Probability Theory and Stochastic Modelling 94

Amarjit Budhiraja
Paul Dupuis

Analysis and Approximation of Rare Events

Representations and Weak Convergence Methods

Springer

Probability Theory and Stochastic Modelling

Volume 94

The **Probability Theory and Stochastic Modelling** series is a merger and continuation of Springer's two well established series Stochastic Modelling and Applied Probability and Probability and Its Applications. It publishes research monographs that make a significant contribution to probability theory or an applications domain in which advanced probability methods are fundamental. Books in this series are expected to follow rigorous mathematical standards, while also displaying the expository quality necessary to make them useful and accessible to advanced students as well as researchers. The series covers all aspects of modern probability theory including

- Gaussian processes
- Markov processes
- Random Fields, point processes and random sets
- Random matrices
- Statistical mechanics and random media
- Stochastic analysis

as well as applications that include (but are not restricted to):

- Branching processes and other models of population growth
- Communications and processing networks
- Computational methods in probability and stochastic processes, including simulation
- Genetics and other stochastic models in biology and the life sciences
- Information theory, signal processing, and image synthesis
- Mathematical economics and finance
- Statistical methods (e.g. empirical processes, MCMC)
- Statistics for stochastic processes
- Stochastic control
- Stochastic models in operations research and stochastic optimization
- Stochastic models in the physical sciences

More information about this series at http://www.springer.com/series/13205

Amarjit Budhiraja · Paul Dupuis

Analysis and Approximation of Rare Events

Representations and Weak Convergence Methods

Springer

Amarjit Budhiraja
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC, USA

Paul Dupuis
Division of Applied Mathematics
Brown University
Providence, RI, USA

ISSN 2199-3130 ISSN 2199-3149 (electronic)
Probability Theory and Stochastic Modelling
ISBN 978-1-4939-9577-6 ISBN 978-1-4939-9579-0 (eBook)
https://doi.org/10.1007/978-1-4939-9579-0

Mathematics Subject Classification (2010): 60F10, 60H10, 60H15, 65C05

To our families:
Satnam Kaur, Surinder Singh, Wei, and Hans
and
Suzanne, Alexander, Nicole, Phillip, and
Emily

Preface

The theory of large deviations is concerned with various approximations involving rare events. It is also concerned with characterizing the circumstances that lead to a given rare event. Precise mathematical statements usually involve some variation on the following. There is a scaling parameter [say $\varepsilon \in (0, \infty)$], and a collection of integrals M_ε indexed by ε. The large deviation properties of measures appearing in these integrals are identified by showing that the limit of scaled nonlinear transformations of the integrals exists, and also expressing the limit as the solution to a variational problem. A typical example of the scaling and transformation is $-\varepsilon \log \mathrm{M}_\varepsilon$, and thus one seeks to establish that $\lim_{\varepsilon \to 0} -\varepsilon \log M_\varepsilon$ exists and characterize the value of the limit.

Suppose that besides the desired variational expression for the *limit*, one also has a variational representation for the *prelimit*, and moreover that quantities appearing in the prelimit and limit representations are naturally related though a scaling limit, such as a law of large numbers or ergodic limit. Then a very natural approach to the convergence issue is to prove the large deviation limit by proving convergence of the variational characterizations.

This approach, which was first introduced by Richard Ellis and one of the authors in [97], is the one that will be taken in this book. The method of analysis involves two steps: first rewrite the quantities such as $-\varepsilon \log M_\varepsilon$ as solutions to variational problems, and then carry out the asymptotic analysis as $\varepsilon \to 0$. Not surprisingly, there are many choices to be made, both with respect to what representation might be most convenient, and also what methods to use for the convergence analysis. We will not go into any details here regarding how one identifies and proves convenient variational representations for the prelimit objects. This is indeed one of the main contributions of the book, and is the focus of much discussion in the pages to follow. However, a discussion on the convergence methods is much easier. One can interpret the prelimit variational representations as the value of a stochastic control problem. Given this interpretation, we will carry out the convergence analysis by proving convergence of value functions using weak convergence, i.e., convergence in distribution, of the underlying controlled processes.

Because we study variational quantities (specifically, the infimum of a "cost" over a family of "controls"), the application of weak convergence methods is more involved than in the study of just limits of integrals. One establishes lower bounds (which correspond to large deviation upper bounds) for the sequence of costs under more or less arbitrary controls, and these lower bounds identify a candidate limit variational problem. An upper bound on costs (a large deviation lower bound) gives the reverse inequality, thus establishing the limit and validating the candidate limit variational problem. The upper bound is shown by taking a nearly optimal control for the proposed variational characterization of the limit and adapting to construct controls for the prelimit that yield the same cost as $\varepsilon \to 0$. In applying this general argument to specific problems of large deviation analysis, the same scaling limit simplifications that appear in random variables and processes relevant to M_ε correspond to analogous scaling simplifications for their controlled counterparts that are relevant to $-\varepsilon \log M_\varepsilon$. This will be illustrated by many examples. Indeed, the method described above will be applied to *all* large deviation problems considered in this book, and can thus be viewed as a unified approach to the analysis of rare events for a large and diverse collection of problem settings.

An important difference between our treatment and those in most other works on this subject is that exponential probability estimates do not play a role in the proofs. Instead of developing bounds for exponential tightness and exponential closeness, we establish (ordinary) tightness of suitable families of controls and controlled processes. Another key difference is in our treatment of continuous time processes, for which we at no point use an approximation of the process model. Instead, the central proof ingredients are stochastic control representations for functionals of noise processes (i.e., Brownian motion and Poisson random measure). The proofs of these representations *do* require approximations of the noise processes, and thus in a sense the "approximation component" of classical proofs is applied at a more fundamental and abstract level. This allows one to study many complex problems where standard approximation methods would be hard to implement.

We note that the use of weak convergence methods for the asymptotic analysis of value functions that appear in stochastic control is not new, and in fact originates with Harold Kushner in his study of convergence of numerical methods, which was first presented in book form in [181]. The closely related idea of Gamma convergence was introduced independently, around the same time, by De Giorgi [74].

We have divided the book into four parts. In the first part we review general results in the theory of large deviations, such as the contraction principle. We also discuss in detail the many interesting and useful properties of relative entropy, also known as Kullback–Leibler divergence. Relative entropy is central to the definition of the appropriate cost structure in all the variational representations given for prelimit quantities, and so a detailed knowledge of its properties is essential in the convergence analysis. The last item in Part I is a chapter on introductory examples. The reader interested in quickly understanding how the general machinery works in both discrete and continuous time will find all the main issues introduced and explored there.

Part II considers discrete time problems. After showing how the chain rule for relative entropy allows an easy derivation of convenient variational representations, we proceed through a number of important process models, stating and proving large and moderate deviation principles. Some of the examples here also were considered in [97], but for these examples both the derivation of the representation and weak convergence analysis are new and, in our opinion, simpler.

In Part III we formulate and apply representations for continuous time problems. The derivation of these representations, which is much more involved than its discrete time counterparts, had just begun when [97] was published, and none of the results of this part have appeared in book form. The representations and related abstract large and moderate deviation results have found wide use, and only a sample of the many different uses is presented.

The last part of the book is concerned with Monte Carlo methods for problems that involve rare events. It is perhaps not surprising that methods for analyzing the impact of rare events on particular integrals can also be used to design and analyze Monte Carlo-type numerical schemes for the approximation of the same integrals. This is in some sense the newest of the topics considered in the book.

The main background assumed in the book is as follows. The reader should be familiar with weak convergence of probability measures on Polish spaces, as discussed for example in Billingsley [24]. General knowledge of Markov processes in both discrete and continuous time is also assumed. Part III assumes familiarity with stochastic differential equations driven by Brownian motion and Poisson random measures. Throughout the book there is much use of concepts and results from finite dimensional convex analysis. For these we refer to the standard book of Rockafeller [217]. Part IV of the book requires some familiarity with basic definitions and concepts of subsolutions of Hamilton–Jacobi equations, which can be found, for example, in the books [14, 135].

We request readers to notify us if they discover any errors in the book. We will maintain an Errata webpage at http://abudhiraja.web.unc.edu/files/2019/07/Errata.pdf.

Chapel Hill, North Carolina, USA — Amarjit Budhiraja
Providence, Rhode Island, USA — Paul Dupuis

Acknowledgements

Our single greatest indebtedness is to Richard Ellis, who graciously allowed us to appropriate much of the material of Chaps. 1 and 2 from a prior book coauthored with PD. It was necessary to include a presentation of these topics in the present volume, and we did not consider that the exposition of the older work could be much improved. Richard passed away in the summer of 2018. His scholarship, optimism, and cheerful demeanor will be greatly missed.

We thank various funding agencies for their support during the time this book was being written, including the NSF, DOE, AFOSR, and ARO. We are particularly indebted to Mou-Hsiung (Harry) Chang of ARO and Fariba Fahroo of AFOSR, early and enthusiastic supporters of this work. A number of graduate students, postdocs, and colleagues have given feedback and otherwise contributed to the content of the book, including Michael Conroy, Arnab Ganguly, Yiyun Luo, Yixiang Mao, David Lipschutz, Pierre Nyquist, Tuhin Sahai, Michael Snarski, and Guo-Jhen Wu. Of particular note are the contributions of Dane Johnson and Ruoyu Wu, who besides collaborating on some material presented here, provided the numerical simulations used in Part IV.

The starting point for this volume was a set of notes prepared for a short course presented at the Università degli Studi di Padova in 2013, and PD thanks the Department of Mathematics, and in particular Markus Fischer, for their hospitality.

Contents

Part II Discrete Time Processes

Part I
Laplace Principle, Relative Entropy, and Elementary Examples

The intent of this book is to explain how variational representations and weak convergence methods can be used for the qualitative and quantitative analysis of rare events (large deviation theory), and to address related questions of numerical analysis (accelerated Monte Carlo). This introductory part consists of three chapters. In the first chapter, the equivalence between a large deviation principle and the corresponding Laplace principle is demonstrated for random variables that take values in a Polish space. The Laplace principle is a "bounded and continuous test function" characterization, and it asserts the convergence of normalized logarithms of certain exponential integrals. With this exponential integral characterization in hand, the rest of the chapter proves a number of general results in the theory.

Chapter 2 discusses relative entropy and its many attractive properties. Relative entropy plays a central role in everything that is done in the book, owing to its appearance in a fundamental variational formula for the exponential integrals. In particular, the chain rule of relative entropy, which is the key to obtaining useful representations for processes with structure (e.g., Markov processes) is stated and proved. The chapter also proves or references various results on tightness of probability measures that will be used in the weak convergence analysis.

Chapter 3 shows how refined variational representations can be combined with weak convergence arguments to prove the large deviation principles for some basic models: Sanov's theorem and Cramér's theorem, and stochastic differential equations driven by Brownian and Poisson noise (the latter using representations for continuous time processes that will be proved in Chap. 8). Although the analysis of models that will be considered in later chapters requires considerably more detail, all the main ideas concerning how the representations should be used can be seen in these simple examples.

Chapter 1
General Theory

Throughout this chapter $\{X^n\}_{n\in\mathbb{N}}$ is a sequence of random variables defined on a probability space $(\Omega, \mathscr{F}, P)$ and taking values in a complete separable metric space $\mathscr{X}$. As is usual, we will refer to such a space as a **Polish space**. The metric of $\mathscr{X}$ is denoted by $d(x, y)$, and expectation with respect to P by E. The theory of large deviations focuses on random variables $\{X^n\}$ for which the probabilities $P\{X^n \in A\}$ converge to 0 exponentially fast for a class of Borel sets A. The exponential decay rate of these probabilities is expressed in terms of a function I mapping $\mathscr{X}$ into $[0, \infty]$. A function I on $\mathscr{X}$ is called a **rate function on** $\mathscr{X}$, or simply a **rate function**, if I maps $\mathscr{X}$ into $[0, \infty]$ and if for each $M < \infty$ the level set $\{x \in \mathscr{X} : I(x) \le M\}$ is a compact subset of $\mathscr{X}$. We summarize the last property by saying that I has **compact level sets**. A function $f : \mathscr{X} \to [0, \infty]$ is called **lower semicontinuous** if for every $x \in \mathscr{X}$, $f(x) \le \liminf_{y\to x} f(y)$. Since a function having compact level sets is automatically lower semicontinuous and it attains its infimum on any nonempty closed set, a rate function I satisfies these properties. A convention used throughout this book is that the infimum of a rate function over the empty set is ∞.

1.1 Large Deviation Principle

We next define the concept of a large deviation principle (LDP). For A a subset of $\mathscr{X}$, we denote $\inf_{x\in A} I(x)$ by $I(A)$.

Definition 1.1 Let I be a rate function on $\mathscr{X}$. The sequence $\{X^n\}$ is said to satisfy the **large deviation principle on** $\mathscr{X}$ **with rate function** I if the following two conditions hold:

(a) **Large deviation upper bound**. For each closed subset F of $\mathscr{X}$,

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{X^n \in F\} \le -I(F).$$

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_1

(b) **Large deviation lower bound**. For each open subset G of $\mathscr{X}$,

$$\liminf_{n\to\infty} \frac{1}{n} \log P\{X^n \in G\} \geq -I(G).$$

As is well known and proved in Theorem 1.15, if a sequence of random variables satisfies the large deviation principle with some rate function, then the rate function is unique. While the normalization or **scaling sequence** is $1/n$ in the definition just given, one also encounters continuous parameter scalings (e.g., $\varepsilon \in (0, 1)$ with $\varepsilon \downarrow 0$) and other sequences (e.g., $b(n) \to \infty$ as $n \to \infty$ in moderate deviations as in Chap. 5).

A number of authors such as [80, 82] use the following slightly different terminology, which we do not adopt. A "rate function" on $\mathscr{X}$ is a function that maps $\mathscr{X}$ into $[0, \infty]$ and is lower semicontinuous. A "good rate function" is a function that maps $\mathscr{X}$ into $[0, \infty]$ and has compact level sets.

Here are two examples of large deviation principles.

Example 1.2 The first example is a special case of **Cramér's theorem** [Theorem 3.8]. Given positive numbers p and q summing to 1, let $\{v_j\}_{j\in\mathbb{N}_0}$ be a sequence of independent, identically distributed (iid) random variables taking values in $\mathbb{R}$ and having the common distribution $P\{v_j = 0\} = q$ and $P\{v_j = 1\} = p$. For $n \in \mathbb{N}$, define the sample means

$$\frac{S^n}{n} \doteq \frac{1}{n} \sum_{j=0}^{n-1} v_j.$$

According to Cramér's theorem, the sequence $\{S^n/n\}$ satisfies the large deviation principle on $\mathbb{R}$ with rate function

$$I(x) \doteq \sup_{\alpha\in\mathbb{R}} \left[\alpha x - \log \int_{\mathbb{R}} e^{\alpha y}\, \rho(dy)\right],$$

where $\rho \doteq q\delta_0 + p\delta_1$ and for $x \in \mathbb{R}$, δ_x denotes the **Dirac measure** at the point x. The supremum can be explicitly evaluated to give

$$I(x) \doteq \begin{cases} x \log\left(\dfrac{x}{p}\right) + (1-x) \log\left(\dfrac{1-x}{q}\right) & \text{if } x \in [0, 1] \\ \infty & \text{if } x \in \mathbb{R} \setminus [0, 1]. \end{cases}$$

Example 1.3 The second example is known as **Schilder's theorem** [230] [also a special case of Theorem 3.19]. Let $\{W(t)\}_{t\in[0,1]}$ denote a standard Brownian motion taking values in $\mathbb{R}^d$, and for $n \in \mathbb{N}$ consider the process $Y^n = \{Y^n(t)\}_{t\in[0,1]}$ defined by

$$Y^n(t) \doteq \frac{1}{\sqrt{n}} W(t).$$

Then Y^n takes values in the space $\mathscr{C}([0,1]:\mathbb{R}^d)$ consisting of all continuous functions φ that map $[0,1]$ into $\mathbb{R}^d$, and Y^n satisfies $Y^n(0)=0$. When equipped with the supremum norm, $\mathscr{C}([0,1]:\mathbb{R}^d)$ is a separable Banach space and thus a Polish space. For $x\in\mathbb{R}^d$ let $\mathscr{AC}_x([0,1]:\mathbb{R}^d)$ denote the subset consisting of all absolutely continuous functions φ satisfying $\varphi(0)=x$. Schilder's theorem states that the sequence $\{Y^n\}$ satisfies the large deviation principle on $\mathscr{C}([0,1]:\mathbb{R}^d)$ with rate function

$$I(\varphi)\doteq\begin{cases}\dfrac{1}{2}\displaystyle\int_0^1\|\dot\varphi(t)\|^2\,dt & \text{if } \varphi\in\mathscr{AC}_0([0,1]:\mathbb{R}^d)\\ \infty & \text{if } \varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)\setminus\mathscr{AC}_0([0,1]:\mathbb{R}^d).\end{cases}$$

There is an important qualitative interpretation of the points that minimize a rate function over a closed set, and in particular, the sense in which they identify the most likely way a rare event occurs, given that is does occur. Although this result is not used in the sequel, it is one of the most interesting and useful aspects of the theory of large deviations.

Theorem 1.4 *Assume that $C\subset\mathscr{X}$ is closed and that $I(C)=I(C^\circ)<\infty$. Let $G=\{x\in C: I(x)=I(C)\}$, and $G^\varepsilon=\{x\in\mathscr{X}: d(x,G)<\varepsilon\}$. If $\{X^n\}$ satisfies the LDP with rate I, then for every $\varepsilon>0$,*

$$P\left\{X^n\in G^\varepsilon\,\middle|\,X^n\in C\right\}\to 1.$$

Proof Fix $\varepsilon>0$. We first claim that there is $\delta>0$ such that $I(C\backslash G^\varepsilon)-I(C)=\delta$. If not, then one can find $x_i\in C\backslash G^\varepsilon$ such that $I(x_i)\le I(C)+1/i$. Then

$$\{x_i, i\in\mathbb{N}\}\subset\{x: I(x)\le I(C)+1\},$$

and so since $I(C)<\infty$, along a subsequence, $x_{i_k}\to x^*\in C$. However, the lower semicontinuity of I implies $I(x^*)\le\liminf_{k\to\infty}I(x_{i_k})=I(C)$, which contradicts the definition of G^ε. Thus such a $\delta>0$ exists. By Bayes's rule,

$$\begin{aligned}P\left\{X^n\in G^\varepsilon\cap C\,\middle|\,X^n\in C\right\}&=1-P\left\{X^n\in C\backslash G^\varepsilon\,\middle|\,X^n\in C\right\}\\&=1-\frac{P\{X^n\in C\backslash G^\varepsilon\}}{P\{X^n\in C\}}.\end{aligned}$$

Since

$$\liminf_{n\to\infty}\frac{1}{n}\log P\left\{X^n\in C\right\}\ge -I(C^\circ)=-I(C)$$

and

$$\limsup_{n\to\infty}\frac{1}{n}\log P\left\{X^n\in C\backslash G^\varepsilon\right\}\le -I(C)-\delta,$$

the result follows. □

Suppose we would like to prove that a sequence $\{X^n\}$ satisfies the large deviation principle. As we will show in the next section, the large deviation principle will follow, provided we can evaluate, for all bounded continuous functions h mapping $\mathscr{X}$ into $\mathbb{R}$, the asymptotics of quantities of the form

$$\frac{1}{n} \log E \exp\left\{-nh(X^n)\right\} \text{ as } n \to \infty. \tag{1.1}$$

The weak convergence approach initiated in [97] and further developed in this book is ideally suited to such evaluations. The main purpose of this chapter is to introduce some of the machinery needed for its implementation. Other key tools, such as the relative entropy representation for exponential integrals and various properties of relative entropy, will be discussed in the next chapter.

The evaluation of the asymptotics of quantities of the form given in Eq. (1.1) leads to the concept of the Laplace principle. In a sense to be made precise in Sect. 1.2, the Laplace principle is equivalent to the large deviation principle. In Sect. 1.3 we make a small but pleasant detour as we reformulate, in terms of the Laplace principle, a number of basic results in the theory.

The techniques used in Sect. 1.2 to prove the equivalence of the Laplace principle and the large deviation principle as well as the techniques used in Sect. 1.3 are basic in large deviation theory. However, they will not be used much in the remainder of the book. In contrast, the techniques used in Chap. 2 to establish properties of the relative entropy will be applied many times.

Notation and Terminology. We will implement the weak convergence approach by studying the asymptotics of

$$V^n \doteq -\frac{1}{n} \log E \exp\{-nh(X^n)\}$$

as $n \to \infty$. Here V^n is the negative of the quantity given in Eq. (1.1). We first reformulate V^n as a variational or stochastic control problem. By inserting the two annoying minus signs in the formula for V^n, we obtain a minimization problem with nonnegative relative entropy costs, which is more natural than a maximization problem with nonpositive costs. The representations express exponential integrals in terms of stochastic control problems. To distinguish various objects of interest we use the term **Laplace upper bound** to refer to a bound of the form $\limsup_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} \le A$, and **variational lower bound** to refer to the corresponding bound $\liminf_{n\to\infty} -\frac{1}{n} \log E \exp\{-nh(X^n)\} \ge -A$. The terms **Laplace lower bound** and **variational upper bound** are defined analogously. Thus a Laplace lower bound will give lower bounds on $\frac{1}{n} \log E \exp\{-nh(X^n)\}$, while the associated variational lower bound gives lower bounds on the control representation for $-\frac{1}{n} \log E \exp\{-nh(X^n)\}$.

