping $\tau(t,\cdot)$ is upper hemi-continuous (u.h.c.). That is, for any $c \in \Theta$, and any open set E such that $\tau(t,c) \subset E \subset \Theta$, there exists a neighbourhood D of c such that for all $c' \in D$, $\tau(t,c') \subset E$ (cf. [Bor85, Definition 11.3]).

Remark 2 The important property of the recursive confidence regions constructed as indicated above is that, in many models, $\lim_{t\to\infty} \Theta_t = \{\theta^*\}$, where the convergence is understood $\mathbb{P}_{\theta^*}$ almost surely, and the limit is in the Hausdorff metric. This is not always the case though in general. In [BCC17] is shown that the convergence holds in probability, for the model setup studied there.

The sequence Θ_t, $t \in \mathscr{T}'$ represents learning about θ^* based on the observation of the history $(Y_0, Y_1 \ldots, Y_t)$, $t \in \mathscr{T}'$, where $Y_t = (X_t, C_t)$, $t \in \mathscr{T}$, is the augmented state process taking values in the augmented state space

$$E_Y = \mathbb{R}^n \times \Theta.$$

We denote by $\mathscr{E}_Y$ the collection of Borel measurable sets in E_Y.

In view of the above, if the control process φ is employed then the process Y has the following dynamics

$$Y_{t+1} = \mathbf{G}(t, Y_t, \varphi_t, Z_{t+1}), \quad t \in \mathscr{T}',$$

where the mapping $\mathbf{G} : \mathbb{N}_0 \times E_Y \times A \times \mathbb{R}^m \to E_Y$ is defined as

$$\mathbf{G}(t, y, a, z) = \big(S(x, a, z), R(t, c, z)\big), \tag{5}$$

with $y = (x, c) \in E_Y$.

We define the corresponding histories

$$H_t = (Y_0, \ldots, Y_t), \quad t \in \mathscr{T}', \tag{6}$$

so that

$$H_t \in \mathbf{H}_t = \underbrace{E_Y \times E_Y \times \ldots \times E_Y}_{t+1 \text{ times}}. \tag{7}$$

Clearly, for any admissible control process φ, the random variable H_t is $\mathscr{F}_t$-measurable. We denote by

$$h_t = (y_0, y_1, \ldots, y_t) = (x_0, c_0, x_1, c_1, \ldots, x_t, c_t) \tag{8}$$

a realization of H_t. Note that $h_0 = y_0 = (x_0, c_0)$.

A control process $\varphi = (\varphi_t,\ t \in \mathscr{T}')$ is called history dependent control process if (with a slight abuse of notation)

$$\varphi_t = \varphi_t(H_t),$$

where (on the right hand side) $\varphi_t : \mathbf{H}_t \to A$, is a measurable mapping. Given our above setup, any history dependent control process is $\mathbb{F}$–adapted, and thus, it is admissible. For any admissible control process φ and for any $t \in \mathscr{T}'$, we denote by $\varphi^t = (\varphi_k,\ k = t, \ldots, T-1)$ the 't-tail' of φ. Accordingly, we denote by $\mathscr{A}^t$ the collection of 't-tails' of φ. In particular, $\varphi^0 = \varphi$ and $\mathscr{A}^0 = \mathscr{A}$. The superscript notation applied to processes should not be confused with power function applied such as β^t.

Let $\psi_t : \mathbf{H}_t \to \Theta$ be a Borel measurable mapping such that $\psi_t(h_t) \in \tau(t, c_t)$, and let us denote by $\psi = (\psi_t,\ t \in \mathscr{T}')$ the sequence of such mappings, and by ψ^t the t-tails of the sequence ψ, in analogy to φ^t. The set of all sequences ψ, and respectively ψ^t , will be denoted by Ψ and Ψ^t, respectively.

Strategies φ and ψ are called *Markovian strategies or policies* if (with some abuse of notation)

$$\varphi_t = \varphi_t(Y_t), \quad \psi_t = \psi_t(Y_t),$$

where (on the right hand side) $\varphi_t : E_Y \to A$, and is a (Borel) measurable mapping, and $\psi_t : E_Y \to \Theta$ is a (Borel) measurable mapping satisfying $\psi_t(x, c) \in \tau(t, c)$.

In order to simplify all the following argument we limit ourselves to Markovian policies. In case of Markovian dynamics settings, such as ours, this comes without

loss of generality, as there typically exist optimal Markovian strategies, if optimal strategies exist at all. Accordingly, $\mathscr{A}$ and Ψ are now sets of Markov strategies.

Next, for each $(t,y,a,\theta) \in \mathscr{T}' \times E_Y \times A \times \Theta$, we define a probability measure on $\mathscr{E}_Y$, given by

$$Q(B \mid t,y,a,\theta) = \mathbb{P}_\theta(Z_{t+1} \in \{z : \mathbf{G}(t,y,a,z) \in B\}) = \mathbb{P}_\theta\left(\mathbf{G}(t,y,a,Z_{t+1}) \in B\right), \quad (9)$$

for any $B \in \mathscr{E}_Y$ We assume that for every $t \in \mathscr{T}$ and every $a \in A$, we have that $Q(dy' \mid t,y,a,\theta)$ is a Borel measurable stochastic kernel with respect to (y,θ). This assumption will be strengthened later on.

Using Ionescu-Tulcea theorem (cf. [BR11, Appendix B]), for every $t = 0,\dots,T-1$, every t-tail $\varphi^t \in \mathscr{A}^t$ and every state $y_t \in E_Y$, we define the family $\mathscr{Q}_{y_t,t}^{\varphi^t,\Psi^t} = \{\mathbb{Q}_{y_t,t}^{\varphi^t,\psi^t},\ \psi^t \in \Psi^t\}$ of probability measures on the concatenated canonical space $\mathsf{X}_{s=t+1}^{T} E_Y$, with

$$\begin{aligned}&\mathbb{Q}_{y_t,t}^{\varphi^t,\psi^t}(B_{t+1} \times \cdots \times B_T)\\ &\quad := \int_{B_{t+1}} \cdots \int_{B_T} \prod_{u=t+1}^{T} Q(\mathrm{d}y_u \mid u-1, y_{u-1}, \varphi_{u-1}(y_{u-1}), \psi_{u-1}(y_{u-1})). \end{aligned} \quad (10)$$

The *discounted, risk-sensitive, adaptive robust control problem* corresponding[2] to (2) is:

$$\sup_{\varphi^0 \in \mathscr{A}^0} \inf_{\mathbb{Q} \in \mathscr{Q}_{y_0,0}^{\varphi^0,\Psi^0}} \mathbb{E}_{\mathbb{Q}} e^{\gamma \Sigma_{t=0}^{T} \beta^t r_t(X_t, \varphi_t(Y_t))}, \quad (11)$$

where, for simplicity of writing, here and everywhere below, with slight abuse of notations, we set $r_T(x,a) = r_T(x)$. In next section we will show that a solution to this problem can be given in terms of the discounted adaptive robust Bellman equations associated to it.

3.1 Adaptive robust Bellman equation

Towards this end we aim our attention at the following adaptive robust Bellman equations

$$\begin{aligned} W_T(y) &= e^{\gamma \beta^T r_T(x)}, \quad y \in E_Y,\\ W_t(y) &= \max_{a \in A} \inf_{\theta \in \tau(t,c)} \int_{E_Y} W_{t+1}(y') e^{\gamma \beta^t r_t(x,a)} Q(\mathrm{d}y' \mid t,y,a,\theta), \\ &\qquad y \in E_Y,\ t = T-1,\dots,0, \end{aligned} \quad (12)$$

where we recall that $y = (x,c)$.