1.2 An Equivalent Formulation of the Large Deviation Principle

In [238] Varadhan proved an important consequence of the large deviation principle that involves the asymptotic behavior of certain expectations. It generalizes the well-known method of Laplace for studying the asymptotics of certain integrals on $\mathbb{R}$. Given h a bounded continuous function mapping $[0, 1]$ into $\mathbb{R}$, **Laplace's method** states that

$$\lim_{n\to\infty} \frac{1}{n} \log \int_0^1 \exp\{-nh(x)\}\, dx = -\min_{x\in[0,1]} h(x).$$

The proof is a straightforward exercise, and if h is smooth, then further analysis yields an asymptotic expansion for the integral as $n \to \infty$. After stating Varadhan's result in the next theorem, we will explain its relationship to the weak convergence approach. We remind the reader that throughout this chapter $\{X^n\}_{n\in\mathbb{N}}$ is a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and taking values in a Polish space $\mathcal{X}$. The metric of $\mathcal{X}$ is denoted by $d(x, y)$, and expectation with respect to P by E.

Theorem 1.5 (VARADHAN) *Assume that the sequence $\{X^n\}$ satisfies the large deviation principle on $\mathcal{X}$ with rate function I. Then for all bounded continuous functions h mapping $\mathcal{X}$ into $\mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} = -\inf_{x\in\mathcal{X}} [h(x) + I(x)]. \tag{1.2}$$

More precisely, the following conclusions hold:

(a) The large deviation upper bound implies that

$$\limsup_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} \le -\inf_{x\in\mathcal{X}} [h(x) + I(x)].$$

(b) The large deviation lower bound implies that

$$\liminf_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} \ge -\inf_{x\in\mathcal{X}} [h(x) + I(x)].$$

Proof (a) There exists $M \in (0, \infty)$ such that $-M \le h(x) \le M$ for all $x \in \mathcal{X}$. For N a positive integer and $j \in \{1, 2, \ldots, N\}$, consider the closed sets

$$F_{N,j} \doteq \left\{ x \in \mathcal{X} : -M + \frac{2(j-1)M}{N} \le -h(x) \le -M + \frac{2jM}{N} \right\}.$$

The large deviation upper bound yields

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} \\
&\quad \le \limsup_{n\to\infty} \frac{1}{n} \log \left(\sum_{j=1}^{N} \int_{F_{N,j}} \exp\{-nh(x)\}\, P\left\{X^n \in dx\right\} \right) \\
&\quad \le \max_{j\in\{1,2,\ldots,N\}} \left[-M + \frac{2jM}{N} - I(F_{N,j}) \right] \\
&\quad \le \max_{j\in\{1,2,\ldots,N\}} \sup_{x\in F_{N,j}} [-h(x) - I(x)] + \frac{2M}{N} \\
&\quad = \sup_{x\in\mathscr{X}} [-h(x) - I(x)] + \frac{2M}{N}.
\end{aligned}$$

In obtaining the second inequality we have used the inequality $\log(\sum_{j=1}^{N} a_j) \le \log N + \max_{j\in\{1,\ldots,N\}} [\log a_j]$ for nonnegative real numbers a_j, $j = 1, \ldots, N$. Sending $N \to \infty$, we obtain

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} &\le \sup_{x\in\mathscr{X}} [-h(x) - I(x)] \\
&= -\inf_{x\in\mathscr{X}} [h(x) + I(x)],
\end{aligned}$$

as claimed.

(b) Given x an arbitrary point in $\mathscr{X}$ and ε an arbitrary positive number, we apply the large deviation lower bound to the open set $G \doteq \{y \in \mathscr{X} : h(y) < h(x) + \varepsilon\}$, obtaining

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} &\ge \liminf_{n\to\infty} \frac{1}{n} \log E \left[1_G(X^n) \exp\{-nh(X^n)\}\right] \\
&\ge -h(x) - \varepsilon + \liminf_{n\to\infty} \frac{1}{n} \log P\{X^n \in G\} \\
&\ge -h(x) - \varepsilon - I(G) \\
&\ge -h(x) - I(x) - \varepsilon.
\end{aligned}$$

Since $x \in \mathscr{X}$ and $\varepsilon > 0$ are arbitrary,

$$\liminf_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} \ge -\inf_{x\in\mathscr{X}} [h(x) + I(x)].$$

This completes the proof of the theorem. □

If we summarize the large deviation principle by the formal notation

$$P\{X^n \in dx\} \asymp \exp\{-nI(x)\}\, dx,$$

then we can write

$$E\exp\{-nh(X^n)\} = \int_{\mathscr{X}} \exp\{-nh(x)\}\, P\{X^n \in dx\}$$
$$\asymp \int_{\mathscr{X}} \exp\{-n(h(x) + I(x))\}\, dx.$$

As in Laplace's method, Varadhan's theorem (Theorem 1.5) states that to exponential order, the main contribution to the integral is due to the largest value of the exponent.

It is convenient to coin phrases to refer to the validity of the limit (1.2) for all bounded continuous functions h as well as to the validity of the upper and lower bounds in parts (a) and (b) of the theorem.

Definition 1.6 Let I be a rate function on $\mathscr{X}$. The sequence $\{X^n\}$ is said to satisfy the **Laplace principle on $\mathscr{X}$ with rate function I** if for all bounded continuous functions h,

$$\lim_{n\to\infty} \frac{1}{n} \log E\exp\{-nh(X^n)\} = -\inf_{x\in\mathscr{X}} [h(x) + I(x)].$$

The term **Laplace principle upper bound** refers to the validity of

$$\limsup_{n\to\infty} \frac{1}{n} \log E\exp\{-nh(X^n)\} \le -\inf_{x\in\mathscr{X}} [h(x) + I(x)]$$

for all bounded continuous functions h, while the term **Laplace principle lower bound** refers to the validity of

$$\liminf_{n\to\infty} \frac{1}{n} \log E\exp\{-nh(X^n)\} \ge -\inf_{x\in\mathscr{X}} [h(x) + I(x)]$$

for all bounded continuous functions h.

Evaluating the Laplace limit for the zero function on $\mathscr{X}$ shows that if a sequence $\{X^n\}$ satisfies a Laplace principle with rate function I, then the infimum of I on $\mathscr{X}$ equals 0. Since a function with compact level sets attains its infimum on a closed set, it follows that there exists a point $x_0 \in \mathscr{X}$ for which $I(x_0) = 0$.

With the last definition, we can express the content of Varadhan's theorem by saying that the large deviation principle implies the Laplace principle with the same rate function. The next theorem, Theorem 1.8, proves the converse. The result is closely related to another converse of Varadhan's theorem due to Bryc [36]. In general, the weak convergence approach directly yields the Laplace principle and thus through Theorem 1.8 can be used to derive the large deviation principle. We will use this technique for proving the large deviation principle throughout the book.

The equivalence between the Laplace principle and the large deviation principle as expressed by Theorems 1.5 and 1.8 can be regarded as an analogue of the portmanteau theorem [Theorem A.2]. The latter states the equivalence between the weak

convergence of probability measures and limits involving closed and open sets. An examination of the proof of the Theorem 1.8 reveals that given a rate function I, the Laplace principle upper bound implies the large deviation upper bound, and the Laplace principle lower bound implies the large deviation lower bound. This and other features of the theorem will be discussed later in the section.

Remark 1.7 Although we assume for convenience throughout this chapter that $\mathscr{X}$ is a Polish space, the properties of completeness and separability are never used in the proofs of Theorems 1.5 and 1.8. Therefore, these results and consequently the equivalence of the Laplace principle and large deviation principle hold for any metric space $\mathscr{X}$.

Theorem 1.8 *The Laplace principle implies the large deviation principle with the same rate function. More precisely, if I is a rate function on $\mathscr{X}$ and the limit*

$$\lim_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} = -\inf_{x\in\mathscr{X}} [h(x) + I(x)]$$

is valid for all bounded continuous functions h, then $\{X^n\}$ satisfies the large deviation principle on $\mathscr{X}$ with rate function I.

Theorem 1.8 states that the large deviation principle follows once we have a suitable asymptotic evaluation of the expectations $E \exp\{-n\, h(X^n)\}$ for all bounded continuous functions h. In the cases that we will treat, we could easily modify our method and prove the large deviation principle by obtaining bounds on the asymptotic behavior of expectations that involve discontinuous functions rather than continuous functions. This class should be large enough to contain suitable approximations to the indicator functions of closed sets and of open balls in $\mathscr{X}$. We also note that in some instances one may find it convenient to consider functions h that are unbounded.

Proof (*of Theorem* 1.8) We assume that I is a rate function on $\mathscr{X}$ and that for all bounded continuous functions h,

$$\lim_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} = -\inf_{x\in\mathscr{X}} [h(x) + I(x)].$$

We want to prove that for each closed set F, the sequence $\{X^n\}$ satisfies the large deviation upper bound

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{X^n \in F\} \le -I(F)$$

and that for each open set G, the sequence $\{X^n\}$ satisfies the large deviation lower bound

$$\liminf_{n\to\infty} \frac{1}{n} \log P\{X^n \in G\} \ge -I(G).$$

Proof of the large deviation upper bound. Given a closed set F, we define the non-negative lower semicontinuous function

$$\varphi(x) \doteq \begin{cases} 0, & \text{if } x \in F, \\ \infty, & \text{if } x \in F^c. \end{cases}$$

Let $d(x, F) \doteq \inf\{d(x, y) : y \in F\}$ denote the distance from x to F, and for $j \in \mathbb{N}$, define

$$h_j(x) \doteq j(d(x, F) \wedge 1). \tag{1.3}$$

Then h_j is a bounded continuous function and $h_j \uparrow \varphi$ as $j \to \infty$. Hence

$$\frac{1}{n} \log P\{X^n \in F\} = \frac{1}{n} \log E \exp\{-n\varphi(X^n)\} \leq \frac{1}{n} \log E \exp\{-nh_j(X^n)\},$$

and so

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \log P\{X^n \in F\} &\leq \lim_{n\to\infty} \frac{1}{n} \log E \exp\{-nh_j(X^n)\} \\ &= -\inf_{x\in\mathcal{X}} [h_j(x) + I(x)]. \end{aligned}$$

We complete the proof by showing that

$$\lim_{j\to\infty} \inf_{x\in\mathcal{X}} [h_j(x) + I(x)] = I(F). \tag{1.4}$$

Half of this is easy. Since $h_j \leq \varphi$,

$$\inf_{x\in\mathcal{X}} [h_j(x) + I(x)] \leq \inf_{x\in\mathcal{X}} [\varphi(x) + I(x)] = \inf_{x\in F} I(x) = I(F),$$

and thus

$$\limsup_{j\to\infty} \inf_{x\in\mathcal{X}} [h_j(x) + I(x)] \leq I(F).$$

The final step is to prove that

$$\liminf_{j\to\infty} \inf_{x\in\mathcal{X}} [h_j(x) + I(x)] \geq I(F).$$

We can assume that $I(F) > 0$, since if $I(F) = 0$, we are done. Since $h_j = 0$ on F,

$$\begin{aligned} \inf_{x\in\mathcal{X}} [h_j(x) + I(x)] &= \min\left(\inf_{x\in F} [h_j(x) + I(x)], \inf_{x\in F^c} [h_j(x) + I(x)] \right) \\ &= \min\left(I(F), \inf_{x\in F^c} [h_j(x) + I(x)] \right). \end{aligned}$$

It suffices to show that

$$\liminf_{j\to\infty} \inf_{x\in F^c} [h_j(x) + I(x)] \geq I(F),$$

which we prove by contradiction. Thus assume that for some $M \in (0, I(F))$,

$$\liminf_{j\to\infty} \inf_{x\in F^c} [h_j(x) + I(x)] < M.$$

Then there exist a subsequence of $j \in \mathbb{N}$ and $\varepsilon \in (0, M/2)$ such that for all j in this subsequence,

$$\inf_{x\in F^c} [h_j(x) + I(x)] \leq M - 2\varepsilon.$$

In addition, for each j there exists $x_j \in F^c$ such that

$$h_j(x_j) + I(x_j) \leq M - \varepsilon.$$

We claim that $d(x_j, F) \to 0$. Indeed, otherwise for some subsubsequence we would have $h_j(x_j) \doteq j(d(x_j, F) \wedge 1) \to \infty$. Since $M - \varepsilon < \infty$, this would contradict the last display. The convergence $d(x_j, F) \to 0$ implies that there exists a sequence $\{y_j\}$ in F such that $d(x_j, y_j) \to 0$. We now use the fact that $\sup_j I(x_j) \leq M - \varepsilon$, which is a consequence of the last display. Since I has compact level sets, it follows that there exist a further subsequence and a point $x^* \in \{x \in \mathscr{X} : I(x) \leq M - \varepsilon\}$ such that $d(x_j, x^*) \to 0$. Of course, $d(x_j, y_j) \to 0$ implies that $d(y_j, x^*) \to 0$. But since the subsequence $\{y_j\}$ lies in F, which is closed, x^* must lie in F, and so $I(x^*) \geq I(F)$. This contradicts the fact that $I(x^*) \leq M - \varepsilon < I(F)$. The contradiction completes the proof of the large deviation upper bound.

Proof of the large deviation lower bound. Let G be an open set. If $I(G) = \infty$, then there is nothing to prove, so we may assume that $I(G) < \infty$. Let x be any point in G such that $I(x) < \infty$ and choose a real number $M > I(x)$. There exists $\delta > 0$ such that $B(x, \delta) \doteq \{y \in \mathscr{X} : d(y, x) < \delta\}$ is a subset of G. In terms of M, x, and δ, we define

$$h(y) \doteq M \left(\frac{d(y, x)}{\delta} \wedge 1 \right). \tag{1.5}$$

This function is bounded and continuous and satisfies $h(x) = 0$, $h(y) = M$ for $y \in B(x, \delta)^c$ and $0 \leq h(z) \leq M$ for all $z \in \mathscr{X}$. We then have

$$\begin{aligned} E \exp\{-nh(X^n)\} &\leq e^{-nM} P\left\{X^n \in B(x, \delta)^c\right\} + P\left\{X^n \in B(x, \delta)\right\} \\ &\leq e^{-nM} + P\left\{X^n \in B(x, \delta)\right\}, \end{aligned}$$

and therefore

$$\begin{aligned}\max\left(\liminf_{n\to\infty}\frac{1}{n}\log P\{X^n\in B(x,\delta)\},-M\right) &\geq \lim_{n\to\infty}\frac{1}{n}\log E\exp\{-nh(X^n)\}\\ &= -\inf_{y\in\mathscr{X}}[h(y)+I(y)]\\ &\geq -h(x)-I(x)\\ &= -I(x).\end{aligned}$$

Since $M > I(x)$ and $B(x,\delta)\subset G$, it follows that

$$\liminf_{n\to\infty}\frac{1}{n}\log P\{X^n\in G\}\geq\liminf_{n\to\infty}\frac{1}{n}\log P\{X^n\in B(x,\delta)\}\geq -I(x)$$

and thus

$$\liminf_{n\to\infty}\frac{1}{n}\log P\{X^n\in G\}\geq -\inf\left[I(x):x\in G, I(x)<\infty\right]=-I(G).$$

This proves the large deviation lower bound. □

Let ξ be any point in $\mathscr{X}$. If we evaluate the limit (1.4) for the closed set $F\doteq\{\xi\}$, then it follows that knowing the Laplace limit for all bounded continuous functions yields the value of the associated rate function at any point. We record this observation in the next lemma. It leads immediately to the fact that a rate function in a Laplace principle is unique [Theorem 1.15].

Corollary 1.9 *Let I be a rate function on $\mathscr{X}$. For every $\xi\in\mathscr{X}$ and $j\in\mathbb{N}$ we define the bounded continuous function h_j on $\mathscr{X}$ by $h_j(x)\doteq j(d(x,\xi)\wedge 1)$. Then*

$$\lim_{j\to\infty}\inf_{x\in\mathscr{X}}[h_j(x)+I(x)]=I(\xi).$$

With the proof of Theorem 1.8 we have completed our basic exposition of the equivalence between the Laplace principle and the large deviation principle. We next present a technical refinement of the Laplace principle that is useful in some circumstances. In order to state it, a definition is needed. Let f be a function mapping $\mathscr{X}$ into $\mathbb{R}$ for which there exists $M<\infty$ such that

$$|f(x)-f(y)|\leq M\,d(x,y)$$

for all x and y in $\mathscr{X}$. Such an f is called **Lipschitz continuous** with constant M. The message of Theorem 1.8 is that if the Laplace limit

$$\lim_{n\to\infty}\frac{1}{n}\log E\exp\{-nh(X^n)\}=-\inf_{x\in\mathscr{X}}[h(x)+I(x)]$$

is valid for all bounded continuous functions h, then the sequence $\{X^n\}$ satisfies the large deviation principle on $\mathscr{X}$ with rate function I. This can easily be strengthened.

Indeed, the functions h_j defined in Eq. (1.3) in the proof of the large deviation upper bound are bounded and Lipschitz continuous as is the function h defined in Eq. (1.5) in the proof of the large deviation lower bound. Thus it suffices if the Laplace limit is valid merely for all bounded Lipschitz continuous functions. As we point out in the next corollary, this can be strengthened even further by considering separately the Laplace principle upper and lower bounds.

Corollary 1.10 *Let I be a rate function on $\mathscr{X}$. If the Laplace limit*

$$\lim_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} = -\inf_{x\in\mathscr{X}} [h(x) + I(x)]$$

is valid for all bounded Lipschitz continuous functions h mapping $\mathscr{X}$ into $\mathbb{R}$, then the sequence $\{X^n\}$ satisfies the Laplace principle on $\mathscr{X}$ with rate function I and the large deviation principle on $\mathscr{X}$ with rate function I. More precisely, the following implications hold.

(a) If the upper bound

$$\limsup_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} \le -\inf_{x\in\mathscr{X}} [h(x) + I(x)]$$

is valid for all bounded Lipschitz continuous functions h, then both the large deviation upper bound and the Laplace principle upper bound are valid with rate function I.

(b) If the lower bound

$$\liminf_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} \ge -\inf_{x\in\mathscr{X}} [h(x) + I(x)]$$

is valid for all bounded Lipschitz continuous functions h, then both the large deviation lower bound and the Laplace principle lower bound are valid with rate function I.

Proof As we have just pointed out, the large deviation bounds are valid simply because the functions h_j and h defined in Eqs. (1.3) and (1.5) are bounded and Lipschitz continuous. The Laplace principle bounds are then obtained by applying Theorem 1.5. □

To introduce the last topic of this section, consider the problem of proving a Laplace principle for a sequence of stochastic processes $\{Y^n(t), t \in [0, T], n \in \mathbb{N}\}$, where T is a fixed positive number and the processes have a fixed initial point $Y^n(0) = y$. Typically the associated rate function will depend on the parameter y. In such cases it is often useful to show that the Laplace principle holds uniformly with respect to the initial point y in compact sets. Another example is the uniformity of large deviation estimates for empirical measures of a Markov chain with respect to the initial condition of the chain, though in this case the rate function often does not depend on the initial condition.

In this section we formulate the concept of a Laplace principle that is uniform with respect to a parameter. We also show that a uniform Laplace principle as formulated here implies the uniform large deviation principle as defined by Freidlin and Wentzell in [140, p. 92] (as the parameter varies over a compact set). In Remark 4.10 we illustrate, for a class of small-noise stochastic differential equations, the ease and naturalness with which this issue of uniformity is handled for the case in which $\mathscr{X} = \mathscr{C}([0, T] : \mathbb{R}^d)$ and $\mathscr{Y} = \mathbb{R}^d$ by the Laplace principle formulation of the theory of large deviations.

A family of rate functions I_y parametrized by $y \in \mathscr{Y}$ is said to have **compact level sets on compacts** if for each compact subset K of $\mathscr{Y}$ and each $M < \infty$, $\cup_{y\in K}\{x \in \mathscr{X} : I_y(x) \leq M\}$ is a compact subset of $\mathscr{X}$.

Definition 1.11 Let I_y be a family of rate functions on $\mathscr{X}$ parametrized by y in a topological space $\mathscr{Y}$ and assume that this family has compact level sets on compacts. Let $\{X^n\}$ be a sequence of $\mathscr{X}$-valued random variables with distributions that depend on $y \in \mathscr{Y}$ and denote the corresponding expectation operator by E_y. The sequence $\{X^n\}$ is said to satisfy the **Laplace principle on $\mathscr{X}$ with rate function I_y uniformly on compacts** if for all compact subsets K of $\mathscr{Y}$ and all bounded continuous functions h mapping $\mathscr{X}$ into $\mathbb{R}$,

$$\lim_{n\to\infty} \sup_{y\in K} \left| \frac{1}{n} \log E_y \exp\{-nh(X^n)\} - F(y, h) \right| = 0, \tag{1.6}$$

where $F(y, h) \doteq -\inf_{x\in\mathscr{X}} [h(x) + I_y(x)]$. The term **uniform Laplace principle upper bound** refers to the validity of

$$\limsup_{n\to\infty} \sup_{y\in K} \left(\frac{1}{n} \log E_y \exp\{-nh(X^n)\} - F(y, h) \right) \leq 0$$

for all compact subsets K of $\mathscr{Y}$ and all bounded continuous functions h. The term **uniform Laplace principle lower bound** refers to the validity of

$$\liminf_{n\to\infty} \inf_{y\in K} \left(\frac{1}{n} \log E_y \exp\{-nh(X^n)\} - F(y, h) \right) \geq 0$$

for all compact subsets K of $\mathscr{Y}$ and all bounded continuous functions h.

It is elementary to show that together, the uniform Laplace principle upper and lower bounds yield the uniform limit (1.6). The following proposition gives a useful criterion for showing these uniform bounds. The Laplace principle formulation is especially convenient for dealing with uniformity properties, because typically the function mapping $y \in \mathscr{Y} \mapsto F(y, h)$ is continuous. This should be contrasted with the discontinuity of the function mapping $y \in \mathscr{Y} \mapsto I_y(A)$, which plays the analogous role in the standard formulation of the theory of large deviations.