[2] Since $\gamma > 0$, we omit the factor $1/\gamma$.

Remark 3 Clearly, in (12), the exponent $e^{\gamma\beta^t r_t(x,a)}$ can be factored out, and W_t can be written as

$$W_t(y) = \max_{a\in A}\left(e^{\gamma\beta^t r_t(x,a)}\cdot \inf_{\theta\in\tau(t,c)}\int_{E_Y} W_{t+1}(y')Q(\mathrm{d}y' \mid t,y,a,\theta)\right).$$

Nevertheless, in what follows, we will keep similar factors inside of the integrals, mostly for the convenience of writing as well as to match the visual appearance of classical Bellman equations.

We will study the solvability of this system. We start with Lemma 1 below, where, under some additional technical assumptions, we show that the optimal selectors in (12) exist; namely, for any $t\in\mathscr{T}'$, and any $y=(x,c)\in E_Y$, there exists a measurable mapping $\varphi_t^* : E_Y\to A$, such that

$$W_t(y) = \inf_{\theta\in\tau(t,c)}\int_{E_Y} W_{t+1}(y')e^{\gamma\beta^t r_t(x,\varphi_t^*(y))}Q(\mathrm{d}y' \mid t,y,\varphi_t^*(y),\theta).$$

In order to proceed, for the sake of simplicity, we will assume that under measure $\mathbb{P}_\theta$, for each $t\in\mathscr{T}$, the random variable Z_t has a density with respect to the Lebesgue measure, say $f_Z(z;\theta)$, $z\in\mathbb{R}^m$. In this case we have

$$\int_{E_Y} W_{t+1}(y')Q(\mathrm{d}y' \mid t,y,a,\theta) = \int_{\mathbb{R}^m} W_{t+1}(\mathbf{G}(t,y,a,z))f_Z(z;\theta)\,\mathrm{d}z,$$

where $\mathbf{G}(t,y,a,z)$ is given in (5).

Additionally, we take the standing assumptions:

(i) for any a and z, the function $S(\cdot,a,z)$ is continuous;
(ii) for each z, the function $f_Z(z;\cdot)$ is continuous;
(iii) for each $t\in\mathscr{T}'$, the function $R(t,\cdot,\cdot)$ is continuous.

Then, the following result holds true.

Lemma 1 *The functions W_t, $t=T,T-1,\ldots,0$, are lower semi-continuous (l.s.c.), and the optimal selectors φ_t^*, $t=T-1,\ldots,0$, realizing maxima in* (12) *exist.*

Proof Since r_T is continuous and bounded, so is the function W_T. Since $\mathbf{G}(T-1,\cdot,a,z)$ is continuous, then, $W_T(\mathbf{G}(T-1,\cdot,a,z))$ is continuous. Consequently, recalling again that $y=(x,c)$, for each a, the function

$$w_{T-1}(y,a,\theta) := \int_{\mathbb{R}} W_T(\mathbf{G}(T-1,y,a,z))e^{\gamma\beta^{T-1}r_{T-1}(x,a)}f_Z(z;\theta)\,\mathrm{d}z$$

$$= e^{\gamma\beta^{T-1}r_{T-1}(x,a)}\int_{\mathbb{R}} e^{\gamma\beta^T r_T(S(x,a,z))}f_Z(z;\theta)\,\mathrm{d}z$$

is continuous in (y,θ).

Next, we will apply [BS78, Proposition 7.33] by taking (in the notations of [BS78])

$$\begin{aligned}
&\mathrm{X} = E_Y \times A = \mathbb{R}^n \times \Theta \times A, \quad \mathrm{x} = (y,a),\\
&\mathrm{Y} = \Theta, \quad \mathrm{y} = \theta,\\
&\mathrm{D} = \bigcup_{(y,a)\in E_Y\times A} \{(y,a)\} \times \tau(T-1,c),\\
&f(\mathrm{x},\mathrm{y}) = w_{T-1}(y,a,\theta).
\end{aligned}$$

Note that in view of the prior assumptions, Y is metrizable and compact. Clearly X is metrizable. From the above, f is continuous, and thus lower semi-continuous. Since $\tau(T-1,\cdot)$ is compact-valued and u.h.c. on $E_Y \times A$, then according to [Bor85, Proposition 11.9], the set-valued function $\tau(T-1,\cdot)$ is closed, which implies that its graph D is closed [Bor85, Definition 11.5]. Also note that the cross section $\mathrm{D}_{\mathrm{x}} = \mathrm{D}_{(y,a)} = \{\theta \in \Theta : (y,a,\theta) \in \mathrm{D}\}$ is given by $\mathrm{D}_{(y,a)}(T-1) = \tau(T-1,c)$. Hence, by [BS78, Proposition 7.33], the function

$$\widetilde{w}_{T-1}(y,a) = \inf_{\theta\in\tau(T-1,c)} (w_{T-1}(y,a,\theta)), \quad (y,a) \in E_Y \times A,$$

is l.s.c. Consequently, the function $\widehat{w}_{T-1}(y,a) = -\widetilde{w}_{T-1}(y,a)$ is u.s.c. (upper semi-continuous). Thus, by [BS78, Proposition 7.34], the function

$$-W_{T-1}(y) = -\max_{a\in A} \widetilde{w}_{T-1}(y,a) = \min_{a\in A} \widehat{w}_{T-1}(y,a)$$

is u.s.c., so that $W_{T-1}(y)$ is l.s.c. Moreover, since A is finite, there exists an optimal selector φ^*_{T-1}, that is $W_{T-1}(y) = \widetilde{w}_{T-1}(y,\varphi^*_{T-1}(y))$.

Proceeding to the next step, note that $W_{T-1}(\mathbf{G}(T-2,y,a,z))e^{\gamma\beta^{T-2}r_{T_2}(x,a)}$ is l.s.c. and positive, hence bounded from below. Therefore, according to [BS78, Proposition 7.31], the function

$$w_{T-2}(y,a,\theta) = \int_{\mathbb{R}} W_{T-1}(\mathbf{G}(T-2,y,a,z))e^{\gamma\beta^{T-2}r_{T-2}(x,a)} f_Z(z;\theta)\,\mathrm{d}z$$

is l.s.c.. The rest of the proof follows in the analogous way. □

Next, we will prove an auxiliary result needed to justify the mathematical operations conducted in the proof of the main result – Theorem 1. Define the functions U_t and U^*_t as follows: for $\varphi^t \in \mathscr{A}^t$ and $y \in E_Y$,

$$U_t(\varphi^t,y) = e^{\gamma\beta^t r_t(x,\varphi_t(y))} \inf_{\mathbb{Q}\in\mathscr{Q}^{\varphi^t,\Psi^t}_{y,t}} \mathbb{E}_{\mathbb{Q}} e^{\gamma\sum_{k=t+1}^T \beta^k r_k(X_k,\varphi_k(Y_k))}, \quad t \in \mathscr{T}', \tag{13}$$

$$U^*_t(y) = \sup_{\varphi^t\in\mathscr{A}^t} U_t(\varphi^t,y), \quad t \in \mathscr{T}', \tag{14}$$

$$U^*_T(y) = e^{\gamma\beta^T r_T(x)}. \tag{15}$$

We now have the following result.

Lemma 2 *For any $t \in \mathscr{T}'$, and for any $\varphi^t \in \mathscr{A}^t$, the function $U_t(\varphi^t, \cdot)$ is lower semi-ananlytic (l.s.a.) on E_Y. Moreover, there exists a sequence of universally measurable functions ψ_k^*, $k = t, \ldots, T-1$ such that*

$$U_t(\varphi^t, y) = e^{\gamma\beta^t r_t(x,\varphi_t(y))} \mathbb{E}_{\mathbb{Q}_{y,t}^{\varphi^t,\psi^{t,*}}} e^{\gamma \Sigma_{k=t+1}^{T} \beta^k r_k(X_k, \varphi_k(Y_k))}. \tag{16}$$

Proof According to (9), and using the definition of $\mathscr{Q}_{y,t}^{\varphi^t,\Psi^t}$, we have that

$$\begin{aligned} U_t(\varphi^t, y) = \inf_{\psi^t \in \Psi^t} \int_{E_Y} \cdots \int_{E_Y} & e^{\gamma \Sigma_{k=t}^{T} \beta^k r_k(x_k, \varphi_k(y_k))} \\ & Q(\mathrm{d}y_T | T-1, y_{T-1}, \varphi_{T-1}(y_{T-1}), \psi_{T-1}(y_{T-1})) \\ & \cdots Q(\mathrm{d}y_{t+1} | t, y, \varphi_t(y), \psi_t(y)). \end{aligned} \tag{17}$$

For a given policy $\varphi \in \mathscr{A}$, define the following functions on E_Y

$$\begin{aligned} V_T(y) &= e^{\gamma\beta^T r_T(x)}, \\ V_t(y) &= \inf_{\theta \in \tau(t,c)} \int_{E_Y} e^{\gamma\beta^t r_t(x,\varphi_t(y))} V_{t+1}(y') Q(\mathrm{d}y' | t, y, \varphi_t(y), \theta), \quad t \in \mathscr{T}'. \end{aligned}$$

We will prove recursively that the functions V_t are l.s.a. in y, and that

$$V_t(y) = U_t(\varphi^t, y), \quad t = 0, \ldots, T-1. \tag{18}$$

Clearly, V_T is l.s.a. in y.

Next, we will prove that $V_{T-1}(y)$ is l.s.a.. By our assumptions, the stochastic kernel $Q(\cdot | T-1, \cdot, \cdot, \cdot)$ is Borel measurable on E_Y given $E_Y \times A \times \Theta$, in the sense of [BS78, Definition 7.2]. Then, the integral $\int_{E_Y} V_T(y') Q(\mathrm{d}y' | T-1, y, a, \theta)$ is l.s.a. on $E_Y \times A \times \Theta$ according to [BS78, Proposition 7.48]. Now, we set (in the notations of [BS78])

$$\begin{aligned} \mathrm{X} &= E_Y \times A, \quad \mathrm{x} = (y, a) \\ \mathrm{Y} &= \Theta, \quad \mathrm{y} = \theta, \\ \mathrm{D} &= \bigcup_{(y,a) \in E_Y \times A} \{y, a\} \times \tau(T-1, c), \\ f(\mathrm{x}, \mathrm{y}) &= \int_{E_Y} V_T(y')\, Q(\mathrm{d}y' | T-1, y, a, \theta). \end{aligned}$$

Note that in view of our assumptions, X and Y are Borel spaces. The set D is closed (see the proof of Lemma 1) and thus analytic. Moreover, $\mathrm{D}_{\mathrm{x}} = \tau(T-1, c)$. Hence, by [BS78, Proposition 7.47], for each $a \in A$ the function

$$\inf_{\theta \in \tau(T-1,c)} \int_{E_Y} V_T(y')\, Q(\mathrm{d}y' | T-1, y, a, \theta)$$

is l.s.a. in y. Thus, it is l.s.a. in (y,a). Moreover, in view of [BS78, Proposition 7.50], for any $\varepsilon > 0$, there exists an analytically measurable function $\psi^{\varepsilon}_{T-1}(y,a)$ such that

$$\inf_{\theta\in\tau(T-1,c)}\int_{E_Y} V_T(y')\,Q(\mathrm{d}y'|T-1,y,a,\theta) = \int_{E_Y} V_T(y')\,Q(\mathrm{d}y'|T-1,y,a,\psi^{\varepsilon}_{T-1}(y,a)) + \varepsilon.$$

Therefore, for any fixed (y,a), we obtain a sequence $\{\psi^{1/n}_{T-1}(y,a), n\in\mathbb{N}\}$ such that

$$\lim_{n\to\infty}\int_{E_Y} V_T(y')\,Q(\mathrm{d}y'|T-1,y,a,\psi^{1/n}_{T-1}(y,a)) = \inf_{\theta\in\tau(T-1,c)}\int_{E_Y} V_T(y')\,Q(\mathrm{d}y'|T-1,y,a,\theta).$$

Due to the assumption that $\tau(T-1,c)$ is compact, there exists a convergent subsequence $\{\psi^{1/n_k}_{T-1}(y,a), k\in\mathbb{N}\}$ such that its limit $\psi^{*}_{T-1}(y,a)$ is universally measurable and satisfies

$$\int_{E_Y} V_T(y')\,Q(\mathrm{d}y'|T-1,y,a,\psi^{*}_{T-1}(y,a)) = \inf_{\theta\in\tau(T-1,c)}\int_{E_Y} V_T(y')\,Q(\mathrm{d}y'|T-1,y,a,\theta).$$

Clearly, the function $e^{\gamma\beta^{T-1}r_{T-1}(x,a)}$ is l.s.a. in (y,a). Thus, since $\varphi_{T-1}(y)$ is a Borel measurable function, using part (3) in [BS78, Lemma 7.30] we conclude that both $e^{\gamma\beta^{T-1}r_{T-1}(x,\varphi_{T-1}(y))}$ and $\inf_{\theta\in\tau(T-1,c)}\int_{E_Y} V_T(y')Q(\mathrm{d}y'|T-1,y,\varphi_{T-1}(y),\theta)$ are l.s.a. in y. Since both these functions are non-negative then, by part (4) in [BS78, Lemma 7.30], we conclude that V_{T-1} is l.s.a. in y. The proof that V_t is l.s.a. in y and ψ^{*}_t exists for $t = 0,\dots,T-2$, follows analogously. We also obtain that

$$\int_{E_Y} V_t(y')\,Q(\mathrm{d}y'|t-1,y,a,\psi^{*}_{t-1}(y,a)) = \inf_{\theta\in\tau(t-1,c)}\int_{E_Y} V_t(y')\,Q(\mathrm{d}y'|t-1,y,a,\theta), \tag{19}$$

for any $t = 1,\dots,T-1$.