Proposition 1.12 *Let I_y be a family of rate functions on $\mathscr{X}$ parametrized by y in a Polish space $\mathscr{Y}$ and assume that this family has compact level sets on compacts. Let*

h be any bounded continuous function mapping $\mathscr{X}$ into $\mathbb{R}$. Assume that whenever $\{y_n\}_{n\in\mathbb{N}}$ is a sequence in $\mathscr{Y}$ converging to a point $y \in \mathscr{Y}$,

$$\lim_{n\to\infty} \frac{1}{n} \log E_{y_n} \exp\{-nh(X^n)\} = F(y,h), \tag{1.7}$$

where $F(y,h) \doteq -\inf_{x\in\mathscr{X}} [h(x) + I_y(x)]$. Then the sequence $\{X^n\}$ satisfies the Laplace principle on $\mathscr{X}$ with rate function I_y uniformly on compacts.

Proof The key observation is that the function mapping $y \in \mathscr{Y} \mapsto F(y,h)$ is continuous. In order to show this, we define for $y \in \mathscr{Y}$,

$$F^n(y,h) \doteq \frac{1}{n} \log E_y \exp\{-nh(X^n)\},$$

and fix $\bar{y} \in \mathscr{Y}$. We claim that for every $\delta > 0$ there exists $N \in \mathbb{N}$ such that for all $n \geq N$ and all $y \in \mathscr{Y}$ satisfying $d(y,\bar{y}) < 1/N$, the inequality $|F^n(y,h) - F(\bar{y},h)| < \delta$ is valid. Sending $n \to \infty$ yields the continuity of $F(\cdot,h)$. We prove the claim by contradiction. Thus suppose that there exists $\delta > 0$ such that for all $N \in \mathbb{N}$ there exist $n_N \geq N$ and a point $y_N \in \mathscr{Y}$ satisfying $d(y_N,\bar{y}) < 1/N$ and $|F^{n_N}(y_N,h) - F(\bar{y},h)| \geq \delta$. The limit (1.7) yields the inequality

$$0 = \lim_{N\to\infty} |F^{n_N}(y_N,h) - F(\bar{y},h)| \geq \delta,$$

which is nonsense. The claim is proved. Now let $\{y_n\}_{n\in\mathbb{N}}$ be any sequence in $\mathscr{Y}$ converging to a point $y \in \mathscr{Y}$. Then (1.7) implies that

$$\lim_{n\to\infty} |F^n(y_n,h) - F(y,h)| = 0. \tag{1.8}$$

We prove that (1.8) yields the uniform convergence in (1.6) for all compact sets K. If this uniform convergence is not valid, then for some compact set K,

$$\limsup_{n\to\infty} \sup_{y\in K} |F^n(y,h) - F(y,h)| > 0.$$

Thus there would exist a sequence $\{y_n\}_{n\in\mathbb{N}}$ in K satisfying

$$\limsup_{n\to\infty} |F^n(y_n,h) - F(y_n,h)| > 0.$$

Without loss of generality we can assume that $y_n \to y$ for some point $y \in K$. Hence by the continuity of $F(\cdot,h)$ we would have

$$\limsup_{n\to\infty} |F^n(y_n,h) - F(y,h)| > 0.$$

Since this contradicts (1.8), the uniform convergence in (1.6) is proved. □

We now recall the notion of a uniform large deviation principle as defined in [140]. We note that the cited reference considers the case in which $\mathscr{X} = \mathscr{C}([0, T] : \mathbb{R}^d)$ and $\mathscr{Y} = \mathbb{R}^d$; however, the definition for general Polish spaces $\mathscr{Y}$ and $\mathscr{X}$ is given similarly. We recall that the distance on the metric space $\mathscr{X}$ is denoted by d.

Definition 1.13 Let I_y be a family of rate functions on $\mathscr{X}$ parametrized by y in a topological space $\mathscr{Y}$ and assume that this family has compact level sets on compacts. Let $\{X^n\}$ be a sequence of $\mathscr{X}$-valued random variables with distributions that depend on $y \in \mathscr{Y}$ with the corresponding probability and expectation operators denoted by P_y and E_y respectively. The sequence $\{X^n\}$ is said to satisfy the **large deviation upper bound on $\mathscr{X}$ with rate function I_y uniformly on compacts** if given $\delta, \gamma \in (0, 1)$, $L \in (0, \infty)$, and compact $K \subset \mathscr{Y}$, there exists $N < \infty$ such that

$$P_y\{d(X^n, \Phi_y(l)) \geq \delta\} \leq \exp\{-n(l - \gamma)\} \tag{1.9}$$

for all $n \geq N$, $y \in K$, and $l \in [0, L]$, where

$$\Phi_y(l) \doteq \{\phi \in \mathscr{X} : I_y(\phi) \leq l\}, \quad l \in [0, \infty), \ y \in \mathscr{Y}.$$

The sequence $\{X^n\}$ is said to satisfy the **large deviation lower bound on $\mathscr{X}$ with rate function I_y uniformly on compacts** if given $\delta, \gamma \in (0, 1)$, $L \in (0, \infty)$, and compact $K \subset \mathscr{Y}$, there exists $N < \infty$ such that

$$P_y\{d(X^n, \phi) < \delta\} \geq \exp\{-n(I_y(\phi) + \gamma)\} \tag{1.10}$$

for all $n \geq N$, $y \in K$, and $\phi \in \Phi_y(L)$.

The sequence $\{X^n\}$ is said to satisfy the **large deviation principle on $\mathscr{X}$ with rate function I_y uniformly on compacts** if it satisfies both the large deviation upper and lower bounds with rate function I_y, uniformly on compacts.

The following result shows that a uniform Laplace principle implies this form of a uniform large deviation principle.

Proposition 1.14 *Let I_y be a family of rate functions on $\mathscr{X}$ parametrized by y in a Polish space $\mathscr{Y}$ and assume that this family has compact level sets on compacts. Let $\{X^n\}$ be a sequence of $\mathscr{X}$-valued random variables with distributions that depend on $y \in \mathscr{Y}$ and suppose that $\{X^n\}$ satisfies the Laplace principle on $\mathscr{X}$ with rate function I_y uniformly on compacts. Then $\{X^n\}$ satisfies the large deviation principle on $\mathscr{X}$ with rate function I_y uniformly on compacts.*

Proof Fix a compact set $K \subset \mathscr{X}$, $\delta, \gamma \in (0, 1)$, and $L < \infty$. It suffices to show that the inequalities in (1.9) and (1.10) are satisfied for all $n \geq N$, $y \in K$, $l \leq L$, and $\phi \in \Phi_y(L)$, for some $N \in \mathbb{N}$. As in the proof of Theorem 1.8 we will approximate certain open and closed sets by suitable Lipschitz continuous functions. We now introduce these functions. For $j \in \mathbb{N}$, $\delta > 0$, and $\phi \in \mathscr{X}$, let

$$h_{j,\delta,\phi}(\psi) \doteq j\left(\frac{d(\psi, \phi)}{\delta} \wedge 1\right), \quad \psi \in \mathscr{X}.$$

These functions will be used to approximate open sets by bounded continuous functions. When clear from context we will suppress j and δ from the notation and write h_ϕ instead of $h_{j,\delta,\phi}$. We will also approximate closed sets of the form

$$F_{y,l,\delta} \doteq \left\{\psi : d(\psi, \Phi_y(l)) \geq \delta\right\},$$

where $y \in K$ and $l \in [0, L]$, by bounded Lipschitz continuous functions $h_{j,\delta,F_{y,l,\delta}}$ defined by

$$h_{j,\delta,F_{y,l,\delta}}(\psi) \doteq j - \bar{h}_{j,\delta,y,l}(\psi), \quad \bar{h}_{j,\delta,y,l}(\psi) \doteq j\left(\frac{d(\psi, \Phi_y(l))}{\delta} \wedge 1\right).$$

Once again we will abbreviate notation and when clear from context write the function $h_{j,\delta,F_{y,l,\delta}}$ as $h_{y,l}$. For the rest of the proof we fix $j = L + 1$.

We claim that for all $y \in K$, $\phi \in \Phi_y(L)$, and $l \in [0, L]$,

$$P_y\{d(X^n, \phi) < \delta\} + e^{-nj} \geq E_y \exp\left\{-nh_\phi(X^n)\right\}, \tag{1.11}$$

$$P_y\{d(X^n, \Phi_y(l)) \geq \delta\} \leq E_y \exp\left\{-nh_{y,l}(X^n)\right\}, \tag{1.12}$$

and

$$\inf_{\psi\in\mathscr{X}}\left[h_\phi(\psi) + I_y(\psi)\right] \leq I_y(\phi) \text{ and } \inf_{\psi\in\mathscr{X}}\left[h_{y,l}(\psi) + I_y(\psi)\right] \geq l. \tag{1.13}$$

Indeed, since $h_\phi(\phi) = 0$, the first inequality in (1.13) holds. Also, because $h_\phi(\psi) = j$ if $d(\psi, \phi) \geq \delta$ and in general $h_\phi(\psi) \geq 0$, (1.11) follows. The inequality in (1.12) holds since $h_{y,l}(\psi) = 0$ whenever $d(\psi, \Phi_y(l)) \geq \delta$ and the expression inside the expectation is nonnegative. Finally, we prove the second inequality in (1.13). If $\psi \notin \Phi_y(l)$, then $I_y(\psi) > l$ and

$$\left[h_{y,l}(\psi) + I_y(\psi)\right] \geq l,$$

since $0 \leq h_{y,l}(\psi)$. If $\psi \in \Phi_y(l)$, then $h_{y,l}(\psi) = j$, and since $j = L + 1 > l$, the inequality in the display holds once more. This verifies the second statement in (1.13) and hence the claim.

Assume for now that the Laplace limit holds uniformly for $y \in K$ and $g \in \mathscr{K}_{K,L}$, where

$$\mathscr{K}_{K,L} \doteq \{g \in \mathscr{C}_b(\mathscr{X}) : g = h_{y,l} \text{ or } g = h_\phi,\ y \in K, l \leq L, \phi \in \Phi_y(L)\}.$$

Thus

$$\sup_{y\in K, g\in\mathscr{K}_{K,L}}\left|\frac{1}{n}\log E_y \exp\left\{-ng(X^n)\right\} + \inf_{\psi\in\mathscr{X}}\left[g(\psi) + I_y(\psi)\right]\right| \to 0 \tag{1.14}$$

as $n \to \infty$. Given $\gamma \in (0, 1)$, there is $N_1 < \infty$ such that for all $g \in \mathscr{K}_{K,L}$ and $y \in K$,

$$\left| \frac{1}{n} \log E_y \exp\left\{-ng(X^n)\right\} + \inf_{\psi \in \mathscr{X}} \left[g(\psi) + I_y(\psi)\right] \right| \leq \gamma/2 \tag{1.15}$$

whenever $n \geq N_1$. Choose $N \geq N_1$ such that for $n \geq N$,

$$e^{-n\gamma/2} \geq e^{-n} + e^{-n\gamma}. \tag{1.16}$$

We claim that (1.9) and (1.10) hold for all $n \geq N$. Indeed, since $j = L + 1$, for $n \geq N$, (1.11) and (1.15) imply

$$\begin{aligned} P_y\{d(X^n, \phi) < \delta\} &\geq \exp\{-n(I_y(\phi) + \gamma/2)\} - e^{-n(L+1)} \\ &\geq \exp\{-n(I_y(\phi) + \gamma)\} \end{aligned}$$

for all $y \in K$ and $\phi \in \Phi_y(L)$, where the second inequality follows from (1.16) on noting that $I_y(\phi) \leq L$ for all $\phi \in \Phi_y(L)$. This proves (1.10).

Now consider (1.9). For all $y \in K$ and $l \leq L$, (1.12) implies that whenever $n \geq N$,

$$\begin{aligned} P_y\{d(X^n, \Phi_y(l)) \geq \delta\} &\leq E_y \exp\left\{-nh_{y,l}(X^n)\right\} \\ &\leq \exp\{-n(l - \gamma/2)\}, \end{aligned}$$

where the second inequality uses (1.15) and (1.13). Thus (1.9) holds for all $n \geq N$, $y \in K$, and $l \leq L$.

Finally, we must prove that (1.14) holds. Note that by the assumption of the proposition, the assertion in (1.14) holds with any finite subset $\mathscr{K}_0$ of $\mathscr{C}_b(\mathscr{X})$ replacing $\mathscr{K}_{K,L}$. Also, if $g_1, g_2 \in \mathscr{C}_b(\mathscr{X})$ are such that $\sup_{x \in \mathscr{X}} |g_1(x) - g_2(x)| \leq \varepsilon$, then the absolute difference between the expressions

$$\left| \frac{1}{n} \log E_y \exp\left\{-ng_i(X^n)\right\} + \inf_{\psi \in \mathscr{X}} \left[g_i(\psi) + I_y(\psi)\right] \right|$$

for $i = 1$ and $i = 2$ is bounded by 2ε.

Thus in order to prove (1.14) it suffices to show that for every $\varepsilon \in (0, 1)$, there is a **finite ε-net** $\mathscr{K}_\varepsilon$ of functions in $\mathscr{C}_b(\mathscr{X})$ for $\mathscr{K}_{K,L}$. In other words, for every $\varepsilon \in (0, 1)$ there is a finite subset $\mathscr{K}_\varepsilon$ in $\mathscr{C}_b(\mathscr{X})$ such that

$$\sup_{g \in \mathscr{K}_{K,L}} \min_{g_\varepsilon \in \mathscr{K}_\varepsilon} \sup_{x \in \mathscr{X}} |g(x) - g_\varepsilon(x)| \leq \varepsilon.$$

Since $\{I_y\}$ has compact level sets on compacts, the space

$$\Phi_{K,L} \doteq \cup_{y \in K} \Phi_y(L)$$

is compact. Then each $g \in \mathscr{K}_{K,L}$ can be identified with a compact subset of this space. In particular, we identify g of the form $h_{y,l}$ with $\Phi_y(l)$, and g of the form h_ϕ with the singleton $\{\phi\}$ for every $\phi \in \Phi_y(L)$, $l \le L$, and $y \in K$. We will use the well-known fact that the collection of closed subsets of $\Phi_{K,L}$, when topologized using the Hausdorff metric, forms a compact Polish space [128, p. 135]. Recall that the **Hausdorff distance** between two closed subsets C_1 and C_2 of $\Phi_{K,L}$ is given by

$$\lambda(C_1, C_2) \doteq \inf \left\{ \varepsilon > 0 : C_1 \subset C_2^\varepsilon \text{ and } C_2 \subset C_1^\varepsilon \right\},$$

where C_i^ε is the ε-fattening $\{\phi : d(\phi, C_i) \le \varepsilon\}$. Since the collection of closed subsets of $\Phi_{K,L}$ is a compact metric space, it admits for every $\varepsilon \in (0,1)$ a finite ε-net, which we denote by $\mathscr{N}_\varepsilon$. Suppose that $\lambda(C_1, C_2) = a$. Then for all $\varepsilon > a$, we have $C_1 \subset C_2^\varepsilon$ and $C_2 \subset C_1^\varepsilon$, which implies for all $\phi \in \mathscr{X}$ that $d(\phi, C_1) \le d(\phi, C_2) + \varepsilon$ and $d(\phi, C_2) \le d(\phi, C_1) + \varepsilon$. Hence

$$F_{C_1}(\psi) \doteq j\left(\frac{d(\psi, C_1)}{\delta} \wedge 1\right) \le j\left(\frac{d(\psi, C_2)}{\delta} \wedge 1\right) + \frac{j\varepsilon}{\delta} = F_{C_2}(\psi) + \frac{j\varepsilon}{\delta},$$

and so by symmetry, $\sup_{x \in \mathscr{X}} |F_{C_1}(x) - F_{C_2}(x)| \le j\varepsilon/\delta$. Also, letting $\tilde{F}_C \doteq j - F_C$, we see that $\sup_{x \in \mathscr{X}} |\tilde{F}_{C_1}(x) - \tilde{F}_{C_2}(x)| \le j\varepsilon/\delta$. Hence an ε-net $\mathscr{N}_\varepsilon$ for closed subsets of $\cup_{y \in K} \Phi_y(L)$ with the Hausdorff metric induces a $j\varepsilon/\delta$-net for

$$\{g \in \mathscr{C}_b(\mathscr{X}) : g = F_C \text{ or } g = \tilde{F}_C, \text{ for some closed subset } C \text{ of } \cup_{y \in K} \Phi_y(L)\}.$$

Since $\tilde{F}_{\Phi_y(l)} = h_{y,l}$ and $F_{\{\phi\}} = h_\phi$, the collection in the last display contains $\mathscr{K}_{K,L}$. This proves that for every $\varepsilon \in (0,1)$, there is a finite ε-net of functions in $\mathscr{C}_b(\mathscr{X})$ for $\mathscr{K}_{K,L}$, and thus the claim follows. □

In the next section we explore the Laplace principle in somewhat more detail, presenting a number of results that are basic to the theory.

1.3 Basic Results in the Theory

The naturalness of formulating the large deviation principle in terms of a Laplace principle can be seen by the relative ease of proof of a number of basic results in the theory.

1. If a sequence of random variables satisfies the Laplace principle with some rate function, then the rate function is unique [Theorem 1.15].
2. The continuous image of a sequence of random variables satisfying the Laplace principle also satisfies the Laplace principle [Theorem 1.16].
3. The Laplace principle is preserved under superexponential approximation [Theorem 1.17].

4. If the Laplace principle is valid, then the Laplace limit holds for certain unbounded continuous functions [Theorem 1.18].
5. If the Laplace principle is valid, then the Laplace limit holds for certain lower semicontinuous functions satisfying a continuity condition [Theorem 1.20].

The first four of these results are standard. Generalizations of the fifth result can be found in [238]. We start by proving that a rate function in a Laplace principle must be unique.

Theorem 1.15 *We assume that $\{X^n\}$ satisfies the Laplace principle on $\mathscr{X}$ with rate function I and with rate function J. Then $I(\xi) = J(\xi)$ for all $\xi \in \mathscr{X}$.*

Proof For $j \in \mathbb{N}$ and any point $\xi \in \mathscr{X}$ we define the bounded continuous function $h_j(x) \doteq j(d(x,\xi) \wedge 1)$. By Corollary 1.9,

$$\lim_{j\to\infty} \lim_{n\to\infty} \frac{1}{n} \log E \exp\{-nh_j(X^n)\} = \lim_{j\to\infty} \inf_{x\in\mathscr{X}} [h_j(x) + I(x)] = I(\xi)$$

and

$$\lim_{j\to\infty} \lim_{n\to\infty} \frac{1}{n} \log E \exp\{-nh_j(X^n)\} = \lim_{j\to\infty} \inf_{x\in\mathscr{X}} [h_j(x) + J(x)] = J(\xi).$$

Thus $I(\xi) = J(\xi)$. □

The next result, known as the **contraction principle**, is a standard tool in the theory.

Theorem 1.16 (CONTRACTION PRINCIPLE) *Let $\mathscr{X}$ and $\mathscr{Y}$ be Polish spaces, I a rate function on $\mathscr{X}$, and f a continuous function mapping $\mathscr{X}$ into $\mathscr{Y}$. The following conclusions hold:*

(a) *For each $y \in \mathscr{Y}$,*

$$J(y) \doteq \inf\left[I(x) : x \in f^{-1}(y)\right]$$

is a rate function on $\mathscr{Y}$.

(b) *If $\{X^n\}$ satisfies the Laplace principle on $\mathscr{X}$ with rate function I, then $\{f(X^n)\}$ satisfies the Laplace principle on $\mathscr{Y}$ with rate function J.*

Proof (a) Given $M < \infty$, we define the level sets

$$L_J(M) \doteq \{y \in \mathscr{Y} : J(y) \le M\} \text{ and } L_I(M) \doteq \{x \in \mathscr{X} : I(x) \le M\}.$$

The definition of J implies that $L_J(M) \supset f(L_I(M))$. On the other hand, since I is a rate function, for each $y \in f(\mathscr{X})$ the infimum in the definition of J is attained at some x in the closed set $f^{-1}(y)$. It follows that $L_J(M) \subset f(L_I(M))$, which when coupled with the opposite inclusion yields that $L_J(M) = f(L_I(M))$. Since f is

continuous and the level sets of I are compact, this formula shows that the level sets of J are compact. Since J is obviously nonnegative, we have shown that J is a rate function on $\mathscr{Y}$.

(b) For every bounded continuous function h mapping $\mathscr{Y}$ into $\mathbb{R}$, the composition $h \circ f$ is a bounded continuous function mapping $\mathscr{X}$ into $\mathbb{R}$. Hence

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(f(X^n))\} &= -\inf_{x\in\mathscr{X}} [h(f(x)) + I(x)] \\ &= -\inf_{y\in\mathscr{Y}} [h(y) + J(y)].\end{aligned}$$

Since we have already checked that J is a rate function on $\mathscr{Y}$, the proof of part (b) is complete. □

Theorem 1.17 *For $n \in \mathbb{N}$ let X^n and Y^n be random variables that are defined on the same probability space $(\Omega, \mathscr{F}, P)$ and take values in $\mathscr{X}$. We assume that $\{X^n\}$ satisfies the Laplace principle on $\mathscr{X}$ with rate function I and that $\{Y^n\}$ is* ***superexponentially close*** *to $\{X^n\}$ in the following sense: for each $\delta > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{d(Y^n, X^n) > \delta\} = -\infty. \tag{1.17}$$

Then $\{Y^n\}$ satisfies the Laplace principle on $\mathscr{X}$ with the same rate function I.