It remains to verify (18). For $t = T-1$, by (17), we have

$$U_{T-1}(\varphi^{T-1},y) = \inf_{\theta\in\tau(T-1,c)}\int_{E_Y} e^{\gamma\beta^{T-1}r_{T-1}(x,\varphi_{T-1}(y))}V_T(y')\,Q(\mathrm{d}y'|T-1,y,\varphi_{T-1}(y),\theta) = V_{T-1}(y).$$

Therefore, $U_{T-1}(\varphi^{T-1},\cdot)$ is l.s.a.. Assume that for $t = 1,\dots,T-1$, $U_t(\varphi^{t},y) = V_t(y)$, and it is l.s.a.. Then, for any $y_{t-1}\in E_Y$, with the notation $\psi^{t-1} = (\psi_{t-1},\psi^{t})$, we get

$$
\begin{aligned}
&U_{t-1}(\varphi^{t-1}, y_{t-1}) \\
&\quad = \inf_{(\psi_{t-1},\psi^t)\in\Psi^{t-1}} \int_{E_Y} \cdots \int_{E_Y} e^{\gamma\Sigma_{k=t-1}^{T-1}\beta^k r_k(x_k,\varphi_k(y_k))+\gamma\beta^T r_T(x_T)} \\
&\qquad\qquad \prod_{k=t}^{T} Q(\mathrm{d}y_k|k-1, y_{k-1}, \varphi_{k-1}(y_{k-1}), \psi_{k-1}(y_{k-1})) \\
&\quad \geq \inf_{(\psi_{t-1},\psi^t)\in\Psi^{t-1}} \int_{E_Y} e^{\gamma\beta^{t-1} r_{t-1}(x_{t-1},\varphi_{t-1}(y_{t-1}))} V_t(y_t) \\
&\qquad\qquad Q(\mathrm{d}y_t|t-1, y_{t-1}, \varphi_{t-1}(y_{t-1}), \psi_{t-1}(y_{t-1})) \\
&\qquad - \inf_{\theta\in\tau(t-1,c)} \int_{E_Y} e^{\gamma\beta^{t-1} r_{t-1}(x_{t-1},\varphi_{t-1}(y_{t-1}))} V_t(y_t) \\
&\qquad\qquad Q(\mathrm{d}y_t|t-1, y_{t-1}, \varphi_{t-1}(y_{t-1}), \psi_{t-1}(y_{t-1})) \\
&\quad = V_{t-1}(y_{t-1}).
\end{aligned}
$$

Next, fix $\varepsilon > 0$, and let $\psi^{t,\varepsilon}$ denote an ε-optimal selectors sequence starting at time t, namely

$$
\begin{aligned}
&\int_{E_Y} \cdots \int_{E_Y} e^{\gamma\Sigma_{k=t}^{T}\beta^k r_k(x_k,\varphi_k(y_k))} \prod_{k=t+1}^{T} Q(\mathrm{d}y_k|k-1, y_{k-1}, \varphi_{k-1}(y_{k-1}), \psi_{k-1}^{t,\varepsilon}(y_{k-1})) \\
&\qquad\qquad \leq U_t(\varphi^t, y_t) + \varepsilon.
\end{aligned}
$$

Consequently, for any $y_{t-1} \in E_Y$,

$$
\begin{aligned}
&U_{t-1}(\varphi^{t-1}, y_{t-1}) = \inf_{(\psi_{t-1},\psi^t)\in\Psi^{t-1}} \int_{E_Y} \cdots \int_{E_Y} e^{\gamma\Sigma_{k=t-1}^{T}\beta^k r_k(x_k,\varphi_k(y_k))} \\
&\qquad\qquad \prod_{k=t}^{T} Q(\mathrm{d}y_k|k-1, y_{k-1}, \varphi_{k-1}(y_{k-1}), \psi_{k-1}(y_{k-1})) \\
&\quad \leq \inf_{\psi_{t-1}\in\tau(t-1,c)} \int_{E_Y} \cdots \int_{E_Y} e^{\gamma\Sigma_{k=t-1}^{T}\beta^k r_k(x_k,\varphi_k(y_k))} \\
&\qquad \prod_{k=t+1}^{T} Q(\mathrm{d}y_k|k-1, y_{k-1}, \varphi_{k-1}(y_{k-1}), \psi_{k-1}^{t,\varepsilon}(y_{k-1})) \\
&\qquad\qquad \cdots Q(\mathrm{d}y_t|t-1, y_{t-1}, \varphi_{t-1}(y_{t-1}), \psi_{t-1}(y_{t-1})) \\
&\quad \leq \inf_{\varphi_{t-1}\in\tau(t-1,c)} \int_{E_Y} U_t(\varphi^t, y_t) Q(\mathrm{d}y_t|t-1, y_{t-1}, \varphi_{t-1}(y_{t-1}), \psi_{t-1}(y_{t-1})) + \varepsilon \\
&\quad = \inf_{\varphi_{t-1}\in\tau(t-1,c)} \int_{E_Y} V_t(y_t) Q(\mathrm{d}y_t|t-1, y_{t-1}, \varphi_{t-1}(y_{t-1}), \psi_{t-1}(y_{t-1})) + \varepsilon \\
&\quad = V_{t-1}(y_{t-1}) + \varepsilon.
\end{aligned}
$$

Since ε is arbitrary, (18) is justified. In particular, $U_t(\varphi^t, \cdot)$ is l.s.a. for any $t \in \mathscr{T}'$. Finally, in view of (19), the equality (16) follows immediately. This concludes the proof. □

Now we are in the position to prove the main result in this paper.

Theorem 1 *For $t = 0, \dots, T$, we have that*

$$U_t^* \equiv W_t. \tag{20}$$

Moreover, the policy φ^* derived in Lemma 1 is adaptive robust-optimal, that is

$$U_t^*(y) = U_t(\varphi^{t,*}, y), \quad t = 0, \dots, T-1. \tag{21}$$

Proof We proceed similarly as in the proof of [Iye05, Theorem 2.1], and via backward induction in $t = T, T-1, \dots, 1, 0$.

For $t = T$, clearly, $U_T^*(y) = W_T(y) = e^{\gamma \beta^T r_T(x)}$ for all $y \in E_Y$. For $t = T-1$ we have, for $y \in E_Y$,

$$\begin{aligned}
U_{T-1}^*(y) &= \sup_{\varphi^{T-1} = \varphi_{T-1} \in \mathscr{A}^{T-1}} \inf_{\theta \in \tau(T-1,c)} \int_{E_Y} e^{\gamma \beta^{T-1} r_{T-1}(x, \varphi_{T-1}(y))} W_T(y') \\
&\qquad\qquad Q(\mathrm{d}y' \mid T-1, y_{T-1}, \varphi_{T-1}(y), \theta) \\
&= \max_{a \in A} \inf_{\theta \in \tau(T-1,c)} \int_{E_Y} e^{\gamma \beta^{T-1} r_{T-1}(x,a)} W_T(y')\, Q(\mathrm{d}y' \mid T-1, y, a, \theta) \\
&= W_{T-1}(y).
\end{aligned}$$

From the above, using Lemma 1, we obtain that U_{T-1}^* is l.s.c. and bounded.