Proof By Corollary 1.10, it suffices to verify the Laplace limit

$$\lim_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(Y^n)\} = -\inf_{x\in\mathscr{X}} [h(x) + I(x)]$$

for all bounded Lipschitz continuous functions h mapping $\mathscr{X}$ into $\mathbb{R}$. Let h be any such function and let M denote its Lipschitz constant. Then for every $\delta > 0$,

$$\begin{aligned}&\limsup_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(Y^n)\} \\ &\quad= \limsup_{n\to\infty} \frac{1}{n} \log \Big(E\big[1_{\{d(Y^n,X^n)\le\delta\}} \exp\{-nh(Y^n)\}\big] \\ &\qquad\qquad + E\big[1_{\{d(Y^n,X^n)>\delta\}} \exp\{-nh(Y^n)\}\big]\Big) \\ &\quad\le \limsup_{n\to\infty} \frac{1}{n} \log \Big(E\exp\{-nh(X^n) + nM\delta\} + e^{n\|h\|_\infty} P\{d(Y^n, X^n) > \delta\}\Big) \\ &\quad= \left(\limsup_{n\to\infty} \frac{1}{n} \log \Big(E\exp\{-nh(X^n)\}\Big) + M\delta\right) \\ &\qquad\vee \left(\|h\|_\infty + \limsup_{n\to\infty} \frac{1}{n} \log P\{d(Y^n, X^n) > \delta\}\right) \\ &\quad= -\inf_{x\in\mathscr{X}} [h(x) + I(x)] + M\delta,\end{aligned}$$

where the last equality follows from (1.17). Sending $\delta \to 0$ gives the Laplace principle upper bound

$$\limsup_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(Y^n)\} \le - \inf_{x\in\mathscr{X}} [h(x) + I(x)].$$

Similarly, for every $\delta > 0$,

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(Y^n)\} \\
&\quad \ge \liminf_{n\to\infty} \frac{1}{n} \log E\big[1_{\{d(Y^n,X^n)\le\delta\}} \exp\{-nh(Y^n)\}\big] \\
&\quad \ge \liminf_{n\to\infty} \frac{1}{n} \log E\big[1_{\{d(Y^n,X^n)\le\delta\}} \exp\{-nh(X^n) - nM\delta\}\big] \\
&\quad \ge \liminf_{n\to\infty} \frac{1}{n} \log \big(E \exp\{-nh(X^n) - nM\delta\} - e^{n\|h\|_\infty - nM\delta} P\{d(Y^n, X^n) > \delta\}\big)^+ \\
&\quad = - \inf_{x\in\mathscr{X}} [h(x) + I(x)] - M\delta,
\end{aligned}$$

where the last equality is due to the fact that if for nonnegative a_n, b_n, and $a \in \mathbb{R}$, one has $\lim_{n\to\infty} \frac{1}{n} \log a_n = a$ and $\limsup_{n\to\infty} \frac{1}{n} \log b_n = -\infty$, then $\liminf_{n\to\infty} \frac{1}{n} \log(a_n - b_n)^+ = a$. Sending $\delta \to 0$ gives the Laplace principle lower bound

$$\liminf_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(Y^n)\} \ge - \inf_{x\in\mathscr{X}} [h(x) + I(x)].$$

This completes the proof of the theorem. □

If the Laplace principle is valid, then a natural question is whether the Laplace limit can be evaluated for certain unbounded continuous functions mapping $\mathscr{X}$ into $\mathbb{R}$. We next point out a class of such functions for which this can be carried out. The following theorem is due to Varadhan [238].

Theorem 1.18 *Assume that $\{X^n\}$ satisfies the Laplace principle on $\mathscr{X}$ with rate function I. Let h be a continuous function mapping $\mathscr{X}$ into $\mathbb{R}$ for which*

$$\lim_{C\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log E\big[1_{\{h\le -C\}}(X^n) \exp\{-nh(X^n)\}\big] = -\infty. \tag{1.18}$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log E \exp\{-nh(X^n)\} = - \inf_{x\in\mathscr{X}} [h(x) + I(x)] \tag{1.19}$$

and the limit is finite. In particular, if h is bounded below on the union of the supports of the $\{X^n\}$, then condition (1.18) is satisfied and the limit (1.19) holds and is finite.

Proof To streamline the proof, we introduce the notation

$$\Lambda^n(A,\varphi) \doteq E\left[1_A(X^n)\,\exp\{-n\varphi(X^n)\}\right]$$

for A a Borel subset of $\mathscr{X}$ and φ a measurable function mapping $\mathscr{X}$ into $\mathbb{R}$. Our goal is to prove that

$$\lim_{n\to\infty}\frac{1}{n}\log\Lambda^n(\mathscr{X},h) = -\inf_{x\in\mathscr{X}}[h(x)+I(x)]$$

and that the limit is finite. We first obtain a lower bound on $\frac{1}{n}\log\Lambda^n(\mathscr{X},h)$ as $n\to\infty$. Given x an arbitrary point in $\mathscr{X}$ and ε an arbitrary positive number, we define the open set $G \doteq \{y\in\mathscr{X} : h(y) < h(x)+\varepsilon\}$. Since the Laplace principle implies the large deviation principle with the same rate function I [Theorem 1.8], the large deviation lower bound yields

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}\log\Lambda^n(\mathscr{X},h) &\geq \liminf_{n\to\infty}\frac{1}{n}\log\Lambda^n(G,h)\\ &\geq -h(x)-\varepsilon+\liminf_{n\to\infty}\frac{1}{n}\log P\{X^n\in G\}\\ &\geq -h(x)-\varepsilon-I(G)\\ &\geq -h(x)-I(x)-\varepsilon.\end{aligned}$$

Since $x\in\mathscr{X}$ and $\varepsilon>0$ are arbitrary, it follows that

$$\liminf_{n\to\infty}\frac{1}{n}\log\Lambda^n(\mathscr{X},h) \geq -\inf_{x\in\mathscr{X}}[h(x)+I(x)].$$

Clearly $\inf_{x\in\mathscr{X}}[h(x)+I(x)] < \infty$. Condition (1.18) guarantees that as $n\to\infty$, the limit superior of $\frac{1}{n}\log\Lambda^n(\mathscr{X},h)$ is finite. Hence the last display implies that $\inf_{x\in\mathscr{X}}[h(x)+I(x)] > -\infty$. We now prove that

$$\limsup_{n\to\infty}\frac{1}{n}\log\Lambda^n(\mathscr{X},h) \leq -\inf_{x\in\mathscr{X}}[h(x)+I(x)].$$

According to (1.18), there exists $C\in(0,\infty)$ satisfying both

$$\limsup_{n\to\infty}\frac{1}{n}\log\Lambda^n(\{h\leq -C\},h) \leq -\inf_{x\in\mathscr{X}}[h(x)+I(x)] \tag{1.20}$$

and $C > \inf_{x\in\mathscr{X}}[h(x)+I(x)]$. In terms of C, we define the bounded continuous function

$$h_C(x) \doteq \begin{cases} h(x), & \text{if } -C\leq h(x)\leq C,\\ C, & \text{if } h(x)\geq C,\\ -C, & \text{if } h(x)\leq -C.\end{cases}$$

By the choice of C and the nonnegativity of I,

$$\inf_{x\in\mathscr{X}}[h_C(x)+I(x)] \geq \inf_{x\in\mathscr{X}}[h(x)+I(x)],$$

and since the Laplace principle is valid,

$$\lim_{n\to\infty}\frac{1}{n}\log\Lambda^n(\mathscr{X},h_C) = -\inf_{x\in\mathscr{X}}[h_C(x)+I(x)].$$

Therefore,

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\log\Lambda^n(\mathscr{X},h)\\
&\quad=\limsup_{n\to\infty}\frac{1}{n}\log\big[\Lambda^n(\{-C\leq h\leq C\},h)+\Lambda^n(\{h>C\},h)\\
&\qquad\qquad\qquad\qquad+\Lambda^n(\{h<-C\},h)\big]\\
&\quad\leq\limsup_{n\to\infty}\frac{1}{n}\log\big[\Lambda^n(\mathscr{X},h_C)+e^{-nC}+\Lambda^n(\{h<-C\},h)\big]\\
&\quad\leq\left(-\inf_{x\in\mathscr{X}}[h_C(x)+I(x)]\right)\vee(-C)\vee\left(-\inf_{x\in\mathscr{X}}[h(x)+I(x)]\right)\\
&\quad=-\inf_{x\in\mathscr{X}}[h(x)+I(x)],
\end{aligned}$$

where the second inequality is from (1.20). This is what we wanted to show. The proof of the theorem is complete. □

In a number of applications, Laplace-type expectations arise that involve discontinuous functions. Theorem 1.20 shows that the Laplace limit can be evaluated if the function is bounded and lower semicontinuous and satisfies a continuity condition. One encounters such functions, for example, in the study of the exit times of processes from smooth regions. Before stating the theorem, we need to know that a bounded lower semicontinuous function can be suitably approximated by a sequence of uniformly bounded Lipschitz continuous functions.

Lemma 1.19 *Let g be a bounded lower semicontinuous function mapping $\mathscr{X}$ into $\mathbb{R}$. Then there exists a sequence $\{g_j\}_{j\in\mathbb{N}}$ of uniformly bounded Lipschitz continuous functions mapping $\mathscr{X}$ into $\mathbb{R}$ with the properties that $g_j\uparrow g$ and that if $\{x_j\}_{j\in\mathbb{N}}$ is a sequence of points in $\mathscr{X}$ converging to some point x^*, then* $\liminf_{j\to\infty} g_j(x_j)\geq g(x^*)$.

Proof We follow the proofs of Lemmas 7.7 and 7.14 in [19]. For $j\in\mathbb{N}$ and $x\in\mathscr{X}$, define

$$g_j(x)\doteq\inf_{y\in\mathscr{X}}[g(y)+j\,d(x,y)].$$

Clearly $g_j\leq g_{j+1}$ and

$$\inf_{y\in\mathscr{X}} g(y) \le g_j(x) \le g(x) + j\,d(x,x) = g(x) \le \sup_{y\in\mathscr{X}} g(y).$$

Thus the sequence $\{g_j\}_{j\in\mathbb{N}}$ is uniformly bounded and nondecreasing, and also $\lim_{j\to\infty} g_j \le g$. For any points x, y, and z in $\mathscr{X}$,

$$g(y) + j\,d(x,y) \le g(y) + j\,d(z,y) + j\,d(x,z),$$

which yields $g_j(x) \le g_j(z) + j\,d(x,z)$, and interchanging x and z gives

$$|g_j(x) - g_j(z)| \le j\,d(x,z).$$

This inequality shows that g_j is Lipschitz continuous on $\mathscr{X}$.

Now let $\{x_j\}_{j\in\mathbb{N}}$ be any sequence of points in $\mathscr{X}$ converging to a point x^* and set $A \doteq \sup_{x\in\mathscr{X}} g(x)$. For $j \in \mathbb{N}$ and $\varepsilon > 0$ there exists $y_j \in \mathscr{X}$ such that

$$A + \varepsilon \ge g_j(x_j) + \varepsilon \ge g(y_j) + j\,d(x_j, y_j) \ge g(y_j).$$

This inequality is violated unless $d(x_j, y_j) \to 0$ as $j \to \infty$, and since $x_j \to x^*$, it follows that $y_j \to x^*$. The lower semicontinuity of g yields

$$\liminf_{j\to\infty} g_j(x_j) + \varepsilon \ge \liminf_{j\to\infty} g(y_j) \ge g(x^*).$$

Since ε is an arbitrary positive number, we have proved that $\liminf_{j\to\infty} g_j(x_j) \ge g(x^*)$. In particular, for each fixed $x \in \mathscr{X}$, $\liminf_{j\to\infty} g_j(x) \ge g(x)$. Since the sequence $\{g_j\}$ is nondecreasing and $\lim_{j\to\infty} g_j \le g$, it follows that $g_j \uparrow g$. This completes the proof. □

We next prove that the Laplace limit is valid for a bounded lower semicontinuous function satisfying a continuity condition. This result can easily be extended to an unbounded function satisfying condition (1.18) in Theorem 1.18. An example of such a function, which will be used in Chap. 17, is $\infty 1_A(x)$ with A open. Generalizations for sequences of functions are given in Sect. 3 of [238].

Theorem 1.20 *Assume that $\{X^n\}$ satisfies the Laplace principle on $\mathscr{X}$ with rate function I. Let g be a bounded lower semicontinuous function mapping $\mathscr{X}$ into $\mathbb{R}$. The following conclusions hold.*

(a) The upper bound

$$\limsup_{n\to\infty} \frac{1}{n} \log E \exp\{-ng(X^n)\} \le -\inf_{x\in\mathscr{X}} [g(x) + I(x)]$$

is valid.

(b) Assume in addition that for each $\varepsilon > 0$, there exists a point $x_\varepsilon \in \mathscr{X}$ such that g is continuous at x_ε and

$$g(x_\varepsilon) + I(x_\varepsilon) \le \inf_{x \in \mathscr{X}} [g(x) + I(x)] + \varepsilon.$$

Then the Laplace limit

$$\lim_{n\to\infty} \frac{1}{n} \log E \exp\{-ng(X^n)\} = -\inf_{x \in \mathscr{X}} [g(x) + I(x)]$$

is valid.

Proof (a) Let $\{g_j\}_{j\in\mathbb{N}}$ be the sequence of functions in Lemma 1.19. Since each function g_j is bounded and continuous and $g_j \le g$, it follows that

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log E \exp\{-n\, g(X^n)\} &\le \limsup_{n\to\infty} \frac{1}{n} \log E \exp\{-n\, g_j(X^n)\} \\ &= -\inf_{x \in \mathscr{X}} [g_j(x) + I(x)].\end{aligned}$$

In order to complete the proof we must show that

$$\lim_{j\to\infty} \inf_{x \in \mathscr{X}} [g_j(x) + I(x)] = \inf_{x \in \mathscr{X}} [g(x) + I(x)].$$

The sequence of infima $\inf_{x\in\mathscr{X}} [g_j(x) + I(x)]$ is nondecreasing, and

$$\limsup_{j\to\infty} \inf_{x \in \mathscr{X}} [g_j(x) + I(x)] \le \inf_{x \in \mathscr{X}} [g(x) + I(x)] < \infty.$$

We now prove that

$$\liminf_{j\to\infty} \inf_{x \in \mathscr{X}} [g_j(x) + I(x)] \ge \inf_{x \in \mathscr{X}} [g(x) + I(x)].$$

Since I is a rate function and each g_j is bounded and continuous, there exists a sequence $\{x_j\}_{j\in\mathbb{N}}$ such that

$$g_j(x_j) + I(x_j) = \inf_{x \in \mathscr{X}} [g_j(x) + I(x)] \le \inf_{x \in \mathscr{X}} [g(x) + I(x)].$$

The uniform boundedness of the sequence $\{g_j\}$ implies that $\sup_{j\in\mathbb{N}} I(x_j) < \infty$, and since I has compact level sets, there exists a subsequence of $\{x_j\}$ converging to some $x^* \in \mathscr{X}$. A property of $\{g_j\}$ stated at the end of Lemma 1.19 gives the required lower limit

$$\begin{aligned}\liminf_{j\to\infty} \inf_{x \in \mathscr{X}} [g_j(x) + I(x)] &= \liminf_{j\to\infty} \left[g_j(x_j) + I(x_j)\right] \\ &\ge g(x^*) + I(x^*) \\ &\ge \inf_{x \in \mathscr{X}} [g(x) + I(x)].\end{aligned}$$

This completes the proof of the upper bound.

(b) In order to prove part (b) we must show that

$$\liminf_{n\to\infty} \frac{1}{n} \log E \exp\{-n\, g(X^n)\} \geq - \inf_{x\in\mathscr{X}} [g(x) + I(x)].$$

By the hypothesis on g, for each $\varepsilon > 0$ there exist $x_\varepsilon \in \mathscr{X}$ and $\delta > 0$ such that

$$g(x_\varepsilon) + I(x_\varepsilon) \leq \inf_{x\in\mathscr{X}} [g(x) + I(x)] + \varepsilon$$

and $|g(y) - g(x_\varepsilon)| < \varepsilon$ whenever y lies in the open ball $B(x_\varepsilon, \delta) \doteq \{y \in \mathscr{X} : d(y, x_\varepsilon) < \delta\}$. Since the Laplace principle implies the large deviation principle with the same rate function I [Theorem 1.8], the large deviation lower bound applied to $B(x_\varepsilon, \delta)$ yields the lower limit

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{n} \log E \exp\{-n\, g(X^n)\} \\
&\quad \geq \liminf_{n\to\infty} \frac{1}{n} \log E[1_{B(x_\varepsilon,\delta)}(X^n) \exp\{-n\, g(X^n)\}] \\
&\quad \geq -g(x_\varepsilon) - \varepsilon + \liminf_{n\to\infty} \frac{1}{n} \log P\{X^n \in B(x_\varepsilon, \delta)\} \\
&\quad \geq -g(x_\varepsilon) - I(B(x_\varepsilon, \delta)) - \varepsilon \\
&\quad \geq -g(x_\varepsilon) - I(x_\varepsilon) - \varepsilon \\
&\quad \geq - \inf_{x\in\mathscr{X}} [g(x) + I(x)] - 2\varepsilon.
\end{aligned}$$

Sending $\varepsilon \to 0$ completes the proof. □

In the next chapter we introduce the relative entropy function and show how to represent, via a variational formula involving the relative entropy and in the setting of discrete time processes, the logarithms of the expectations appearing in the Laplace limit

$$\lim_{n\to\infty} \frac{1}{n} \log E \exp\{-n\, h(X^n)\} = - \inf_{x\in\mathscr{X}} [h(x) + I(x)].$$

This is a key step in representing integrals of exponentials in terms of associated stochastic control problems. The representation in terms of minimal cost functions is the basis for analyzing the asymptotic behavior of the expectations and thus, through Theorem 1.8, the basis for proving the large deviation principle for the discrete time models of Chaps. 4–7. The same representations are also the starting point for the analysis of the variance of accelerated Monte Carlo schemes in the last part of the book (Chaps. 15–17). Continuous time models as in Chaps. 10–14 will use related representations proved in Chap. 8 and tailored to the continuous time setting.

1.4 Notes

The material of this chapter is mostly taken with little modification from [97]. The main exception is Proposition 1.14, which shows that the uniform Laplace principle implies the form of uniform large deviation principle used in [140]. This is the formulation that is used in the proofs of various large-time and metastability properties of finite dimensional processes that satisfy a "small-noise" large deviation principle on path space. Although the notion of a uniform Laplace principle was introduced in [97], it was not shown that the corresponding uniform large deviation principle as in [140] followed.

There are a number of important parts of the theory that will not be addressed in this book. One is the Gärtner–Ellis theorem [120, 144], which gives conditions for the LDP to hold for $\mathbb{R}^d$-valued random variables that are phrased in terms of convergence of the corresponding moment-generating functions. Another is the just mentioned extensive and widely used theory for the large-time behavior of Markov processes, due to Freidlin and Wentzell [140], that can be viewed as small perturbations of deterministic systems. Although we do not present any aspects of that theory, as noted, our formulation of uniform large deviation and Laplace principles in the present chapter is tailored to the type of uniformity needed for their theory, and that these issues are still under investigation for processes with infinite dimensional state variables [189, 227]. Other directions in which the theory has developed but which we do not pursue include higher-order corrections to the large deviation approximation (e.g., [7, 9, 136]) and a special focus on Gaussian models (see, e.g., [58] and references therein).

There are a number of monographs that develop other perspectives and aspects, including [80–82, 97, 121, 132, 211, 231, 237, 239]. The book [140] is still at this time the most comprehensive in its treatment of the large-time theory mentioned previously, though other presentations of certain aspects of that theory appear in [80, 231].

Chapter 2
Relative Entropy and Tightness of Measures

In this chapter we will collect results on relative entropy and tightness of probability measures that will be used many times in this book.

2.1 Properties of Relative Entropy

Relative entropy will play a key role in the definition of every rate function that we encounter. It arises in the weak convergence approach to large deviations via the variational formula given in part (a) of Proposition 2.2. The derivation of this formula requires only a measurable space $(\mathscr{V}, \mathscr{A})$. We denote by $\mathscr{P}(\mathscr{V})$ the set of probability measures on $(\mathscr{V}, \mathscr{A})$. The probability measure $\gamma \in \mathscr{P}(\mathscr{V})$ is **absolutely continuous** with respect to $\theta \in \mathscr{P}(\mathscr{V})$ if $\theta(A) = 0$ for $A \in \mathscr{A}$ implies $\gamma(A) = 0$, and this relation is denoted by $\gamma \ll \theta$.

For $\theta \in \mathscr{P}(\mathscr{V})$, the **relative entropy** $R(\cdot\|\theta)$ is a mapping from $\mathscr{P}(\mathscr{V})$ into the extended real numbers. It is defined by

$$R(\gamma\|\theta) \doteq \int_{\mathscr{V}} \left(\log \frac{d\gamma}{d\theta}\right) d\gamma$$

whenever $\gamma \in \mathscr{P}(\mathscr{V})$ is absolutely continuous with respect to θ. Otherwise, we set $R(\gamma\|\theta) \doteq \infty$. Occasionally we will refer to $R(\gamma\|\theta)$ as the relative entropy of γ with respect to θ. For $t \in \mathbb{R}$, define $t^- \doteq -(t \wedge 0)$. Since $s(\log s)^-$ is bounded for $s \in [0, \infty)$, it follows that whenever $\gamma \in \mathscr{P}(\mathscr{V})$ is absolutely continuous with respect to θ, we have

$$\int_{\mathscr{V}} \left(\log \frac{d\gamma}{d\theta}\right)^- d\gamma = \int_{\mathscr{V}} \frac{d\gamma}{d\theta} \left(\log \frac{d\gamma}{d\theta}\right)^- d\theta < \infty.$$

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_2

It follows that $\int_{\mathscr{V}} (\log \frac{d\gamma}{d\theta}) d\gamma$ is well defined and that

$$R(\gamma\|\theta) = \int_{\mathscr{V}} \frac{d\gamma}{d\theta} \left(\log \frac{d\gamma}{d\theta} \right) d\theta.$$

The proof of the variational formula in Proposition 2.2 requires the two properties of relative entropy given in the next lemma.

Lemma 2.1 *Let $(\mathscr{V}, \mathscr{A})$ be a measurable space and γ and θ probability measures on $\mathscr{V}$. Then $R(\gamma\|\theta) \geq 0$ and $R(\gamma\|\theta) = 0$ if and only if $\gamma = \theta$.*

Proof In order to prove the nonnegativity, it suffices to consider the case $R(\gamma\|\theta) < \infty$. Since $s \log s \geq s - 1$ with equality if and only if $s = 1$,

$$R(\gamma\|\theta) = \int_{\mathscr{V}} \frac{d\gamma}{d\theta} \left(\log \frac{d\gamma}{d\theta} \right) d\theta \geq \int_{\mathscr{V}} \left(\frac{d\gamma}{d\theta} - 1 \right) d\theta = 0.$$

In this display, equality holds if and only if $d\gamma/d\theta = 1$ θ-a.e., i.e., if and only if $\gamma = \theta$. This completes the proof. □

Part (a) of the next proposition states the variational formula, and part (b) indicates where the infimum in the variational formula is attained. The proposition, though completely elementary, is the cornerstone of the weak convergence approach and will be applied on numerous occasions throughout the book. Its first applications will be in Chap. 3, where we will prove the Laplace principle for some basic examples.

Proposition 2.2 *Let $(\mathscr{V}, \mathscr{A})$ be a measurable space, k a bounded measurable function mapping $\mathscr{V}$ into $\mathbb{R}$, and θ a probability measure on $\mathscr{V}$. The following conclusions hold.*

(a) *We have the variational formula*

$$-\log \int_{\mathscr{V}} e^{-k}\, d\theta = \inf_{\gamma \in \mathscr{P}(\mathscr{V})} \left[R(\gamma\|\theta) + \int_{\mathscr{V}} k\, d\gamma \right]. \tag{2.1}$$

(b) *Let γ_0 denote the probability measure on $\mathscr{V}$ that is absolutely continuous with respect to θ and satisfies*

$$\frac{d\gamma_0}{d\theta}(x) \doteq e^{-k(x)} \cdot \frac{1}{\displaystyle\int_{\mathscr{V}} e^{-k}\, d\theta}.$$

Then the infimum in the variational formula (2.1) is uniquely attained at γ_0.

Proof For part (a) it suffices to prove that

$$-\log \int_{\mathscr{V}} e^{-k}\, d\theta = \inf \left[R(\gamma\|\theta) + \int_{\mathscr{V}} k\, d\gamma : \gamma \in \mathscr{P}(\mathscr{V}), R(\gamma\|\theta) < \infty \right].$$

If $R(\gamma\|\theta) < \infty$, then γ is absolutely continuous with respect to θ, and since θ is absolutely continuous with respect to γ_0, γ is also absolutely continuous with respect to γ_0. Thus

$$\begin{aligned} R(\gamma\|\theta) + \int_{\mathscr{Y}} k\, d\gamma &= \int_{\mathscr{Y}} \left(\log \frac{d\gamma}{d\theta}\right) d\gamma + \int_{\mathscr{Y}} k\, d\gamma \\ &= \int_{\mathscr{Y}} \left(\log \frac{d\gamma}{d\gamma_0}\right) d\gamma + \int_{\mathscr{Y}} \left(\log \frac{d\gamma_0}{d\theta}\right) d\gamma + \int_{\mathscr{Y}} k\, d\gamma \\ &= R(\gamma\|\gamma_0) - \log \int_{\mathscr{Y}} e^{-k}\, d\theta. \end{aligned}$$

We now use the fact that $R(\gamma\|\gamma_0) \geq 0$ and $R(\gamma\|\gamma_0) = 0$ if and only if $\gamma = \gamma_0$. This not only completes the proof of the variational formula in part (a) but also shows that the infimum in the variational formula is uniquely attained at γ_0, as claimed in part (b). □

It is useful in various situations to know that the representation holds for unbounded functionals. One example can be found when $h(x) \doteq \infty 1_{A^c}(x)$, where A is a Borel set and the Laplace functional corresponds to the probability of A. Another that occurs when $\mathscr{X}$ is a normed space is $h(x) = -c\|x\|^2$ with $c > 0$. The following proposition accommodates these sorts of functions. Its proof, which appears in Appendix C, uses a more structured underlying space, and following our convention we assume a Polish space structure.

Proposition 2.3 *Suppose that $\mathscr{X}$ is a Polish space and θ a probability measure on $\mathscr{X}$. The following conclusions hold.*

(a) *If k is a measurable function mapping $\mathscr{X}$ into $\mathbb{R}$ that is bounded from below, then*

$$-\log \int_{\mathscr{X}} e^{-k}\, d\theta = \inf_{\gamma \in \mathscr{P}(\mathscr{X})} \left[R(\gamma\|\theta) + \int_{\mathscr{X}} k\, d\gamma \right]. \tag{2.2}$$

(b) *If k is a measurable function mapping $\mathscr{X}$ into $\mathbb{R}$ that is bounded from above, then*

$$-\log \int_{\mathscr{X}} e^{-k}\, d\theta = \inf_{\gamma \in \Delta(\mathscr{X})} \left[R(\gamma\|\theta) + \int_{\mathscr{X}} k\, d\gamma \right], \tag{2.3}$$

where $\Delta(\mathscr{X}) \doteq \{\gamma \in \mathscr{P}(\mathscr{X}) : R(\gamma\|\theta) < \infty\}$.

(c) *Suppose that $\mathscr{X} = \mathbb{R}^d$ and that for some $\delta > 0$, $\int_{\mathbb{R}^d} \exp\langle \alpha, x\rangle\, \theta(dx) < \infty$ for all $\alpha \in \mathbb{R}^d$ with $\|\alpha\| < \delta$. If there is $\sigma < \infty$ such that $|k(x)| \leq \sigma(1 + \|x\|)$, then*

$$-\log \int_{\mathbb{R}^d} e^{-k}\, d\theta = \inf_{\gamma \in \Delta(\mathbb{R}^d)} \left[R(\gamma\|\theta) + \int_{\mathbb{R}^d} k\, d\gamma \right]. \tag{2.4}$$

In particular, for all $\alpha \in \mathbb{R}^d$,

$$-\log \int_{\mathbb{R}^d} e^{-\langle \alpha, x\rangle}\, d\theta = \inf_{\gamma \in \Delta(\mathbb{R}^d)} \left[R(\gamma \| \theta) + \int_{\mathbb{R}^d} \langle \alpha, x\rangle \, d\gamma \right]. \tag{2.5}$$

We spend the rest of this section investigating seven extremely pleasant properties of relative entropy: convexity, lower semicontinuity, compactness of level sets, uniform integrability of sequences of measures satisfying a suitable relative entropy bound, approximation by sums, monotonicity under mappings, and a very important factorization property known as the "chain rule." These properties will be used repeatedly throughout the book. Several of the proofs rely on the Donsker–Varadhan variational formula for the relative entropy [87], which is stated in part (a) of the next lemma. This formula is dual to the variational formula (2.1). Relative entropy has been the subject of deep study by numerous authors in probability, statistics, information theory, and other areas. In the present section we develop only those properties of the relative entropy that we need.

Let $\mathscr{X}$ be a Polish space and $\{\theta_n\}_{n\in\mathbb{N}}$ a sequence in $\mathscr{P}(\mathscr{X})$. We say that $\{\theta_n\}$ **converges weakly** to a probability measure θ on $\mathscr{X}$, and write $\theta_n \Rightarrow \theta$, if for each bounded continuous function g mapping $\mathscr{X}$ into $\mathbb{R}$,

$$\lim_{n\to\infty} \int_{\mathscr{X}} g\, d\theta_n = \int_{\mathscr{X}} g\, d\theta.$$

We endow $\mathscr{P}(\mathscr{X})$ with the weak topology, which is defined to be the topology corresponding to the weak convergence of probability measures. As we discuss in Appendix A, there exists a metric on $\mathscr{P}(\mathscr{X})$ that is compatible with the weak topology, and with respect to this metric $\mathscr{P}(\mathscr{X})$ is also a Polish space.

The proof that relative entropy has compact level sets relies on Prohorov's theorem, a key result in weak convergence theory that we now state. A family Φ of probability measures on $\mathscr{X}$ is said to be **tight** if for each $\varepsilon > 0$ there exists a compact set K such that

$$\inf_{\gamma\in\Phi} \gamma(K) \geq 1 - \varepsilon.$$

According to Prohorov's theorem [Theorem A.4], Φ is tight if and only if it is **relatively compact** (i.e., it has a compact closure) with respect to weak convergence.

The chain rule asserts a factorization property of relative entropy. In order to formulate it, we introduce a fundamental concept. Let $(\mathscr{V}, \mathscr{A})$ be a measurable space and $\mathscr{Y}$ a Polish space and let $\tau(dy|x)$ be a family of probability measures on $\mathscr{Y}$ parametrized by $x \in \mathscr{V}$. We call $\tau(dy|x)$ a **stochastic kernel** on $\mathscr{Y}$ given $\mathscr{V}$ if for every Borel subset E of $\mathscr{Y}$, the function mapping $x \in \mathscr{V} \mapsto \tau(E|x) \in [0,1]$ is measurable. A basic result is that a family $\tau(dy|x)$ of probability measures on $\mathscr{Y}$ parametrized by $x \in \mathscr{V}$ is a stochastic kernel if and only if the function mapping $x \in \mathscr{V} \mapsto \tau(\cdot|x) \in \mathscr{P}(\mathscr{Y})$ is measurable [Theorem B.4], i.e., if and only if $\tau(\cdot|x)$ is a random variable mapping $\mathscr{V}$ into $\mathscr{P}(\mathscr{Y})$. We have the following useful fact.

If $\tau(dy|x)$ is a stochastic kernel on $\mathscr{Y}$ given $\mathscr{V}$, and f is a measurable function mapping a measurable space $(\mathscr{Z}, \mathscr{D})$ into $(\mathscr{V}, \mathscr{A})$, then $\sigma(dy|z) \doteq \tau(dy|f(z))$ is a stochastic kernel on $\mathscr{Y}$ given $\mathscr{Z}$. We adopt the convention that if $\mathscr{V}$ is empty, then a stochastic kernel on $\mathscr{Y}$ given $\mathscr{V}$ is a probability measure on $\mathscr{Y}$.

Stochastic kernels are commonly encountered in probability. Indeed, let X and Y be random variables taking values in $\mathscr{V}$ and $\mathscr{Y}$, respectively. Then a regular conditional distribution for Y given $X = x$ defines a stochastic kernel on $\mathscr{Y}$ given $\mathscr{V}$. A basic example of a stochastic kernel on $\mathscr{Y}$ given $\mathscr{Y}$ is a transition probability function $\tau(x, dy)$ of a Markov chain taking values in $\mathscr{Y}$. In this example or in cases in which we want to suggest an interpretation of a particular stochastic kernel as a transition probability function, we will deviate from the notation $\tau(dy|x)$. In this context, if X is a random variable mapping a probability space $(\Omega, \mathscr{F}, P)$ into $\mathscr{Y}$, then $\tau(X(\omega), dy)$ is a stochastic kernel on $\mathscr{Y}$ given Ω. Similar examples will arise in the book. Since in general probability spaces are not Polish spaces, this example motivates our general definition of a stochastic kernel.

A **finite measurable partition** of $\mathscr{X}$ is a finite sequence $\pi \doteq \{A_i, i = 1, 2, \ldots, r\}$ consisting of disjoint Borel sets whose union is $\mathscr{X}$. Part (e) of the next lemma gives an approximation property of the relative entropy in terms of sums over finite measurable partitions.

All of the results in the lemma are formulated for arbitrary Polish spaces except part (d), which is stated for $\mathbb{R}^d$. A number of the results are valid for more general spaces, but for ease of exposition we will not point out which ones. Parts (b)–(d) of the lemma are proved in the present section, since similar techniques will be used throughout the book. The proof of part (a) can be found in [97]. Parts (e) and (f) are proved in Appendix C.

Let $\mathscr{C}_b(\mathscr{X})$ [respectively $\mathscr{M}_b(\mathscr{X})$] denote the space of bounded continuous [respectively Borel-measurable] functions mapping $\mathscr{X}$ into $\mathbb{R}$.

Lemma 2.4 *Let $\mathscr{X}$ and $\mathscr{Y}$ be Polish spaces. The relative entropy $R(\cdot\|\cdot)$ has the following properties.*

(a) (***Donsker–Varadhan variational formula***) *For each γ and θ in $\mathscr{P}(\mathscr{X})$,*

$$R(\gamma\|\theta) = \sup_{g\in\mathscr{C}_b(\mathscr{X})} \left[\int_{\mathscr{X}} g\, d\gamma - \log \int_{\mathscr{X}} e^g\, d\theta\right]$$
$$= \sup_{\psi\in\mathscr{M}_b(\mathscr{X})} \left[\int_{\mathscr{X}} \psi\, d\gamma - \log \int_{\mathscr{X}} e^\psi\, d\theta\right].$$

(b) *$R(\gamma\|\theta)$ is a convex, lower semicontinuous function of $(\gamma, \theta) \in \mathscr{P}(\mathscr{X}) \times \mathscr{P}(\mathscr{X})$. In particular, $R(\gamma\|\theta)$ is a convex lower semicontinuous function of each variable γ and θ separately. In addition, for fixed $\theta \in \mathscr{P}(\mathscr{X})$, $R(\cdot\|\theta)$ is strictly convex on the set $\{\gamma \in \mathscr{P}(\mathscr{X}) : R(\gamma\|\theta) < \infty\}$.*

(c) *For each $\theta \in \mathscr{P}(\mathscr{X})$, $R(\cdot\|\theta)$ has compact level sets. That is, for each $M < \infty$, the set $\{\gamma \in \mathscr{P}(\mathscr{X}) : R(\gamma\|\theta) \le M\}$ is a compact subset of $\mathscr{P}(\mathscr{X})$.*

(d) *Let $\mathscr{X} = \mathbb{R}^d$ and let $\{\gamma_n\}_{n\in\mathbb{N}}$ and $\{\theta_n\}_{n\in\mathbb{N}}$ be sequences in $\mathscr{P}(\mathbb{R}^d)$. Assume that for each $\alpha \in \mathbb{R}^d$,*

$$\sup_{n\in\mathbb{N}} \int_{\mathbb{R}^d} \exp\langle \alpha, y\rangle\, \theta_n(dy) < \infty \text{ and } \Delta \doteq \sup_{n\in\mathbb{N}} R(\gamma_n \| \theta_n) < \infty.$$

Then $\{\gamma_n\}$ is uniformly integrable in the sense that

$$\lim_{C\to\infty} \sup_{n\in\mathbb{N}} \int_{\{y\in\mathbb{R}^d : \|y\|>C\}} \|y\|\, \gamma_n(dy) = 0.$$

Furthermore,

$$\sup_{n\in\mathbb{N}} \int_{\mathbb{R}^d} \|y\|\, \gamma_n(dy) < \infty. \tag{2.6}$$

In particular, $\{\gamma_n\}$ is tight.

(e) *We denote by Π the class of all finite measurable partitions of $\mathscr{X}$. Then for each γ and θ in $\mathscr{P}(\mathscr{X})$*

$$R(\gamma\|\theta) = \sup_{\pi\in\Pi} \sum_{A\in\pi} \gamma(A)\, \log \frac{\gamma(A)}{\theta(A)},$$

where the summand equals 0 if $\gamma(A) = 0$ and equals ∞ if $\gamma(A) > 0$ and $\theta(A) = 0$. In addition, if A is any Borel subset of $\mathscr{X}$, then

$$R(\gamma\|\theta) \geq \gamma(A)\, \log \frac{\gamma(A)}{\theta(A)} - 1.$$

(f) *Let ψ be a measurable mapping from $\mathscr{X}$ to $\mathscr{Y}$ and let Δ_ψ denote the function mapping $\mathscr{P}(\mathscr{X})$ into $\mathscr{P}(\mathscr{Y})$ that is given by $\Delta_\psi \alpha \doteq \alpha \circ \psi^{-1}$. Then for each γ and θ in $\mathscr{P}(\mathscr{X})$,*

$$R(\Delta_\psi \gamma \| \Delta_\psi \theta) \leq R(\gamma\|\theta).$$

If ψ is one-to-one and ψ^{-1} is measurable, then the inequality can be replaced by an equality.

Proof (a) A proof can be found in [97, Appendix C.2].

(b) To prove the first assertion, we use the variational formula stated in part (a). For each fixed $g \in \mathscr{C}_b(\mathscr{X})$, the function mapping

$$(\gamma, \theta) \in \mathscr{P}(\mathscr{X}) \times \mathscr{P}(\mathscr{X}) \mapsto \int_{\mathscr{X}} g\, d\gamma - \log \int_{\mathscr{X}} e^g\, d\theta$$

is convex and continuous. As the supremum over $g \in \mathscr{C}_b(\mathscr{X})$ of such functions, $R(\gamma\|\theta)$ is a convex lower semicontinuous function of $(\gamma, \theta) \in \mathscr{P}(\mathscr{X}) \times \mathscr{P}(\mathscr{X})$. To prove the strict convexity, we use

$$R(\gamma\|\theta) = \int_{\mathscr{X}} \frac{d\gamma}{d\theta} \left(\log \frac{d\gamma}{d\theta} \right) d\theta,$$

which is valid for any $\gamma \in \mathscr{P}(\mathscr{X})$ satisfying $R(\gamma\|\theta) < \infty$. The strict convexity of $R(\cdot\|\theta)$ on the set $\{\gamma \in \mathscr{P}(\mathscr{X}) : R(\gamma\|\theta) < \infty\}$ follows from the strict convexity of $s \log s$ for $s \in [0, \infty)$. This completes the proof of part (b).

(c) We follow the proof of Lemma 5.1 in [88]. Let $\{\gamma_n\}_{n\in\mathbb{N}}$ be any sequence in $\mathscr{P}(\mathscr{X})$ satisfying $\sup_{n\in\mathbb{N}} R(\gamma_n\|\theta) \le M < \infty$. According to the variational formula stated in part (a), for any bounded measurable function ψ mapping $\mathscr{X}$ into $\mathbb{R}$ we have for each $n \in \mathbb{N}$,

$$\int_{\mathscr{X}} \psi \, d\gamma_n - \log \int_{\mathscr{X}} e^{\psi} \, d\theta \le R(\gamma_n\|\theta) \le M.$$

Let $\delta > 0$ and $\varepsilon > 0$ be given. It follows from Prohorov's theorem [Theorem A.4] that a single probability measure θ is tight. This guarantees that there exists a compact set K such that $\theta(K^c) \le \varepsilon$. Substituting into the last display the function ψ that equals 0 on K and equals $\log(1 + 1/\varepsilon)$ on K^c, we have for each $n \in \mathbb{N}$,

$$\begin{aligned}
\gamma_n(K^c) &\le \frac{1}{\log(1+1/\varepsilon)} \left(M + \log \left[\theta(K) + \left(1 + \frac{1}{\varepsilon}\right) \theta(K^c) \right] \right) \\
&= \frac{1}{\log(1+1/\varepsilon)} \left(M + \log \left[1 + \frac{1}{\varepsilon} \theta(K^c) \right] \right) \\
&\le \frac{1}{\log(1+1/\varepsilon)} (M + \log 2).
\end{aligned}$$

Since $\varepsilon > 0$ can be chosen so that $\frac{1}{\log(1+1/\varepsilon)}(M + \log 2) \le \delta$, this formula implies that $\{\gamma_n\}$ is tight. By Prohorov's theorem there exist $\gamma \in \mathscr{P}(\mathscr{X})$ and a subsequence of $n \in \mathbb{N}$ such that $\gamma_n \Rightarrow \gamma$. The lower semicontinuity of $R(\cdot\|\theta)$ yields

$$R(\gamma\|\theta) \le \liminf_{n\to\infty} R(\gamma_n\|\theta) \le M.$$

This completes the proof that $\{\gamma \in \mathscr{P}(\mathscr{X}) : R(\gamma\|\theta) \le M\}$ is compact.

(d) As in the proof of part (c), we again follow the proof of Lemma 5.1 in [88]. Let σ be any positive number. By considering each coordinate direction separately, we obtain from the assumed bound

$$\sup_{n\in\mathbb{N}} \int_{\mathbb{R}^d} \exp\langle \alpha, y\rangle \, \theta_n(dy) < \infty$$

for each $\alpha \in \mathbb{R}^d$ the limit

$$\lim_{C\to\infty} \sup_{n\in\mathbb{N}} \int_{\{y\in\mathbb{R}^d : \|y\| > C\}} e^{\sigma\|y\|} \, \theta_n(dy) = 0. \tag{2.7}$$

Using this together with the bound $\Delta \doteq \sup_{n\in\mathbb{N}} R(\gamma_n\|\theta_n) < \infty$, we will prove that $\{\gamma_n\}$ is uniformly integrable. Since the latter property implies that $\{\gamma_n\}$ is tight and also (2.6), the proof of part (d) will be done.

Since for each $n \in \mathbb{N}$, $R(\gamma_n \| \theta_n)$ is finite, γ_n is absolutely continuous with respect to θ_n. Thus we can consider the Radon–Nikodym derivative

$$f_n \doteq \frac{d\gamma_n}{d\theta_n},$$

in terms of which

$$R(\gamma_n \| \theta_n) = \int_{\mathscr{X}} f_n \log f_n \, d\theta_n. \tag{2.8}$$

We need the inequality

$$ab \le e^{\sigma a} + \frac{1}{\sigma}(b \log b - b + 1), \tag{2.9}$$

valid for $a \ge 0$, $b \ge 0$, and $\sigma \ge 1$. This follows from the fact that

$$\sup_{a \in \mathbb{R}} \left[ab - e^{\sigma a}\right] = \frac{b}{\sigma}\left(\log \frac{b}{\sigma} - 1\right) \le \frac{1}{\sigma}(b \log b - b + 1).$$

Since $b \log b - b + 1 \ge 0$ for all $b \ge 0$, we find that for $\sigma \ge 1$ and $C < \infty$,

$$\begin{aligned}
\sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} \|y\| \, \gamma_n(dy) &= \sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} \|y\| \, f_n(y) \, \theta_n(dy) \\
&\le \sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} e^{\sigma \|y\|} \, \theta_n(dy) \\
&\quad + \frac{1}{\sigma} \sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d} (f_n \log f_n - f_n + 1) \, d\theta_n.
\end{aligned}$$

Equation (2.8) yields

$$\sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d} (f_n \log f_n - f_n + 1) \, d\theta_n = \sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d} f_n \log f_n \, d\theta_n = \sup_{n \in \mathbb{N}} R(\gamma_n \| \theta_n) = \Delta.$$

Combining this with (2.7), we obtain

$$\lim_{C \to \infty} \sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} \|y\| \, \gamma_n(dy) \le \frac{\Delta}{\sigma}.$$

Since $\sigma \ge 1$ can be taken arbitrarily large, we have completed the proof that $\{\gamma_n\}$ is uniformly integrable.