For $t = T-2, \dots, 1, 0$, assume that U_{t+1}^* is l.s.c. and bounded. Recalling the notation $\varphi^t = (\varphi_t, \varphi^{t+1})$, we thus have, $y \in E_Y$,

$$\begin{aligned}
U_t^*(y) &= \sup_{(\varphi_t, \varphi^{t+1}) \in \mathscr{A}^t} \inf_{\theta \in \tau(t,c)} \int_{E_Y} e^{\gamma \beta^t r_t(x, \varphi_t(y))} U_{t+1}(\varphi^{t+1}, y')\, Q(\mathrm{d}y' \mid t, y, \varphi_t(y), \theta) \\
&\le \sup_{(\varphi_t, \varphi^{t+1}) \in \mathscr{A}^t} \inf_{\theta \in \tau(c_t,t)} \int_{E_Y} e^{\gamma \beta^t r_t(x, \varphi_t(y))} U_{t+1}^*(y')\, Q(\mathrm{d}y' \mid t, y, \varphi_t(y), \theta) \\
&= \max_{a \in A} \inf_{\theta \in \tau(t,c)} \int_{E_Y} e^{\gamma \beta^t r_t(x,a)} U_{t+1}^*(y')\, Q(\mathrm{d}y \mid t, y_t, a, \theta) \\
&= \max_{a \in A} \inf_{\theta \in \tau(t,c)} \int_{E_Y} e^{\gamma \beta^t r_t(y,a)} W_{t+1}(y')\, Q(\mathrm{d}y' \mid t, y, a, \theta) \\
&= W_t(y).
\end{aligned}$$

Now, fix $\varepsilon > 0$, and let $\varphi^{t+1,\varepsilon}$ denote an ε-optimal control strategy starting at time $t+1$, that is

$$U_{t+1}(\varphi^{t+1,\varepsilon}, y) \ge U_{t+1}^*(y) - \varepsilon, \quad y \in E_y.$$

Then, for $y \in E_Y$, we have

$$
\begin{aligned}
U_t^*(y) &= \sup_{(\varphi_t,\varphi^{t+1})\in\mathscr{A}^t} \inf_{\theta\in\tau(t,c)} \int_{E_Y} e^{\gamma\beta^t r_t(x,\varphi_t(y))} U_{t+1}(\varphi^{t+1},y')\, Q(\mathrm{d}y' \mid t,y,\varphi_t(y),\theta) \\
&\geq \sup_{(\varphi_t,\varphi^{t+1})\in\mathscr{A}^t} \inf_{\theta\in\tau(t,c)} \int_{E_Y} e^{\gamma\beta^t r_t(x,\varphi_t(y))} U_{t+1}(\varphi^{t+1,\varepsilon},y')\, Q(\mathrm{d}y' \mid t,y,\varphi_t(y),\theta) \\
&\geq \max_{a\in A} \inf_{\theta\in\tau(t,c)} \int_{E_Y} e^{\gamma\beta^t r_t(x,a)} U_{t+1}^*(y')\, Q(\mathrm{d}y' \mid t,y,a,\theta) - \varepsilon \\
&= \max_{a\in A} \inf_{\theta\in\tau(t,c)} \int_{E_Y} W_{t+1}(y')\, Q(\mathrm{d}y' \mid t,y,a,\theta) - \varepsilon \\
&= W_t(y) - \varepsilon.
\end{aligned}
$$

Since ε was arbitrary, the proof of (20) is done. In particular, we have that for any $t \in \mathscr{T}$, the function $U_t^*(\cdot)$ is l.s.c. as well as bounded.

It remains to justify the validity of equality (21). We will proceed again by (backward) induction in t. For $t = T-1$, using (20), we have that

$$
\begin{aligned}
U_{T-1}^*(y) = W_{T-1}(y) &= e^{\gamma\beta^{T-1} r_{T-1}(x,\varphi_{T-1}^*(y))} \\
&\qquad \inf_{\theta\in\tau(t,c)} \int_{E_Y} e^{\gamma\beta^T r_T(x')}\, Q(\mathrm{d}y' \mid T-1,y,\varphi_{T-1}^*(y),\theta) \\
&= e^{\gamma\beta^{T-1} r_{T-1}(x,\varphi_{T-1}^*(y))} \inf_{\mathbb{Q}\in\mathscr{D}_{y,T-1}^{\varphi^{T-1,*},\psi^{T-1}}} \left(\mathbb{E}_{\mathbb{Q}} e^{\gamma\beta^T r_T(X_T)}\right) \\
&= U_{T-1}(\varphi^{T-1,*},y).
\end{aligned}
$$

Moreover, by Lemma 2, we get that

$$
U_{T-1}^*(y) = U_{T-1}(\varphi^{T-1,*},y) = \mathbb{E}_{\mathbb{Q}_{y,T-1}^{\varphi^{T-1,*},\psi^{T-1,*}}} e^{\gamma\beta^{T-1} r_{T-1}(x,\varphi_{T-1}^*(y)) + \gamma\beta^T r_T(X_T)}.
$$

For $t = T-2$, using again (20), Lemma 1, and Lemma 2, we have

$$
\begin{aligned}
U_{T-2}^*(y) = W_{T-2}(y) &= e^{\gamma\beta^{T-2} r_{T-2}(x,\varphi_{T-2}^*(y))} \\
&\quad \times \int_{E_Y} W_{T-1}(y')\, Q(\mathrm{d}y' \mid T-2,y,\varphi_{T-2}^*(y),\psi_{T-2}^*(y,\varphi_{T-2}^*(y))) \\
&= e^{\gamma\beta^{T-2} r_{T-2}(x,\varphi_{T-2}^*(y))} \\
&\quad \times \int_{E_Y} U_{T-1}(\varphi^{T-1,*},y')\, Q(\mathrm{d}y' \mid T-2,y,\varphi_{T-2}^*(y),\psi_{T-2}^*(y,\varphi_{T-2}^*(y))) \\
&= e^{\gamma\beta^{T-2} r_{T-2}(x,\varphi_{T-2}^*(y))} \\
&\quad \times \int_{E_Y} \left(\mathbb{E}_{\mathbb{Q}_{y',T-1}^{\varphi^{T-1,*},\psi^{T-1,*}}} e^{\gamma\beta^{T-1} r_{T-1}(x',\varphi_{T-1}^*(y')) + \gamma\beta^T r_T(X_T)} \right) \\
&\qquad\qquad Q(\mathrm{d}y' \mid T-2,y,\varphi_{T-2}^*(y),\psi_{T-2}^*(y,\varphi_{T-2}^*(y))) \\
&= \mathbb{E}_{\mathbb{Q}_{y,T-2}^{\varphi^{T-2,*},\psi^{T-2,*}}} e^{\gamma\beta^{T-2} r_{T-2}(x,\varphi_{T-2}^*(y)) + \gamma\beta^{T-1} r_{T-1}(x',\varphi_{T-1}^*(y')) + \gamma\beta^T r_T(X_T)}.
\end{aligned}
$$

Hence, we have that $U^*_{T-2}(y)$ is attained at $\varphi^{T-2,*}$, and therefore $U^*_{T-2}(y) = U_{T-2}(\varphi^{T-2,*}, y)$. The rest of the proof of (21) proceeds in an analogous way. The proof is complete. □

4 Exponential Discounted Tamed Quadratic Criterion Example

In this section, we consider a linear quadratic control problem under model uncertainty as a numerical demonstration of the adaptive robust method. To this end, we consider the 2-dimensional controlled process

$$X_{t+1} = B_1 X_t + B_2 \varphi_t + Z_{t+1},$$

where B_1 and B_2 are two 2×2 matrices and Z_{t+1} is a 2-dimensional normal random variable with mean 0 and convariance matrix

$$\Sigma^* = \begin{pmatrix} \sigma_1^{*,2} & \sigma_{12}^{*,2} \\ \sigma_{12}^{*,2} & \sigma_2^{*,2} \end{pmatrix},$$

where $\sigma_1^{*,2}$, $\sigma_{12}^{*,2}$, and $\sigma_2^{*,2}$ are unknown. Given observations $Z_1, \dots, Z_t$, we consider an unbiased estimator, say $\widehat{\Sigma}_t = \begin{pmatrix} \widehat{\sigma}^2_{1,t} & \widehat{\sigma}^2_{12,t} \\ \widehat{\sigma}^2_{12,t} & \widehat{\sigma}^2_{2,t} \end{pmatrix}$, of the covariance matrix Σ^*, given as

$$\widehat{\Sigma}_t = \frac{1}{t+1} \sum_{i=1}^{t} Z_i Z_i^\top,$$

which can be updated recursively as

$$\widehat{\Sigma}_t = \frac{t(t+1)\widehat{\Sigma}_{t-1} + t Z_t Z_t^\top}{(t+1)^2}.$$

With slight abuse of notations, we denote by Σ, Σ^*, and $\widehat{\Sigma}_t$ the column vectors

$$\begin{aligned} \Sigma^\top &= (\sigma_1^2, \sigma_{12}^2, \sigma_2^2) \\ \Sigma^{*,\top} &= (\sigma_1^{*,2}, \sigma_{12}^{*,2}, \sigma_2^{*,2}) \\ \widehat{\Sigma}_t^\top &= (\widehat{\sigma}^2_{1,t}, \widehat{\sigma}^2_{12,t}, \widehat{\sigma}^2_{2,t}). \end{aligned}$$

The corresponding parameter set is defined as

$$\Theta := \left\{ \Sigma^\top = (\Sigma_1, \Sigma_{12}, \Sigma_2) \in \mathbb{R}^3 : \ 0 \le \Sigma_1, \ \Sigma_2 \le \overline{\Sigma}, \ \Sigma_{12}^2 \le \Sigma_1 \Sigma_2 \right\},$$

where $\overline{\Sigma}$ is some fixed positive constant. Note that the set Θ is a compact subset of $\mathbb{R}^3$.

Putting the above together and considering the augmented state process $Y_t = (X_t, \widehat{\Sigma}_t)$, $t \in \mathscr{T}$, and some finite control set $A \subset \mathbb{R}^2$, we get that the function S defined in (1) is given by

$$S(x,a,z) = B_1 x + B_2 a + z, \quad x,z \in \mathbb{R}^2,\ a \in A,$$

and the function $R(t,c,z)$ showing in (3) satisfies that

$$R(t,c,z) = (\bar{c}_1, \bar{c}_2, \bar{c}_3)^\top, \quad \begin{pmatrix} \bar{c}_1 & \bar{c}_3 \\ \bar{c}_3 & \bar{c}_2 \end{pmatrix} = \frac{(t+1)(t+2)\begin{pmatrix} c_1 & c_3 \\ c_3 & c_2 \end{pmatrix} + (t+1)zz^\top}{(t+2)^2},$$

where $z \in \mathbb{R}^2$, $t \in \mathscr{T}'$, $c = (c_1, c_2, c_3)$. Then, function $\mathbf{G}$ defined in (5) is specified accordingly.

It is well-known that $\sqrt{t+1}(\widehat{\Sigma}_t - \Sigma^*)$ converges weakly to 0-mean normal dsitribution with covariance matrix

$$M_\Sigma = \begin{pmatrix} 2\sigma_1^{*,4} & 2\sigma_1^{*,2}\sigma_{12}^{*,2} & 2\sigma_{12}^{*,4} \\ 2\sigma_1^{*,2}\sigma_{12}^{*,2} & \sigma_1^{*,2}\sigma_2^{*,2} + \sigma_{12}^{*,4} & 2\sigma_{12}^{*,2}\sigma_2^{*,2} \\ 2\sigma_{12}^{*,4} & 2\sigma_{12}^{*,2}\sigma_2^{*,2} & 2\sigma_2^{*,4} \end{pmatrix}.$$

We replace every entry in M_Σ with the corresponding estimator at time $t \in \mathscr{T}'$ and denote by $\widehat{M}_t(\widehat{\Sigma}_t)$ the resulting matrix. With probability one, the matrix $\widehat{M}_t(\widehat{\Sigma}_t)$ is positive-definite. Therefore, we get the confidence region for $\sigma_1^{*,2}$, $\sigma_{12}^{*,2}$, and $\sigma_2^{*,2}$ as

$$\tau(t,c) = \left\{ \Sigma \in \Theta : (t+1)(\Sigma - c)^\top \widehat{M}_t^{-1}(c)(\Sigma - c) \le \kappa \right\},$$

where κ is the $1-\alpha$ quantile of χ^2 distribution with 3 degrees of freedom for some confidence level $0 < \alpha < 1$.

We further take functions $r_T(x) = \min\{b_1, \max\{b_2, x^\top K_1 x\}\}$ and

$$r_t(x,a) = \min\{b_1, \max\{b_2, x^\top K_1 x + a^\top K_2 a\}\},$$

$t \in \mathscr{T}'$, where $x, a \in \mathbb{R}^2$, $b_1 > 0$, $b_2 < 0$, and K_1 and K_2 are two fixed 2-by-2 matrices with negative trace.

For this example, all conditions of the adaptive robust framework of Section 2 are easy to verify, except for the u.h.c. property of set-valued function $\tau(t,\cdot)$, which we establish in the following lemma.

Lemma 3 *For any $t \in \mathscr{T}'$, the set valued function $\tau(t,\cdot)$ is upper hemi-continuous.*

Proof Fix any $t \in \mathscr{T}'$ and $c_0 \in \Theta$. According to our earlier discussion, the matrix $\widehat{M}_t(c_0)$ is positive-definite. Hence, its inverse admits the Cholesky decomposition $\widehat{M}_t^{-1}(c_0) = L_t(c_0)L_t^\top(c_0)$. Consider the change of coordinate system via the linear transformation $\mathscr{L}c = L_t^\top(c_0)c$, and we name it system-$\mathscr{L}$. Let $E \subset \Theta$ be open and such that $\tau(t,c_0) \subset E$. Note that $\mathscr{L}\tau(t,c_0)$ is a closed ball centered at $\mathscr{L}c_0$ in the

system-$\mathscr{L}$. Also, the mapping $\mathscr{L}$ is continuous and one-to-one, hence $\mathscr{L}E$ is an open set and $\mathscr{L}\tau(t,c_0) \subset \mathscr{L}E$. Then, we have that there exists an open ball $B_r(\mathscr{L}c_0)$ in the system-$\mathscr{L}$ centered at $\mathscr{L}c_0$ with radius r such that $\mathscr{L}\tau(t,c_0) \subset B_r(\mathscr{L}c_0) \subset \mathscr{L}E$.

Any ellipsoid centered at c' in the original coordinate system has representation $(c-c')^\top F(c-c') = 1$ which can be written as $(L_t^\top c - L_t^\top c')L^{-1}F(L^\top)^{-1}(L^\top c - L^\top c') = 1$. Hence, it is still an ellipsoid in the $\mathscr{L}$-system after transformation. To this end, we define on Θ a function $h(c) := \|\mathscr{L}c - \mathscr{L}c_0\| + \max\{r_i(c), i = 1,2,3\}$, where $\|\cdot\|$ is the Euclidean norm in the system-$\mathscr{L}$, and $r_i(c)$, $i = 1,2,3$, are the lengths of the three semi axes of the ellipsoid $\mathscr{L}\tau(t,c)$. It is clear that $r_i(c)$, $i = 1,2,3$ are continuous functions.