(e) This is proved in Appendix C.

(f) This is proved in Appendix C. □

For many stochastic systems, the mapping that takes the noise process into the system state may be only measurable and not continuous. A basic example is a stochastic differential equation with state-dependent diffusion matrix for which there are existence and uniqueness in the strong sense. For such a process, the mapping from the noise space (Wiener process) into the state space (diffusion process) is Borel measurable but not, in general, continuous. In such circumstances the following lemma, which will guarantee convergence of integrals with respect to bounded and measurable functions when a uniform bound on relative entropies holds, is very convenient. In particular, it will often be used in the study of continuous time processes. The lemma first appeared in [32].

Lemma 2.5 *Let $\mathscr{X}$ be a Polish space, $\theta \in \mathscr{P}(\mathscr{X})$, and let $f : \mathscr{X} \to \mathbb{R}$ be a bounded Borel-measurable function. Consider a sequence $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathscr{P}(\mathscr{X})$ satisfying $\sup_{n\in\mathbb{N}} R(\mu_n \| \theta) \le \alpha < \infty$, and assume that μ_n converges weakly to $\mu \in \mathscr{P}(\mathscr{X})$. Then:*

(a) $\lim_{n\to\infty} \int_{\mathscr{X}} f d\mu_n = \int_{\mathscr{X}} f d\mu$;

(b) *if $\{f_n\}_{n\in\mathbb{N}}$ is a sequence of uniformly bounded functions (i.e., $\sup_{n\in\mathbb{N}} \|f\|_\infty < \infty$) converging θ-a.s. to f, then*

$$\lim_{j\to\infty} \sup_{n\in\mathbb{N}} \int_{\mathscr{X}} |f_j - f| d\mu_n = 0 \text{ and } \lim_{n\to\infty} \int_{\mathscr{X}} f_n d\mu_n = \int_{\mathscr{X}} f d\mu.$$

Proof As a first step we verify that the limit measure μ is absolutely continuous with respect to θ. Indeed, by the weak convergence of μ_n to μ and the lower semicontinuity of $R(\cdot \| \theta)$ [part (b) of Lemma 2.4],

$$R(\mu \| \theta) \le \liminf_{n\to\infty} R(\mu_n \| \theta) \le \alpha < \infty.$$

From the definition of relative entropy, this implies that μ is absolutely continuous with respect to θ. Theorem E.4 guarantees that there is a sequence $\{\bar{f}_j\}_{j\in\mathbb{N}}$ of uniformly bounded and continuous functions such that $\lim_{j\to\infty} \bar{f}_j = f$ θ-a.s. Since $\mu \ll \theta$, the limit also holds μ-a.e. Thus by the dominated convergence theorem, $\int_{\mathscr{X}} \bar{f}_j d\mu$ converges to $\int_{\mathscr{X}} f d\mu$. For each fixed $j \in \mathbb{N}$, $\int_{\mathscr{X}} \bar{f}_j d\mu_n$ converges to $\int_{\mathscr{X}} \bar{f}_j d\mu$ because of the weak convergence of μ_n to μ. Hence to prove part (a) of the lemma it remains only to verify

$$\lim_{j\to\infty} \sup_{n\in\mathbb{N}} \int_{\mathscr{X}} \left| \bar{f}_j - f \right| d\mu_n = 0. \tag{2.10}$$

In the proof of (2.10) we will not use the continuity of $\bar{f}_j$, only the θ-a.s. convergence to f, and so the first statement in part (b) will also follow. Fix $\varepsilon > 0$. Let $M < \infty$ be such that $\|f\|_\infty \le M$ and $\sup_{j\in\mathbb{N}} \|\bar{f}_j\|_\infty \le M$. Then

$$\begin{aligned}\int_{\mathscr{X}} |\bar{f}_j - f| \, d\mu_n &= \int_{\{|\bar{f}_j - f| > \varepsilon\}} |\bar{f}_j - f| \, d\mu_n + \int_{\{|\bar{f}_j - f| \le \varepsilon\}} |\bar{f}_j - f| \, d\mu_n \\ &\le \int_{\{|\bar{f}_j - f| > \varepsilon\}} |\bar{f}_j - f| \, d\mu_n + \varepsilon \\ &\le 2M\mu_n\{|\bar{f}_j - f| > \varepsilon\} + \varepsilon.\end{aligned}$$

Using the inequality (2.9) with $\sigma = c$ and $a = 1$ we have that for every $c \in [1, \infty)$,

$$\begin{aligned}\mu_n\{|\bar{f}_j - f| > \varepsilon\} &= \int_{\{|\bar{f}_j - f| > \varepsilon\}} \frac{d\mu_n}{d\theta} d\theta \\ &\le e^c \theta\{|\bar{f}_j - f| > \varepsilon\} + \frac{1}{c} R(\mu_n \| \theta).\end{aligned}$$

Thus

$$\sup_{n \in \mathbb{N}} \int_{\mathscr{X}} |\bar{f}_j - f| \, d\mu_n \le 2Me^c \theta\{|\bar{f}_j - f| > \varepsilon\} + \frac{2M\alpha}{c} + \varepsilon.$$

The convergence in (2.10) now follows on sending $j \to \infty$ followed by $c \to \infty$ and $\varepsilon \to 0$ in the last display.

To prove the second claim in part (b), observe that

$$\int_{\mathscr{X}} f_n d\mu_n = \int_{\mathscr{X}} f d\mu_n + \left\{ \int_{\mathscr{X}} f_n d\mu_n - \int_{\mathscr{X}} f d\mu_n \right\}.$$

From part (a), the first term on the right side converges to $\int_{\mathscr{X}} f d\mu$, while the second term converges to 0 from the first statement in part (b). □

Let $\mathscr{X}$ and $\mathscr{Y}$ be Polish spaces, $\sigma(dy|x)$ a stochastic kernel on $\mathscr{Y}$ given $\mathscr{X}$, and θ a probability measure on $\mathscr{X}$. Then we define $\theta \otimes \sigma$ to be the unique probability measure on $(\mathscr{X} \times \mathscr{Y}, \mathscr{B}(\mathscr{X} \times \mathscr{Y}))$ with the property that for Borel subsets A of $\mathscr{X}$ and B of $\mathscr{Y}$,

$$\theta \otimes \sigma(A \times B) \doteq \int_{A \times B} \theta(dx) \, \sigma(dy|x) = \int_A \sigma(B|x) \, \theta(dx).$$

This formula is summarized by the notation $\theta \otimes \sigma(dx \times dy) = \theta(dx) \otimes \sigma(dy|x)$. If, for example, $\sigma(dy|x)$ is independent of x, thus defining a probability measure $\sigma(dy)$ on $\mathscr{Y}$, then $\theta \otimes \sigma$ equals the product measure $\theta \times \sigma$ on $\mathscr{X} \times \mathscr{Y}$. Conversely, given a probability measure α on $\mathscr{X} \times \mathscr{Y}$ and denoting by $[\alpha]_1$ the first marginal of α, there is a stochastic kernel $[\alpha]_{2|1}(dy|x)$ on $\mathscr{Y}$ given $\mathscr{X}$ such that $\alpha = [\alpha]_1 \otimes [\alpha]_{2|1}$. This is a consequence of the existence of regular conditional probabilities for random variables that take values in a Polish space [Theorem B.2].

Using the Laplace formulation, the variational representation (2.1) involving $R(\gamma \| \theta)$ becomes the starting point for large deviations and related analysis for systems driven by the "base" measure θ. When θ has a nice structure (for example if θ is

product measure or a Markovian measure), this property can be exploited to rewrite (2.1) in a form that is much more convenient for analysis. This is done in the discrete time setting using the chain rule, which we now describe. The analogous rewriting of relative entropy is more involved but still possible in continuous time, and will be discussed in detail in Chap. 8.

Theorem 2.6 (CHAIN RULE) *Let $\mathscr{X}$ and $\mathscr{Y}$ be Polish spaces and α and β probability measures on $\mathscr{X} \times \mathscr{Y}$. We denote by $[\alpha]_1$ and $[\beta]_1$ the first marginals of α and β and by $[\alpha]_{2|1}(dy|x)$ and $[\beta]_{2|1}(dy|x)$ the stochastic kernels on $\mathscr{Y}$ given $\mathscr{X}$, for which we have the decompositions*

$$\alpha(dx \times dy) = [\alpha]_1(dx) \otimes [\alpha]_{2|1}(dy|x)$$
$$\beta(dx \times dy) = [\beta]_1(dx) \otimes [\beta]_{2|1}(dy|x).$$

Then the function mapping $x \in \mathscr{X} \mapsto R([\alpha]_{2|1}(\cdot|x) \| [\beta]_{2|1}(\cdot|x))$ is measurable and

$$R(\alpha\|\beta) = R([\alpha]_1\|[\beta]_1) + \int_{\mathscr{X}} R([\alpha]_{2|1}(\cdot|x)\|[\beta]_{2|1}(\cdot|x))\,[\alpha]_1(dx).$$

Before giving the proof of Theorem 2.6 we present two corollaries of this result.

Corollary 2.7 *Let $\mathscr{X}$ and $\mathscr{Y}$ be Polish spaces, $\sigma(dy|x)$ and $\tau(dy|x)$ stochastic kernels on $\mathscr{Y}$ given $\mathscr{X}$, and θ a probability measure on $\mathscr{X}$. Then the function mapping $x \in \mathscr{X} \mapsto R(\sigma(\cdot|x)\|\tau(\cdot|x))$ is measurable, and*

$$\int_{\mathscr{X}} R(\sigma(\cdot|x)\|\tau(\cdot|x))\ \theta(dx) = R(\theta \otimes \sigma\|\theta \otimes \tau). \tag{2.11}$$

Proof The measurability of the function mapping $x \in \mathscr{X} \mapsto R(\sigma(\cdot|x)\|\tau(\cdot|x))$ is shown in the proof of Theorem 2.6. In order to prove formula (2.11), we make the following identifications in Theorem 2.6: $\alpha(dx \times dy) \doteq \theta(dx) \otimes \sigma(dy|x)$ and $\beta(dx \times dy) \doteq \theta(dx) \otimes \tau(dy|x)$. Then the first marginals $[\alpha]_1(dx)$ and $[\beta]_1(dx)$ both equal $\theta(dx)$, and Theorem 2.6 implies that

$$\begin{aligned} R(\theta \otimes \sigma\|\theta \otimes \tau) &= R(\theta\|\theta) + \int_{\mathscr{X}} R(\sigma(\cdot|x)\|\tau(\cdot|x))\,\theta(dx) \\ &= \int_{\mathscr{X}} R(\sigma(\cdot|x)\|\tau(\cdot|x))\,\theta(dx). \end{aligned}$$

This is formula (2.11). □

For use elsewhere in the text, we record another corollary of Theorem 2.6 that applies to product measures.

Corollary 2.8 *Let $\mathscr{X}$ and $\mathscr{Y}$ be Polish spaces, γ and θ probability measures on $\mathscr{X}$, and λ and μ probability measures on $\mathscr{Y}$. Then*

$$R(\gamma \times \lambda \| \theta \times \mu) = R(\gamma \| \theta) + R(\lambda \| \mu).$$

Proof (*of Theorem* 2.6) The stochastic kernels $[\alpha]_{2|1}(dy|x)$ and $[\beta]_{2|1}(dy|x)$ are measurable functions mapping $\mathscr{X}$ into $\mathscr{P}(\mathscr{Y})$ [Theorem B.4]. Since $R(\cdot\|\cdot)$ is lower semicontinuous on $\mathscr{P}(\mathscr{Y}) \times \mathscr{P}(\mathscr{Y})$ [Lemma 2.1(b)], the measurability of the function mapping $x \in \mathscr{X} \mapsto R\big([\alpha]_{2|1}(\cdot|x)\|[\beta]_{2|1}(\cdot|x)\big)$ follows. We now prove that

$$R(\alpha\|\beta) = R([\alpha]_1\|[\beta]_1) + \int_{\mathscr{X}} R\big([\alpha]_{2|1}(\cdot|x)\|[\beta]_{2|1}(\cdot|x)\big)\,[\alpha]_1(dx), \tag{2.12}$$

assuming first that the right-hand side is finite. Under this assumption, $[\alpha]_1 \ll [\beta]_1$ and there exists an $[\alpha]_1$-null set Γ in $\mathscr{X}$ such that $R\big([\alpha]_{2|1}(\cdot|x)\|[\beta]_{2|1}(\cdot|x)\big)$ is finite for $x \in \Gamma^c$. Hence for $x \in \Gamma^c$, we have that $[\alpha]_{2|1}(\cdot|x) \ll [\beta]_{2|1}(\cdot|x)$ as measures on $\mathscr{Y}$. By redefining $[\alpha]_{2|1}(\cdot|x)$ for x in the null set Γ, we can ensure that $[\alpha]_{2|1}(\cdot|x) \ll [\beta]_{2|1}(\cdot|x)$ for all $x \in \mathscr{X}$. Let

$$\psi(x) \doteq \frac{d[\alpha]_1}{d[\beta]_1}(x).$$

Theorem A.5.7 in [97] guarantees that there exists a version of the Radon–Nikodym derivative

$$\zeta(x, y) \doteq \frac{d[\alpha]_{2|1}(\cdot|x)}{d[\beta]_{2|1}(\cdot|x)}(y)$$

that is a nonnegative measurable function on $\mathscr{X} \times \mathscr{Y}$. For any Borel subsets A of $\mathscr{X}$ and B of $\mathscr{Y}$,

$$\begin{aligned}
\alpha(A \times B) &= \int_A [\alpha]_{2|1}(B|x)[\alpha]_1(dx) \\
&= \int_A \left(\int_B \zeta(x, y)\,[\beta]_{2|1}(dy|x) \right) \psi(x)\,[\beta]_1(dx) \\
&= \int_{A\times B} \psi(x)\,\zeta(x, y)\,\beta(dx \times dy).
\end{aligned}$$

This implies that $\alpha \ll \beta$ and that for β-a.e. $(x, y) \in \mathscr{X} \times \mathscr{Y}$,

$$\frac{d\alpha}{d\beta}(x, y) = \psi(x)\,\zeta(x, y).$$

As a consequence,

$$
\begin{aligned}
&R([\alpha]_1 \| [\beta]_1) + \int_{\mathscr{X}} R([\alpha]_{2|1}(\cdot|x) \| [\beta]_{2|1}(\cdot|x))\, [\alpha]_1(dx) \\
&= \int_{\mathscr{X}} \log \psi(x)\, [\alpha]_1(dx) + \int_{\mathscr{X}} \left(\int_{\mathscr{Y}} \log \zeta(x, y)\, [\alpha]_{2|1}(dy|x) \right) [\alpha]_1(dx) \\
&= \int_{\mathscr{X} \times \mathscr{Y}} \log \psi(x)\, \alpha(dx \times dy) + \int_{\mathscr{X} \times \mathscr{Y}} \log \zeta(x, y)\, [\alpha]_1(dx) \otimes [\alpha]_{2|1}(dy|x) \\
&= \int_{\mathscr{X} \times \mathscr{Y}} \log[\psi(x)\, \zeta(x, y)]\, \alpha(dx \times dy) \\
&= R(\alpha \| \beta).
\end{aligned}
$$

This is formula (2.12).

We now prove (2.12) under the assumption that the left-hand side is finite. Under this assumption $\alpha \ll \beta$, and so for $(x, y) \in \mathscr{X} \times \mathscr{Y}$ we can define

$$
\varphi(x, y) \doteq \frac{d\alpha}{d\beta}(x, y).
$$

Since $\alpha \ll \beta$ we have $[\alpha]_1 \ll [\beta]_1$, and also $[\beta]_1$ a.s. $\psi(x) \doteq \frac{d[\alpha]_1}{d[\beta]_1}(x)$ is equal to $\int_{\mathscr{Y}} \varphi(x, y)[\beta]_{2|1}(dy|x)$,

For any Borel subsets A of $\mathscr{X}$ and B of $\mathscr{Y}$,

$$
\begin{aligned}
\int_A [\alpha]_{2|1}(B|x)\, \psi(x)\, [\beta]_1(dx) &= \int_A [\alpha]_{2|1}(B|x)\, [\alpha]_1(dx) \\
&= \alpha(A \times B) \\
&= \int_{A \times B} \varphi(x, y)\, \beta(dx \times dy) \\
&= \int_A \left(\int_B \varphi(x, y)\, [\beta]_{2|1}(dy|x) \right) [\beta]_1(dx).
\end{aligned}
$$

This implies that there exists a $[\beta]_1$-null set Γ such that for all $x \in \Gamma^c$,

$$
\psi(x)\, [\alpha]_{2|1}(B|x) = \int_B \varphi(x, y)\, [\beta]_{2|1}(dy|x).
$$

In fact, using the separability of $\mathscr{X}$, we can find a $[\beta]_1$-null set (denoted again by Γ) such that for $x \in \Gamma^c$, the above equality holds for all Borel subsets B of $\mathscr{Y}$.

Thus for all $x \in \Gamma^c \cap \{\psi > 0\}$, $[\alpha]_{2|1}(\cdot|x) \ll [\beta]_{2|1}(\cdot|x)$, and for such x and $[\beta]_{2|1}(\cdot|x)$-a.e. $y \in \mathscr{Y}$,

$$
\zeta(x, y) \doteq \frac{d[\alpha]_{2|1}(\cdot|x)}{d[\beta]_{2|1}(\cdot|x)}(y) \text{ equals } \frac{\varphi(x, y)}{\psi(x)}.
$$

In other words, for all $x \in \Gamma^c \cap \{\psi > 0\}$, the various Radon–Nikodym derivatives are related by

$$\varphi(x, y) = \psi(x)\, \zeta(x, y), \ [\beta]_{2|1}(\cdot|x)\text{-a.e.}$$

We have $[\alpha]_1\{\psi > 0\} = 1$, and since $[\alpha]_1 \ll [\beta]_1$ and $[\beta]_1\{\Gamma^c\} = 1$, we also have $[\alpha]_1\{\Gamma^c\} = 1$. It now follows that

$$\begin{aligned}
&R(\alpha\|\beta) \\
&\quad = \int_{\mathscr{X}\times\mathscr{Y}} \log \varphi(x, y)\, \alpha(dx \times dy) \\
&\quad = \int_{\mathscr{X}\times\mathscr{Y}} \log \varphi(x, y)\, [\alpha]_1(dx) \otimes [\alpha]_{2|1}(dy|x) \\
&\quad = \int_{(\Gamma^c\cap\{\psi>0\})\times\mathscr{Y}} \log[\psi(x)\, \zeta(x, y)]\, [\alpha]_1(dx) \otimes [\alpha]_{2|1}(dy|x) \\
&\quad = \int_{\Gamma^c\cap\{\psi>0\}} \log \psi(x)\, [\alpha]_1(dx) \\
&\qquad + \int_{\Gamma^c\cap\{\psi>0\}} \left(\int_{\mathscr{Y}} \log \zeta(x, y)\, [\alpha]_{2|1}(dy|x) \right) [\alpha]_1(dx) \\
&\quad = R([\alpha]_1\|[\beta]_1) + \int_{\mathscr{X}} R([\alpha]_{2|1}(\cdot|x)\|[\beta]_{2|1}(\cdot|x))\, [\alpha]_1(dx),
\end{aligned}$$

which again is Eq. (2.12). This finishes the proof of the theorem. □

Throughout the remainder of the book we will make frequent use of relative entropy. It will sometimes be necessary to work simultaneously with the relative entropy associated with a number of different spaces. We will abuse notation and simply write $R(\cdot\|\cdot)$ to denote the relative entropy in all the different cases. The particular space involved in each circumstance will be obvious.

2.2 Tightness of Probability Measures

Many proofs in this book are based on the asymptotic analysis of variational representations for exponential integrals. We will evaluate limits of expected values for continuous functions of various types of random objects, or equivalently limits of integrals with respect to the distributions of these random objects. The notion of weak convergence, which was introduced in the last section, is ideally suited to such analysis, in part because one can often establish precompactness of the collection of distributions under mild conditions. In this section we review some characterizations of precompactness that will be useful later in the book.

The two types of random variables most commonly encountered in this book will be random processes and random probability or subprobability measures. Let $\mathscr{E}$ be a Polish space. Random processes will usually take values in a space of the form

$\mathscr{D}([0,T]:\mathscr{E})$, the space of functions from $[0,T]$ to $\mathscr{E}$ that are right continuous and with limits from the left at each $t \in (0,T]$. The space $\mathscr{D}([0,T]:\mathscr{E})$ is equipped with the usual Skorohod topology, which can be metrized so that it is a Polish space [24, Chap. 3, Sect. 12]. We will also encounter processes taking values in the subset $\mathscr{C}([0,T]:\mathscr{E})$ of continuous functions. When restricted to $\mathscr{C}([0,T]:\mathscr{E})$, convergence in the Skorohod topology is equivalent to convergence with respect to the supremum metric typically used on $\mathscr{C}([0,T]:\mathscr{E})$.

Let A be an index set and let $\{\lambda_a\}_{a\in A} \subset \mathscr{P}(\mathscr{X})$, where $\mathscr{X}$ is a Polish space. Recall that the collection $\{\lambda_a\}$ is tight if for each $\varepsilon > 0$ there is a compact set $K_\varepsilon \subset \mathscr{X}$ such that $\inf_{a\in A} \lambda_a(K_\varepsilon) \geq 1-\varepsilon$. If random variables $\{X_a\}_{a\in A}$ have the distributions $\{\lambda_a\}_{a\in A}$, we say that $\{X_a\}$ is **tight** if $\{\lambda_a\}$ is tight. According to Prohorov's theorem, $\{\lambda_a\}$ is **precompact** [i.e., has compact closure] in the topology of weak convergence if and only if it is tight.