Next, it is straightforward to check that f is a non-constant continuous function. Therefore, we consider the set $D := \{c \in \Theta : h(c) < r\}$ and see that it is an open set in Θ and non-empty as $c_0 \in D$. Moreover, for any $c \in D$, we get that the ellipsoid $\mathscr{L}\tau(t,c) \subset B_r(\mathscr{L}c_0)$. Hence, $\tau(t,c) \subset E$, and we conclude that $\tau(t,\cdot)$ is u.h.c.. □

Thus, according to Theorem 1, the dynamic risk sensitive optimization problem under model uncertainty can be reduced to the Bellman equations given in (12):

$$W_T(y) = e^{\gamma\beta^T r_T(x)}, \tag{22}$$

$$W_t(y) = \sup_{a\in A}\inf_{\theta\in\tau(t,c)}\int_{\mathbb{R}^2} W_{t+1}(\mathbf{G}(t,y,a,z))e^{\gamma\beta^t(r_t(x,a))}f_Z(z;\theta)dz, \tag{23}$$

$$y = (x,c_1,c_2,c_3) \in E_Y,\ t = T-1,\dots,0,$$

where $f_Z(\cdot;\theta)$ is the density function for two dimensional normal random variable with mean 0 and covariance parameter θ. In the next section, using (22)-(23), we will compute numerically W_t by a machine learning based method. Note that the dimension of the state space E_Y is five in the present case, for which the traditional grid-based numerical method becomes extremely inefficient. Hence, we employ the new approach introduced in [CL19] to overcome the challenges met in our high dimensional robust stochastic control problem.

5 Machine Learning Algorithm and Numerical Results

In this section, we describe our machine learning based method and present the numerical results for our example. Similarly to [CL19], we discretize the state space the relevant state space in the spirit of the regression Monte Carlo method and adaptive design by creating a random (non-gridded) mesh for the process $Y = (X,C)$. Note that the component X depends on the control process, hence at each time t we randomly select from the set A a value of φ_t, and we randomly generate a value of Z_{t+1}, so to simulate the value of X_{t+1}. Next, for each t, we construct the convex hull of simulated Y_t and uniformly generate in-sample points from the convex hull to

obtain a random mesh of Y_t. Then, we solve the equations (22)–(23), and compute the optimal trading strategies at all mesh points.

The key idea of our machine learning based method is to utilize a non-parametric value function approximation strategy called Gaussian process surrogate. For the purpose of solving the Bellman equations (22)–(23), we build GP regression model for the value function $W_{t+1}(\cdot)$ so that we can evaluate

$$\int_{\mathbb{R}^2} W_{t+1}(\mathbf{G}(t,y,a,z))e^{\gamma\alpha^t(r_t(x,a))}f_Z(z;\theta)dz.$$

We also construct GP regression model for the optimal control φ^*. It permits us to apply the optimal strategy to out-of-sample paths without actual optimization, which allows for a significant reduction of the computational cost.

As the GP surrogate for the value function W_t we consider a regression model $\widetilde{W}_t(y)$ such that for any $y^1, \ldots, y^N \in E_Y$, with $y^i \neq y^j$ for $i \neq j$, the random variables $\widetilde{W}_t(y^1), \ldots, \widetilde{W}_t(y^N)$ are jointly normally distributed. Then, given training data $(y^i, W_t(Y^i))$, $i = 1, \ldots, N$, for any $y \in E_Y$, the predicted value $\widetilde{W}_t(y)$, providing an estimate (approximation) of $W_t(y)$ is given by

$$\widetilde{W}(y) = \left(k(y,y^1),\ldots,k(y,y^N)\right)\left[\mathbf{K}+\varepsilon^2\mathbf{I}\right]^{-1}\left(W_t(y^1),\ldots,W_t(y^N)\right)^T,$$

where ε is a tuning parameter, $\mathbf{I}$ is the $N \times N$ identity matrix and the matrix $\mathbf{K}$ is defined as $\mathbf{K}_{i,j} = k(y^i, y^j)$, $i, j = 1, \ldots, N$. The function k is the kernel function for the GP model, and in this work we choose the kernel as the Matern-5/2. Fitting the GP surrogate $\widetilde{W}_t$ means to estimate the hyperparameters inside k through the training data $(y^i, W_t(y^i))$, $i = 1, \ldots, N$ for which we take $\varepsilon = 10^{-5}$. The GP surrogates for φ^* is obtained in an analogous way.

Given the mesh points $\{y_t^i,\ i = 1, \ldots, N_t,\ t \in \mathscr{T}'\}$, the overall algorithm proceeds as follows:

Part A: Time backward recursion for $t = T-1, \ldots, 0$.

1. Assume that $W_{t+1}(y_{t+1}^i)$, and $\varphi_{t+1}^*(y_{t+1}^i) = (\varphi_{t+1}^{1,*}(y_{t+1}^i), \varphi_{t+1}^{2,*}(y_{t+1}^i))$, $i = 1, \ldots, N_t$, are numerically approximated as $\overline{W}_{t+1}(y_{t+1}^i)$, $\overline{\varphi}_{t+1}^{1,*}(y_{t+1}^i)$ and $\overline{\varphi}_{t+1}^{2,*}(y_{t+1}^i)$, $i = 1, \ldots, N_t$, respectively. Also suppose that the corresponding GP surrogates $\widetilde{W}_{t+1}$, $\widetilde{\varphi}_{t+1}^{1,*}$, and $\widetilde{\varphi}_{t+1}^{2,*}$ are fitted through training data $(y_{t+1}^i, \overline{W}_{t+1}(y_{t+1}^i))$, $(y_{t+1}^i, \overline{\varphi}_{t+1}^{1,*}(y_{t+1}^i))$, and $(y_{t+1}^i, \overline{\varphi}_{t+1}^{2,*}(y_{t+1}^i))$, $i = 1, \ldots, N_t$, respectively.
2. For time t, any $a \in A$, $\theta \in \tau(t,c)$ and each y_t^i, $i = 1, \ldots, N_t$, use one-step Monte Carlo simulation to estimate the integral

$$w_t(y,a,\theta) = \int_{\mathbb{R}^2} W_{t+1}(\mathbf{G}(t,y,a,z))e^{\gamma\alpha^t(r_t(x,a))}f_Z(z;\theta)dz.$$

For that, if $Z_{t+1}^1, \ldots, Z_{t+1}^M$ is a sample of Z_{t+1} drawn from the normal distribution corresponding to parameter θ, where $M > 0$ is a positive integer, then estimate the above integral as

$$\widetilde{w}_t(y,a,\theta) = \frac{1}{M}\sum_{i=1}^{M} \widetilde{W}_{t+1}(\mathbf{G}(t,y,a,Z_{t+1}^i))e^{\gamma\alpha^t(r_t(x,a))}.$$

3. For each y_t^i, $i = 1, \ldots, N_t$, and any $a \in A$, compute

$$\overline{w}_t(y_t^i,a) = \inf_{\theta\in\tau(t,c)} \widetilde{w}_t(y_t^i,a,\theta).$$

4. Compute

$$\overline{W}_t(y_t^i) = \max_{a\in A} \overline{w}_t(y_t^i,a),$$

and obtain a maximizer $\overline{\varphi}_t^*(y_t^i) = (\overline{\varphi}_t^{1,*}(y_t^i), \overline{\varphi}_t^{2,*}(y_t^i))$, $i = 1,\ldots,N_t$.
5. Fit a GP regression model for $V_t(\cdot)$ using the results from Step 4 above. Fit GP models for $\varphi_t^{1,*}(\cdot)$ and $\varphi_t^{2,*}(\cdot)$ as well; these are needed for obtaining values of the optimal strategies for out-of-sample paths in Part B of the algorithm.
6. Goto 1: Start the next recursion for $t-1$.

Part B: Forward simulation to evaluate the performance of the GP surrogates $\varphi_t^{1,*}(\cdot)$ and $\varphi_t^{2,*}(\cdot)$, $t = 0,\ldots,T-1$, over the out-of-sample paths.

1. Draw $K > 0$ samples of i.i.d. $Z_1^{*,i},\ldots,Z_T^{*,i}$, $i = 1,\ldots,K$, from the normal distribution corresponding to the assumed true parameter θ^*.
2. All paths will start from the initial state y_0. The state along each path i is updated according to $\mathbf{G}(t,y_t^i,\widetilde{\varphi}_t^*(y_t^i),Z_{t+1}^{*,i})$, where $\widetilde{\varphi}_t^* = (\widetilde{\varphi}_t^{1,*},\widetilde{\varphi}_t^{2,*})$ is the GP surrogate fitted in Part A. Also, compute the running reward $r_t(x_t^i,\widetilde{\varphi}_t^*(y_t^i))$.
3. Obtain the terminal reward $r_T(x_T^i)$, generated by $\widetilde{\varphi}^*$ along the path corresponding to the sample of $Z_1^{*,i}, \ldots, Z_T^{*,i}$, $i = 1, \ldots, K$, and compute

$$W^{\mathrm{ar}} := \frac{1}{\gamma}\ln\left(\frac{1}{K}\sum_{i=1}^{K} e^{\gamma(\Sigma_{t=0}^{T-1}\beta^t r_t(x_t^i,\widetilde{\varphi}_t^*(y_t^i))+\beta^T r_T(x_T^i))}\right) \tag{24}$$

as an estimate of the performance of the optimal adaptive robust risk sensitive strategy φ^*.

For comparison, we also analyze the optimial risk sensitive strategies of the adaptive and strong robust control methods. In (23), if we take $\tau(t,c) = \{c\}$ for any t, then we obtain the adaptive risk sensitive strategy. On the other hand, by taking $\tau(t,c) = \Theta$ for any t and c, we get the strong robust strategy. We will compute W^{ad} and W^{sr} the risk sensitive criteria of adaptive and strong robust, respectively, in analogy to (24).

Next, we apply the machine learning algorithm described above by solving (22)–(23) for a specific set of parameters. In particular, we take: $T = 10$ with one period of time corresponding to one-tenth of a year; the discount factor being equal to 0.3 or equivalently $\beta = 0.3$; the initial state $X_0^\top = (2,2)$; the confidence level $\alpha = 0.1$; in Part A of our algorithm the number of one-step Monte Carlo simulations is $M = 100$; the number of forward simluations in Part B is taken $K = 2000$; the control set A is approximated by the compact set $[-1,1]^2$; the relevant matrices are

$$B_1 = B_2 = \begin{pmatrix} 0.5 & -0.1 \\ -0.1 & 0.5 \end{pmatrix}, \quad K_1 = \begin{pmatrix} 0.7 & -0.2 \\ -0.2 & 0.7 \end{pmatrix}, \quad K_2 = \begin{pmatrix} -200 & 100 \\ 100 & -200 \end{pmatrix}.$$

The assumed true covariance matrix for Z_t, $t \in \mathscr{T}$, as well as initital guess are

$$\Sigma^* = \begin{pmatrix} 0.009 & 0.006 \\ 0.006 & 0.016 \end{pmatrix}, \quad \widehat{\Sigma}_0 = \begin{pmatrix} 0.00625 & 0.004 \\ 0.004 & 0.02025 \end{pmatrix},$$

respectively. The parameter set is chosen as $\Theta = \tau(0, c_0)$, where $c_0^\top = (0.00625, 0.004, 0.02025)$. For all three control approaches, we compute W^{ar}, W^{ad}, and W^{sr}, respectively, for the risk sensitive parameters $\gamma = 0.2$ and $\gamma = 1.5$.

Finally, we report on the computed values of the optimality criterion corresponding to three different methods: adaptive robust (AR), adaptive (AD) and strong robust (SR).

	W^{ar}	W^{ad}	W^{sr}
$\gamma = 0.2$	-319.81	-323.19	-329.53
$\gamma = 1.5$	-427.76	-427.97	-442.97

Table 1 Risk sensitive criteria for AR, AD, and SR.

Acknowledgements

Tomasz R. Bielecki and Igor Cialenco acknowledge support from the National Science Foundation grant DMS-1907568.

References

[BCC17] T. R. Bielecki, I. Cialenco, and T. Chen. Recursive construction of confidence regions. *Electron. J. Statist.*, 11(2):4674–4700, 2017.

[BCC+19] T. R. Bielecki, I. Cialenco, T. Chen, A. Cousin, and M. Jeanblanc. Adaptive Robust Hedging Under Model Uncertainty. *SIAM J. Control Optim.*, 57(2):925–946, 2019.

[Bor85] K. Border. *Fixed Point Theorems with Applications to Economics and Game Theory*. Cambridge University Press, 9 edition, 1985.

[BP03] T. R. Bielecki and S.R. Pliska. Economic properties of the risk sensitive criterion for portfolio management. *Review of Accounting and Finance*, 2:3–17, 2003.

[BR11] N. Bäuerle and U. Rieder. *Markov decision processes with applications to finance*. Universitext. Springer, Heidelberg, 2011.

[BR17] N. Bäuerle and U. Rieder. Zero-sum risk-sensitive stochastic games. *Stochastic Processes and their Applications*, 127(2):622 – 642, 2017.

[BS78] D. P. Bertsekas and S. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, 1978.

[CG91] H. F. Chen and L. Guo. *Identification and stochastic adaptive control*. Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 1991.

[CL19] T. Chen and M. Ludkovski. A machine learning approach to adaptive robust utility maximization and hedging. *Preprint, arXiv:1912.00244*, 2019.

[DL14] M. Davis and S. Lleo. *Risk-Sensitive Investment Management*, volume 19 of *Advanced Series on Statistical Science & Applied Probability*. World Sci., 2014.

[GS89] I. Gilboa and D. Schmeidler. Maxmin expected utility with nonunique prior. *J. Math. Econom.*, 18(2):141–153, 1989.

[HS95] L. P. Hansen and T. J. Sargent. Discounted linear exponential quadratic Gaussian control. *IEEE Transactions on Automatic Control*, 40(5):968–971, 1995.

[HS08] P. L. Hansen and T. J. Sargent. *Robustness*. Princeton University Press, 2008.

[HSTW06] L. P. Hansen, T. J. Sargent, G. Turmuhambetova, and N. Williams. Robust control and model misspecification. *J. Econom. Theory*, 128(1):45–90, 2006.

[Iye05] G. N. Iyengar. Robust Dynamic Programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[KV15] P. R. Kumar and P. Varaiya. *Stochastic systems: estimation, identification and adaptive control*, volume 75 of *Classics in applied mathematics*. SIAM, 2015.