The notion of a tightness function will be useful. A measurable function $g : \mathscr{X} \to [0,\infty]$ is called a **tightness function** if it has **precompact level sets**: for every $M \in [0,\infty)$, the set $\{x \in \mathscr{X} : g(x) \leq M\}$ has compact closure. Thus rate functions are tightness functions. We have the following elementary result.

Lemma 2.9 *A collection $\{\lambda_a\}_{a\in A} \subset \mathscr{P}(\mathscr{X})$ is tight if and only if there is a tightness function g such that $\sup_{a\in A} \int_{\mathscr{X}} g(x)\lambda_a(dx) < \infty$. If $\{\lambda_a\}$ is tight, then one can assume without loss that g is lower semicontinuous.*

Proof The "if" part follows from Chebyshev's inequality. Let $\bar{K}_M$ be the closure of $\{x \in \mathscr{X} : g(x) \leq M\}$. Since g is a tightness function, $\bar{K}_M$ is compact. If $B \doteq \sup_{a\in A} \int_{\mathscr{X}} g(x)\lambda_a(dx)$, then $\lambda_a(\bar{K}_M^c) \leq B/M$. Thus K_ε in the definition of tightness of $\{\lambda_a\}$ can be taken to be $\bar{K}_{B/\varepsilon}$. To argue the reverse direction, let K_ε satisfy the requirement in the definition of tightness for $\{\lambda_a\}_{a\in A}$. We can assume without loss that $K_{2^{-i}}$ is increasing, since if this is not true, we can always use the compact set $\cup_{j=1}^{i} K_{2^{-j}}$ in place of $K_{2^{-i}}$. Let

$$g(x) \doteq \sum_{i=1}^{\infty} 1_{K_{2^{-i}}^c}(x). \tag{2.13}$$

Then $g(x) \leq M$ implies $x \notin K_{2^{-i}}^c$ when $i > M$, and thus $\{x \in \mathscr{X} : g(x) \leq M\} \subset K_{2^{-M-1}}$. Hence the level sets of g have compact closure. Since for all $a \in A$,

$$\int_{\mathscr{X}} g(x)\lambda_a(dx) = \sum_{i=1}^{\infty} \lambda_a(K_{2^{-i}}^c) \leq \sum_{i=1}^{\infty} 2^{-i} < \infty,$$

g serves as a tightness function with the desired uniform bound on integrals. To prove the last claim, note that since K_ε^c is open, $x \mapsto 1_{K_\varepsilon^c}(x)$ is lower semicontinuous, and hence by Fatou's lemma the same is true for g. □

The next result shows that tightness functions have a useful "bootstrap" property.

Theorem 2.10 *Let g be a tightness function on $\mathscr{X}$. Define $G : \mathscr{P}(\mathscr{X}) \to [0, \infty]$ by*

$$G(\mu) \doteq \int_{\mathscr{X}} g(x)\mu(dx).$$

Then G is a tightness function on $\mathscr{P}(\mathscr{X})$.

Proof The preceding lemma shows that for every $M \in [0, \infty)$, the set

$$\{\mu \in \mathscr{P}(\mathscr{X}) : G(\mu) \leq M\}$$

is tight. By Prohorov's theorem, the same set is precompact, and thus G is a tightness function. □

The next result shows that a every member of a collection of random probability measures is tight (as random variables!) if and only if their "means" are tight as deterministic probability measures.

Theorem 2.11 *Let $\{\Lambda_a\}_{a\in A}$ be random variables taking values in $\mathscr{P}(\mathscr{X})$ (i.e., random probability measures), and let $\lambda_a = E\Lambda_a$. Then $\{\Lambda_a\}_{a\in A}$ is tight if and only if $\{\lambda_a\}_{a\in A}$ is tight.*

Proof Fix $\varepsilon > 0$. Let $\eta_a \in \mathscr{P}(\mathscr{P}(\mathscr{X}))$ denote the distribution of Λ_a and let $\varepsilon > 0$ be given. Assuming that the random measures $\{\Lambda_a\}$ are tight, there is a compact set $K \subset \mathscr{P}(\mathscr{X})$ such that $\eta_a(K^c) \leq \varepsilon$ for all $a \in A$. Since K is compact, by Prohorov's theorem there is a compact $K_1 \subset \mathscr{X}$ such that $\lambda \in K$ implies $\lambda(K_1^c) \leq \varepsilon$. Therefore

$$\begin{aligned}\lambda_a(K_1^c) &= \int_{\mathscr{P}(\mathscr{X})} \lambda(K_1^c)\eta_a(d\lambda) \\ &= \int_K \lambda(K_1^c)\eta_a(d\lambda) + \int_{K^c} \lambda(K_1^c)\eta_a(d\lambda) \\ &\leq 2\varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we have that $\{\lambda_a\}_{a\in A}$ is tight. To prove the reverse direction, it suffices from Lemma 2.9 to find a tightness function $\bar{G} : \mathscr{P}(\mathscr{X}) \to [0, \infty]$ such that

$$\sup_{a\in A} E\bar{G}(\Lambda_a) < \infty. \tag{2.14}$$

Since $\{\lambda_a\}_{a\in A}$ is tight by Lemma 2.9, there is a tightness function $\bar{g} : \mathscr{X} \to [0, \infty]$ such that

$$\sup_{a\in A} \int_{\mathscr{X}} \bar{g}(x)\lambda_a(dx) < \infty.$$

By Theorem 2.10, $\bar{G}(\rho) \doteq \int_{\mathscr{X}} \bar{g}(x)\rho(dx)$ is a tightness function on $\mathscr{P}(\mathscr{X})$. Also,

$$\sup_{a\in A} E\bar{G}(\Lambda_a) = \sup_{a\in A} E\int_{\mathscr{X}} \bar{g}(x)\Lambda_a(dx) = \sup_{a\in A}\int_{\mathscr{X}} \bar{g}(x)\lambda_a(dx) < \infty.$$

This proves (2.14), and hence tightness of $\{\Lambda_a\}_{a\in A}$ follows. □

2.3 Notes

Much of the material in this chapter is taken from [97]. Exceptions include parts of Proposition 2.3 and Lemma 2.5, which are from [32]. Relative entropy originated in information theory, and it is also heavily used in statistics and computer science. In those disciplines it often goes by the name Kullback–Leibler divergence, and indeed, early references to the topic are Kullback and Leibler [177] and Kullback [176]. An introduction to its properties in a non-measure-theoretic setting can be found in [67]. Although it is not a metric owing to the lack of symmetry, in many uses relative entropy is treated as though it were a metric, and in recent years there has been interest in its relation to genuine metrics on the space of probability measures [240]. The Donsker–Varadhan variational formula and its dual (sometimes called the Gibbs variational formula) can be considered an infinite dimensional analogue of the pairing of convex functions through Legendre–Fenchel duality. Although we do not appeal significantly to infinite dimensional convex analysis in this book, a good reference for this topic is the book [13].

Chapter 3
Examples of Representations and Their Application

Our approach to the study of large deviations is based on convenient variational representations for expected values of nonnegative functionals. In this chapter we give three examples of such representations and show how they allow easy proofs of some classical results.

In Sect. 3.1 we present a representation for stochastic processes in discrete time. To illustrate the main idea we consider the simple setting in which the stochastic process is an iid sequence of random variables [Proposition 3.1]. We then show how this representation can be used to prove Sanov's theorem and Cramér's theorem. Analogous representations for more general noise models will be used many times in later chapters. In Sect. 3.2 we state a variational representation for functionals of a k-dimensional Brownian motion [Theorem 3.14]. This result will be generalized and proved in the setting of an infinite dimensional Brownian motion in Chap. 8, and we apply it here to give an elementary proof of the large deviation principle for small noise diffusions. Section 3.3 states a variational representation for functionals of a standard Poisson process [Theorem 3.23]. This result will also be extended in Chap. 8 to the setting of Poisson random measures with points in an arbitrary locally compact Polish space. As an application of Theorem 3.23 we prove the large deviation principle for stochastic differential equations driven by Poisson processes.

3.1 Representation for an IID Sequence

Owing to the role it plays in the representations, we sometimes refer to the measure appearing in the second position in relative entropy, i.e., θ in $R(\mu \| \theta)$, as the "base" measure. The starting point of all large deviation results in the book is the relative entropy representation in part (a) of Proposition 2.2. When the base measure is structured, for example when θ is a product measure or a Markov measure, a more useful, control-theoretic, representation can be found in terms of the component

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_3

measures that make up θ. Here is an example. Suppose that (X_1, X_2) is an $(S_1 \times S_2)$-valued random variable with joint distribution $\theta(dx_1 \times dx_2) = \theta_1(dx_1)\theta_2(dx_2)$. Then the variational formula (2.1) says that if $G \in \mathscr{M}_b(S_1 \times S_2)$, then

$$-\log Ee^{-G(X_1,X_2)} = \inf_{\mu \in \mathscr{P}(S_1 \times S_2)} \left[\int_{S_1 \times S_2} G d\mu + R(\mu \| \theta) \right].$$

One can always disintegrate μ in the form

$$\mu(dx_1 \times dx_2) = [\mu]_1(dx_1)[\mu]_{2|1}(dx_2|x_1),$$

where $[\mu]_1$ is the marginal on S_1 and $[\mu]_{2|1}$ is the conditional distribution on S_2 given x_1. Suppose that $(\bar{X}_1, \bar{X}_2)$ is distributed according to μ, $\bar{\mu}_1(\cdot) = [\mu]_1(\cdot)$ and $\bar{\mu}_2(\cdot) = [\mu]_{2|1}(\cdot \,|\bar{X}_1)$ (and note that $\bar{\mu}_2$ is a *random* measure). It follows from the chain rule [Theorem 2.6] that

$$\begin{aligned} R(\mu \| \theta) &= R([\mu]_1 \| \theta_1) + \int_{S_1} R([\mu]_{2|1}(\cdot|x_1) \| \theta_2(\cdot))[\mu]_1(dx_1) \\ &= E\left[R(\bar{\mu}_1 \| \theta_1) + R(\bar{\mu}_2 \| \theta_2)\right]. \end{aligned}$$

Here we have used that $\bar{X}_1$ has distribution $[\mu]_1$ to account for integration with respect to this measure. Then we can rewrite the representation as

$$-\log Ee^{-G(X_1,X_2)} = \inf_{\mu \in \mathscr{P}(S_1 \times S_2)} E\left[G(\bar{X}_1, \bar{X}_2) + \sum_{i=1}^{2} R(\bar{\mu}_i \| \theta_i) \right]. \tag{3.1}$$

There is an obvious extension of (3.1) to any finite collection of independent random variables. The extension for the special case in which the random variables are iid is as follows. Let S^n denote the product space of n copies of S.

Proposition 3.1 *Let $\{X_i\}_{i \in \mathbb{N}}$ be iid S-valued random variables with distribution θ and let $n \in \mathbb{N}$. If $G \in \mathscr{M}_b(S^n)$, then*

$$-\frac{1}{n}\log Ee^{-nG(X_1,\ldots,X_n)} = \inf E\left[G(\bar{X}_1^n, \ldots, \bar{X}_n^n) + \frac{1}{n}\sum_{i=1}^{n} R\left(\bar{\mu}_i^n \| \theta\right) \right], \tag{3.2}$$

with the infimum over all collections of random probability measures $\left\{\bar{\mu}_i^n\right\}_{i \in \{1,\ldots,n\}}$ that satisfy the following two conditions:

1. *$\bar{\mu}_i^n$ is measurable with respect to the σ-algebra $\mathscr{F}_{i-1}^n$, where $\mathscr{F}_0^n = \{\emptyset, \Omega\}$ and for $i \in \{1, \ldots, n\}$, $\mathscr{F}_i^n = \sigma\{\bar{X}_1^n, \ldots, \bar{X}_i^n\}$;*
2. *the conditional distribution of $\bar{X}_i^n$, given $\mathscr{F}_{i-1}^n$, is $\bar{\mu}_i^n$.*

Given any measure $\mu \in \mathscr{P}(S^n)$, if $\{\bar{X}_i^n\}_{i=1,\ldots,n}$ has distribution μ, then $\bar{\mu}_i^n$ in the statement of the proposition would equal $[\mu]_{i|1,\ldots,i-1}(\cdot|\bar{X}_1^n, \ldots, \bar{X}_{i-1}^n)$. On the other

hand, given $\{\bar{X}_i^n\}$ and $\{\bar{\mu}_i^n\}$ as in the statement of the proposition, one can identify a $\mu \in \mathscr{P}(S^n)$ that corresponds to these conditional distributions. We consider $\{\bar{X}_i^n\}_{i=1,\dots,n}$ to be a *controlled* version of the original sequence $\{X_i\}_{i=1,\dots,n}$, with control $\bar{\mu}_i^n$ selecting the (conditional) distribution of $\bar{X}_i^n$.

Notational convention. Throughout, we will use overbars to indicate the controlled analogue of any uncontrolled process.

3.1.1 Sanov's and Cramér's Theorems

First we recall the statement of the Glivenko–Cantelli lemma. The space of probability measures on S is denoted by $\mathscr{P}(S)$ and is equipped with the weak topology (see Appendix A).

Lemma 3.2 (GLIVENKO–CANTELLI LEMMA) *Let $\{X_i\}_{i\in\mathbb{N}}$ be iid S-valued random variables with distribution γ, and let L^n be the empirical measure of the first n variables:*

$$L^n(dx) \doteq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(dx).$$

Then with probability one (w.p.1), L^n converges to γ.

The proof is a special case of the arguments we will use for Sanov's theorem, and in particular, the result follows from Lemmas 3.4 and 3.5. Sanov's theorem itself is the large deviation refinement of this law of large numbers (LLN) result.

Theorem 3.3 (SANOV'S THEOREM) *Let $\{X_i\}_{i\in\mathbb{N}}$ be iid S-valued random variables with distribution γ. Then $\{L^n\}_{n\in\mathbb{N}}$ satisfies the LDP on $\mathscr{P}(S)$ with rate function $I(\mu) = R(\mu\,\|\gamma)$.*

By Theorem 1.8, to prove Theorem 3.3 it is enough to show that

$$\lim_{n\to\infty} -\frac{1}{n}\log E\exp\{-nF(L^n)\} = \inf_{\mu\in\mathscr{P}(S)}\left[F(\mu) + R(\mu\,\|\gamma)\right]$$

for every $F \in C_b(\mathscr{P}(S))$. The proof will use the control representation in Proposition 3.1 and will be completed in two steps. First, we will show that the left side in the last display is bounded below by the right side (which gives the Laplace upper bound), and then we will prove the reverse inequality (Laplace lower bound). The first inequality is proved in Sect. 3.1.3, while the second is proved in Sect. 3.1.4.

Taking $G(x_1,\dots,x_n) = F\left(\sum_{i=1}^n \delta_{x_i}(dx)/n\right)$ in the representation (3.2) gives

$$-\frac{1}{n}\log E\exp\{-nF(L^n)\} = \inf_{\{\bar{\mu}_i^n\}} E\left[F\left(\bar{L}^n\right) + \frac{1}{n}\sum_{i=1}^{n} R\left(\bar{\mu}_i^n\,\|\gamma\right)\right], \tag{3.3}$$

where $\bar{L}^n = \frac{1}{n}\sum_{i=1}^n \delta_{\bar{X}_i}$. Thus in order to prove Theorem 3.3, we need to show that

$$\inf_{\{\bar{\mu}_i^n\}} E\left[F\left(\bar{L}^n\right) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n \,\|\gamma\right)\right] \to \inf_{\mu\in\mathscr{P}(S)}\left[F(\mu) + R\left(\mu\,\|\gamma\right)\right].$$

Since F is bounded, the infimum in the representation is always bounded above by $\|F\|_\infty \doteq \sup_{x\in S}|F(x)| < \infty$. It follows that in the infimum in (3.3) we can always restrict to control sequences $\{\bar{\mu}_i^n\}_{i=1,\ldots,n}$ for which

$$\sup_{n\in\mathbb{N}} E\left[\frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n \,\|\gamma\right)\right] \le 2\|F\|_\infty + 1. \tag{3.4}$$

3.1.2 Tightness and Weak Convergence

The bound (3.4) on relative entropy costs is all that is *available*, but also all that is *needed*, to prove tightness.

Lemma 3.4 *Consider any collection of controls $\{\bar{\mu}_i^n, i = 1,\ldots,n\}_{n\in\mathbb{N}}$ for which (3.4) is satisfied, and let $\hat{\mu}^n = \frac{1}{n}\sum_{i=1}^n \bar{\mu}_i^n$. Then $\{(\bar{L}^n, \hat{\mu}^n)\}_{n\in\mathbb{N}}$ is tight.*

Proof By the convexity of relative entropy and Jensen's inequality,

$$2\|F\|_\infty + 1 \ge E\left[\frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n \,\|\gamma\right)\right] \ge E\left[R\left(\hat{\mu}^n \,\|\gamma\right)\right].$$

Since $\mu \mapsto R\left(\mu\,\|\gamma\right)$ has compact level sets, it is a tightness function, and so the bound in the last display along with Lemmas 2.9 and 2.11 shows that both $\left\{\hat{\mu}^n\right\}_{n\in\mathbb{N}}$ and $\left\{E\hat{\mu}^n\right\}_{n\in\mathbb{N}}$ are tight. Since $\bar{\mu}_i^n$ is the conditional distribution used to select $\bar{X}_i^n$, it follows that for every bounded measurable function f,

$$\begin{aligned} E\int_S f(x)\bar{L}^n(dx) &= E\left[\frac{1}{n}\sum_{i=1}^n f(\bar{X}_i^n)\right] \\ &= E\left[\frac{1}{n}\sum_{i=1}^n \int_S f(x)\bar{\mu}_i^n(dx)\right] \\ &= E\int_S f(x)\hat{\mu}^n(dx). \end{aligned}$$

Thus $E\bar{L}^n = E\hat{\mu}^n$, and so $\{\bar{L}^n\}$ and hence $\{(\bar{L}^n, \hat{\mu}^n)\}$ are tight. Here once more we have used Lemma 2.11. □

Thus $(\bar{L}^n, \hat{\mu}^n)$ will converge, at least along subsequences. To prove the LDP we need to relate the limits of the controls $\hat{\mu}^n$ and the controlled process $\bar{L}^n$.

Lemma 3.5 *Suppose* $\{(\bar{L}^n, \hat{\mu}^n)\}_{n\in\mathbb{N}}$ *converges along a subsequence to* $(\bar{L}, \hat{\mu})$. *Then* $\bar{L} = \hat{\mu}$ *w.p.1.*

The proof of this result, which is a martingale version of the proof of the Glivenko–Cantelli lemma, will be given in Sect. 3.1.5 after we complete the proof of Sanov's theorem.

3.1.3 Laplace Upper Bound

The proof of Theorem 3.3 is partitioned into upper and lower bounds. In this section we will prove the Laplace upper bound, which is the same as the variational lower bound

$$\liminf_{n\to\infty} -\frac{1}{n}\log E\exp\{-nF(L^n)\} \geq \inf_{\mu\in\mathscr{P}(S)}\left[F(\mu) + R\left(\mu\,\|\gamma\right)\right]. \tag{3.5}$$

For $\varepsilon > 0$, let $\left\{\bar{\mu}_i^n\right\}_{i=1,\ldots,n}$ and $\{\bar{X}_i^n\}_{i=1,\ldots,n}$ come within ε of the infimum in (3.3):

$$-\frac{1}{n}\log E\exp\{-nF(L^n)\} + \varepsilon \geq E\left[F\left(\bar{L}^n\right) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n\,\|\gamma\right)\right].$$

Recall that we assume (without loss of generality) that the uniform bound in (3.4) holds, and thus by Lemma 3.4, $\{(\bar{L}^n, \hat{\mu}^n)\}$ is tight.

Owing to tightness, for every subsequence of $\{(\bar{L}^n, \hat{\mu}^n)\}$ we can extract a further subsequence that converges weakly. It suffices to prove (3.5) for such a subsubsequence. To simplify notation, we denote this subsubsequence by n, and its limit by $\left(\bar{L}, \hat{\mu}\right)$. According to Lemma 3.5, $\bar{L} = \hat{\mu}$ a.s. Using Jensen's inequality for the second inequality, the convergence in distribution, Fatou's lemma and lower semicontinuity of relative entropy for the third inequality, and the w.p.1 relation $\bar{L} = \hat{\mu}$ for the last inequality, we obtain

$$\begin{aligned}
\liminf_{n\to\infty} -\frac{1}{n}\log Ee^{-nF(L^n)} + \varepsilon &\geq \liminf_{n\to\infty} E\left[F\left(\bar{L}^n\right) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n\,\|\gamma\right)\right] \\
&\geq \liminf_{n\to\infty} E\left[F\left(\bar{L}^n\right) + R\left(\hat{\mu}^n\,\|\gamma\right)\right] \\
&\geq E\left[F\left(\bar{L}\right) + R\left(\hat{\mu}\,\|\gamma\right)\right] \\
&\geq \inf_{\mu\in\mathscr{P}(S)}\left[F(\mu) + R\left(\mu\,\|\gamma\right)\right].
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, (3.5) follows. □

3.1.4 Laplace Lower Bound

Next we prove the variational upper bound

$$\limsup_{n\to\infty} -\frac{1}{n}\log E\exp\{-nF(L^n)\} \le \inf_{\mu\in\mathscr{P}(S)}\left[F(\mu)+R(\mu\,\|\,\gamma)\right], \tag{3.6}$$

which establishes the Laplace lower bound. For $\varepsilon > 0$ let μ^* satisfy

$$F(\mu^*)+R\left(\mu^*\,\|\,\gamma\right) \le \inf_{\mu\in\mathscr{P}(S)}\left[F(\mu)+R(\mu\,\|\,\gamma)\right]+\varepsilon.$$

Then let $\bar{\mu}_i^n = \mu^*$ for all $n \in \mathbb{N}$ and $i \in \{1,\dots,n\}$. By either Lemma 3.5 or the ordinary Glivenko–Cantelli lemma, the weak limit of $\bar{L}^n$ equals μ^* w.p.1. The representation in Proposition 3.1 gives the first inequality below, and the dominated convergence theorem gives the equality

$$\begin{aligned}\limsup_{n\to\infty} -\frac{1}{n}\log Ee^{-nF(L^n)} &\le \limsup_{n\to\infty} E\left[F\left(\bar{L}^n\right)+\frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n\,\|\,\gamma\right)\right]\\ &= \left[F\left(\mu^*\right)+R\left(\mu^*\,\|\,\gamma\right)\right]\\ &\le \inf_{\mu\in\mathscr{P}(S)}\left[F(\mu)+R(\mu\,\|\,\gamma)\right]+\varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, the bound (3.6) follows. □

Remark 3.6 When combined with the previous subsection, the argument just given shows that for asymptotic optimality one can restrict to controls of the form $\bar{\mu}_i^n = \mu^*$, i.e., product measure.

3.1.5 Proof of Lemma 3.5 and Remarks on the Proof of Sanov's Theorem

Since S is Polish, there exists a countable separating class $\{f_m\}_{m\in\mathbb{N}}$ of bounded continuous functions (see Appendix A). Define $K_m \doteq \|f_m\|_\infty$ and $\Delta_{m,i}^n \doteq f_m\left(\bar{X}_i^n\right) - \int_S f_m(x)\,\bar{\mu}_i^n(dx)$. For every $\varepsilon > 0$,

$$\begin{aligned}&P\left\{\left|\frac{1}{n}\sum_{i=1}^n f_m\left(\bar{X}_i^n\right)-\frac{1}{n}\sum_{i=1}^n\int_S f_m(x)\,\bar{\mu}_i^n(dx)\right|>\varepsilon\right\}\\ &\qquad\le \frac{1}{\varepsilon^2}E\left[\frac{1}{n^2}\sum_{i,j=1}^n \Delta_{m,i}^n\Delta_{m,j}^n\right].\end{aligned}$$

Recall that $\mathscr{F}_j^n = \sigma(\bar{X}_i^n, i = 1, \ldots, j)$. By a standard conditioning argument, the off-diagonal terms vanish: for $i > j$,

$$E\left[\Delta_{m,i}^n \Delta_{m,j}^n\right] = E\left[E\left[\Delta_{m,i}^n \Delta_{m,j}^n \middle| \mathscr{F}_{i-1}^n\right]\right] = E\left[E\left[\Delta_{m,i}^n \middle| \mathscr{F}_{i-1}^n\right] \Delta_{m,j}^n\right] = 0.$$

Since $|\Delta_{m,i}^n| \leq 2K_m$,

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^n f_m(\bar{X}_i^n) - \frac{1}{n}\sum_{i=1}^n \int_S f_m(x)\bar{\mu}_i^n(dx)\right| > \varepsilon\right\} \leq \frac{4K_m^2}{n\varepsilon^2}.$$

Since $(\bar{L}^n, \hat{\mu}^n) \Rightarrow \left(\bar{L}, \hat{\mu}\right)$ and $\varepsilon > 0$ is arbitrary, by Fatou's lemma, we have

$$P\left\{\int_S f_m(x)\,\bar{L}(dx) = \int_S f_m(x)\,\hat{\mu}(dx)\right\} = 1.$$

Now use that $\{f_m\}$ is countable and separating to conclude that $\bar{L} = \hat{\mu}$ w.p.1. □

Remark 3.7 There is a close relationship between the legitimate use of Jensen's inequality in the proof of any particular Laplace upper bound and the asymptotic independence of optimal controls with respect to one or more parameters. In the context of Sanov's theorem, the parameter is the time index i. In the proof of the upper bound, the inequality

$$E\left[\frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n \| \gamma\right)\right] \geq E\left[R\left(\hat{\mu}^n \| \gamma\right)\right]$$

was used, where $\hat{\mu}^n$ is the average (over i) of $\bar{\mu}_i^n$. In general, Jensen's inequality holds with a strict inequality. There is an exception when the quantity being averaged is independent of the parameter over which the averaging occurs. Since we consider the limit $n \to \infty$, this means that there should be no loss due to the use of Jensen's inequality if one restricts to controls that are independent of i in this limit. In any particular instance, a use of Jensen's inequality is appropriate only when one proves the corresponding lower bound with the same rate function, i.e., in the proof of the lower bound one should be able to restrict to controls that do not depend on the parameter being averaged. This of course occurs in the proof of Sanov's theorem, since for the lower bound we consider controls of the form $\bar{\mu}_i^n = \mu^*$ for a fixed measure μ^*.

Information on what control dependencies are asymptotically unimportant can be useful in various ways, including the construction of importance sampling schemes, which is considered later in the book. It typically simplifies the large deviation proofs considerably, since one needs to keep track in the weak convergence analysis of only the nontrivial dependencies, and often one has some a priori insight into which parameters should be unimportant. However, as noted previously, it is only after the proof of upper and lower bounds with the same rate function that one can claim that the use of Jensen's inequality was without loss.

3.1.6 Cramér's Theorem

Cramér's theorem states the LDP for the empirical mean of $\mathbb{R}^d$-valued iid random variables: $S_n \doteq \frac{1}{n}(X_1 + \cdots + X_n)$. Of course, one can recover the empirical mean from the empirical measure via $S_n = \int_{\mathbb{R}^d} yL^n(dy)$. If the underlying distribution γ has compact support, then the mapping $\mu \to \int_{\mathbb{R}^d} y\mu(dy)$ is continuous on a subset of $\mathscr{P}(\mathbb{R}^d)$ that contains L^n w.p.1. In this case, the LDP for $\{S_n\}_{n\in\mathbb{N}}$ follows directly from the contraction principle [Theorem 1.16], with the rate function I given by

$$I(\beta) \doteq \inf\left[R(\mu\,\|\gamma) : \int_{\mathbb{R}^d} y\mu(dy) = \beta\right] \tag{3.7}$$

for $\beta \in \mathbb{R}^d$. However, in general the mapping $\mu \mapsto \int_{\mathbb{R}^d} y\mu(dy)$ is not continuous, and the contraction principle does not suffice. As we will see, the issue is that the conditions of Sanov's theorem are too weak to force continuity with high probability. They are sufficient to imply tightness of controls, but no more. Once the conditions are appropriately strengthened, the weak convergence arguments can be carried out just as before, with the only difference being in the qualitative properties of the convergence. For $\alpha \in \mathbb{R}^d$ let

$$H(\alpha) \doteq \log \int_{\mathbb{R}^d} e^{\langle \alpha, y\rangle}\gamma(dy).$$

Theorem 3.8 (CRAMÉR'S THEOREM) *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of iid $\mathbb{R}^d$-valued random variables with common distribution γ, and let $S_n \doteq \frac{1}{n}\sum_{i=1}^n X_i$. Assume that $H(\alpha) < \infty$ for all $\alpha \in \mathbb{R}^d$. Then $\{S_n\}_{n\in\mathbb{N}}$ satisfies the LDP with rate function I defined in (3.7).*

To prove the LDP we need to calculate the limits of

$$-\frac{1}{n}\log E\exp\left\{-nF\left(\int_{\mathbb{R}^d} yL^n(dy)\right)\right\}, \tag{3.8}$$

where $F \in \mathscr{C}_b(\mathbb{R}^d)$. From the representation in Proposition 3.1 we see that (3.8) equals

$$\inf_{\{\bar{\mu}_i^n\}} E\left[F\left(\int_{\mathbb{R}^d} y\bar{L}^n(dy)\right) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n\,\|\gamma\right)\right].$$

Once more, without loss of generality we can assume that the relative entropy cost is uniformly bounded, and in particular that (3.4) holds. The next lemma shows that as a consequence of this uniform bound and our assumption on H, the collection $\left\{\bar{L}^n\right\}_{n\in\mathbb{N}}$ is uniformly integrable.

Lemma 3.9 *Assume* (3.4) *and that* $H(\alpha) < \infty$ *for all* $\alpha \in \mathbb{R}^d$. *Then*

$$\lim_{M\to\infty} \limsup_{n\to\infty} E\left[\int_{\mathbb{R}^d} \|y\| \, 1_{\{\|y\|\geq M\}} \bar{L}^n(dy)\right] = 0.$$

Before proving the lemma we complete the proof of Theorem 3.8.

Proof (of Theorem 3.8) The uniform integrability of Lemma 3.9 implies that if $\bar{L}^n$ converges in distribution to $\bar{L}$ and (3.4) holds, then

$$E\left[F\left(\int_{\mathbb{R}^d} y\bar{L}^n(dy)\right)\right] \to E\left[F\left(\int_{\mathbb{R}^d} y\bar{L}(dy)\right)\right]. \tag{3.9}$$

The limit of (3.8) will now be calculated using essentially the same argument as that used to prove Sanov's theorem.

Variational lower bound. For $\varepsilon > 0$ let $\left\{\bar{\mu}_i^n\right\}_{i=1,\ldots,n}$ and $\{\bar{X}_i^n\}_{i=1,\ldots,n}$ satisfy

$$-\frac{1}{n}\log E e^{-nF\left(\int_{\mathbb{R}^d} yL^n(dy)\right)} + \varepsilon \geq E\left[F\left(\int_{\mathbb{R}^d} y\bar{L}^n(dy)\right) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n \,\|\gamma\right)\right].$$

Consider a subsubsequence as in Sect. 3.1.3 (denoted again by n) along which $\left(\bar{L}^n, \hat{\mu}^n\right)$ converges weakly to $\left(\bar{L}, \hat{\mu}\right)$. Then as in Sect. 3.1.3, we have

$$\begin{aligned}
&\liminf_{n\to\infty} -\frac{1}{n}\log E\exp\left\{-nF\left(\int_{\mathbb{R}^d} yL^n(dy)\right)\right\} + \varepsilon \\
&\geq \liminf_{n\to\infty} E\left[F\left(\int_{\mathbb{R}^d} y\bar{L}^n(dy)\right) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n \,\|\gamma\right)\right] \\
&\geq E\left[F\left(\int_{\mathbb{R}^d} y\bar{L}(dy)\right) + R\left(\hat{\mu}\,\|\gamma\right)\right] \\
&\geq E\left[F\left(\int_{\mathbb{R}^d} y\bar{L}(dy)\right) + I\left(\int_{\mathbb{R}^d} y\bar{L}(dy)\right)\right] \\
&\geq \inf_{\beta\in\mathbb{R}^d}\left[F(\beta) + I\left(\beta\right)\right].
\end{aligned}$$

Here the second inequality follows from (3.9), and the third follows from the definition of I and $\bar{L} = \hat{\mu}$ a.s. Since $\varepsilon > 0$ is arbitrary, the lower bound follows.

Variational upper bound. For $\varepsilon \in (0, 1)$ let $\beta^* \in \mathbb{R}^d$ satisfy

$$F(\beta^*) + I\left(\beta^*\right) \leq \inf_{\beta\in\mathbb{R}^d}\left[F(\beta) + I\left(\beta\right)\right] + \varepsilon.$$

Next let $\mu^* \in \mathscr{P}(\mathbb{R}^d)$ be such that $\int_{\mathbb{R}^d} x\mu^*(dx) = \beta^*$ and

$$F(\beta^*) + R\left(\mu^* \,\|\gamma\right) \le F(\beta^*) + I\left(\beta^*\right) + \varepsilon.$$

As in Sect. 3.1.4, let $\bar{\mu}_i^n = \mu^*$ for all $n \in \mathbb{N}$ and $i \in \{1, \ldots, n\}$. Then the weak limit of $\bar{L}^n$ equals μ^* a.s., and (3.4) is satisfied. Thus

$$\begin{aligned}
&\limsup_{n\to\infty} -\frac{1}{n} \log E \exp\left\{-nF\left(\int_{\mathbb{R}^d} yL^n(dy)\right)\right\} \\
&\le \limsup_{n\to\infty} E\left[F\left(\int_{\mathbb{R}^d} y\bar{L}^n(dy)\right) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n \,\|\gamma\right)\right] \\
&= F\left(\beta^*\right) + R\left(\mu^* \,\|\gamma\right) \\
&\le F(\beta^*) + I\left(\beta^*\right) + \varepsilon \\
&\le \inf_{\beta\in\mathbb{R}^d} [F(\beta) + I\,(\beta)] + 2\varepsilon.
\end{aligned}$$

Here the equality follows from (3.9) and the a.s. convergence of $\bar{L}^n$ to μ^*. Since $\varepsilon \in (0, 1)$ is arbitrary, the upper bound follows. □

Finally, we give the proof of Lemma 3.9.

Proof (of Lemma 3.9) The uniform integrability stated in this lemma is essentially a consequence of the bound on relative entropy costs and the assumption $H(\alpha) < \infty$. For $b \ge 0$ let

$$\ell(b) \doteq b \log b - b + 1. \tag{3.10}$$

We recall a bound already used frequently in Chap. 2 [see (2.9)]: for $a \ge 0$, $b \ge 0$, and $\sigma \ge 1$,

$$ab \le e^{\sigma a} + \frac{1}{\sigma}\left(b \log b - b + 1\right) = e^{\sigma a} + \frac{1}{\sigma}\ell(b).$$

Thus if $\theta \in \mathscr{P}(\mathbb{R}^d)$ satisfies $\theta \ll \gamma$, then for every $\sigma \ge 1$,

$$\begin{aligned}
\int_{\mathbb{R}^d} \|y\|\, 1_{\{\|y\|\ge M\}}\theta(dy) &= \int_{\mathbb{R}^d} \|y\|\, 1_{\{\|y\|\ge M\}} \frac{d\theta}{d\gamma}(y)\gamma(dy) \\
&\le \int_{\mathbb{R}^d} e^{\sigma\|y\|} 1_{\{\|y\|\ge M\}}\gamma(dy) + \frac{1}{\sigma}\int_{\mathbb{R}^d} \ell\left(\frac{d\theta}{d\gamma}(y)\right)\gamma(dy) \\
&= \int_{\mathbb{R}^d} e^{\sigma\|y\|} 1_{\{\|y\|\ge M\}}\gamma(dy) + \frac{1}{\sigma} R\left(\theta \,\|\gamma\right).
\end{aligned}$$

Note that the inequality holds trivially if $\theta \not\ll \gamma$. Therefore,

$$\begin{aligned}
E\int_{\mathbb{R}^d} \|y\|\, 1_{\{\|y\|\ge M\}}\bar{L}^n(dy) &= E\int_{\mathbb{R}^d} \|y\|\, 1_{\{\|y\|\ge M\}}\hat{\mu}^n(dy) \\
&\le \int_{\mathbb{R}^d} e^{\sigma\|y\|} 1_{\{\|y\|\ge M\}}\gamma(dy) + \frac{1}{\sigma} E R\left(\hat{\mu}^n \,\|\gamma\right).
\end{aligned} \tag{3.11}$$

Since $H(\alpha) < \infty$ for all $\alpha \in \mathbb{R}^d$, for each fixed σ the mapping $y \mapsto \exp\{\sigma \|y\|\}$ is integrable with respect to γ. To see this, for $\lambda > 0$ let

$$m(\lambda) \doteq \sup_{\alpha \in \mathbb{R}^d : \|\alpha\| \le \lambda} e^{H(\alpha)} = \sup_{\alpha \in \mathbb{R}^d : \|\alpha\| \le \lambda} \int_{\mathbb{R}^d} e^{\langle \alpha, y \rangle} \gamma(dy).$$

From the continuity of $\alpha \mapsto H(\alpha)$ it follows that $m(\lambda) < \infty$. For $J \subset \{1, \ldots, d\}$ let $\mathbb{R}^d_J \doteq \{x \in \mathbb{R}^d : x_i \ge 0 \text{ if and only if } i \in J\}$, and define $\alpha^J \in \mathbb{R}^d$ by

$$\alpha^J_i \doteq \frac{\lambda}{\sqrt{d}} \text{ if } i \in J \text{ and } \alpha^J_i \doteq -\frac{\lambda}{\sqrt{d}} \text{ if } i \in J^c.$$

Then $\|\alpha^J\| = \lambda$ for all J, and for all $y \in \mathbb{R}^d_J$,

$$\langle \alpha^J, y \rangle = \frac{\lambda}{\sqrt{d}} \sum_{i=1}^{d} |y_i| \ge \frac{\lambda}{\sqrt{d}} \|y\|.$$

Thus

$$m(\lambda) \ge \int_{\mathbb{R}^d} e^{\langle \alpha^J, y \rangle} \gamma(dy) \ge \int_{\mathbb{R}^d_J} e^{\langle \alpha^J, y \rangle} \gamma(dy) \ge \int_{\mathbb{R}^d_J} e^{\frac{\lambda}{\sqrt{d}} \|y\|} \gamma(dy),$$

and therefore

$$\int_{\mathbb{R}^d} e^{\frac{\lambda}{\sqrt{d}} \|y\|} \gamma(dy) = \sum_J \int_{\mathbb{R}^d_J} e^{\frac{\lambda}{\sqrt{d}} \|y\|} \gamma(dy) \le 2^d m(\lambda). \tag{3.12}$$

Since $\lambda > 0$ is arbitrary, we get $\int_{\mathbb{R}^d} \exp\{\sigma \|y\|\} \gamma(dy) < \infty$ for every $\sigma \in \mathbb{R}$, as asserted.

The bound (3.4) on the relative entropy and Jensen's inequality imply that the last term in (3.11) is bounded by $(2\|F\|_\infty + 1)/\sigma$. The conclusion of Lemma 3.9 follows by taking limits in (3.11), in the order $n \to \infty$, $M \to \infty$, and then $\sigma \to \infty$. □

Remark 3.10 The proof most often given of Cramér's theorem (e.g., as in [239]) uses a change of measure argument for the large deviation lower bound and Chebyshev's inequality for the upper bound. This line of argument naturally produces the following alternative form of the rate function as the **Legendre-Fenchel transform** of H:

$$L(\beta) = \sup_{\alpha \in \mathbb{R}^d} \left[\langle \alpha, \beta \rangle - H(\alpha) \right].$$

By the uniqueness of rate functions [Theorem 1.15] it must be that $I = L$, though one can also directly verify that the two coincide [Lemma 4.16]. Both characterizations of the rate are useful. For example, the description as a Legendre transform easily shows that I is convex, while the characterization in terms of relative entropy allows

an easy calculation of the domain of finiteness of I. Note also that in principle, the two different expressions can be used to obtain upper and lower bounds on $I(\beta)$ for any given β. The two descriptions are in fact dual to each other.

Remark 3.11 It is possible to prove Cramér's theorem under just the condition that there is $\delta > 0$ such that $H(\alpha) < \infty$ for all α with $\|\alpha\| \leq \delta$. The main difficulty imposed by this weaker condition is that boundedness of costs does not imply the uniform integrability of controls that is used in the proof of the variational lower bound. This can be bypassed by the use of unbounded test functions of the form $F(x) = \infty 1_{C^c}(x)$, where C is convex. In the proof of the variational upper bound (large deviation lower bound) we can take C to be an open ball of radius $\delta > 0$ about a point x. Tightness follows, since here one picks controls that correspond to product measure. For the lower bound one must first establish that lower bounds for convex sets, which correspond to large deviation upper bounds, suffice to establish the full large deviation upper bound. This can be shown by approximating the complement of a level set of the rate function by a finite union of half-spaces (see the proof of Cramér's theorem in [239]), which uses the compactness of the level sets and an open covering argument. Given that it is sufficient to prove the variational lower bound for just convex sets, Jensen's inequality can be used to move the expected value inside F in the representation, and all that is required to complete the proof is boundedness of $E \int_{\mathbb{R}^d} x \hat{\mu}^n(dx)$ when costs are bounded. Since boundedness of costs implies boundedness of $L(E \int_{\mathbb{R}^d} x \hat{\mu}^n(dx))$, this follows, since L has compact level sets.

3.2 Representation for Functionals of Brownian Motion

Let $(\Omega, \mathscr{F}, P)$ be a probability space and $T \in (0, \infty)$. A filtration $\{\mathscr{F}_t\}_{0 \leq t \leq T}$ is a collection of sub-sigma fields of $\mathscr{F}$ with the property $\mathscr{F}_s \subset \mathscr{F}_t$ for $s \leq t$. A filtration $\{\mathscr{F}_t\}_{0 \leq t \leq T}$ is called right continuous if $\cap_{s>t} \mathscr{F}_s = \mathscr{F}_t$ for every $t \in [0, T)$. A filtration $\{\mathscr{F}_t\}_{t \in [0,T]}$ is said to satisfy **the usual conditions** if it is right continuous and for every $t \in [0, T]$, $\mathscr{F}_t$ contains all P-null sets in $\mathscr{F}$. All filtrations in this book will satisfy the usual conditions. Suppose we are given such a filtration $\{\mathscr{F}_t\}$ on $(\Omega, \mathscr{F}, P)$ and that $\{W(t)\}_{0 \leq t \leq T}$ is a k-dimensional $\mathscr{F}_t$-**Brownian motion**, i.e., $W(0) = 0$; W has continuous trajectories; $W(t)$ is $\mathscr{F}_t$-measurable for every $t \in [0, T]$; and $W(t) - W(s)$ is independent of $\mathscr{F}_s$ for all $0 \leq s \leq t \leq T$ and is normally distributed with mean zero and variance $(t - s)$. A standard choice of $\mathscr{F}_t$ is the sigma-field $\sigma\{W(s) : 0 \leq s \leq t\}$, augmented with all P-null sets, i.e.,

$$\mathscr{G}_t \doteq \sigma\left\{\sigma\{W(s) : 0 \leq s \leq t\} \vee \mathscr{N}\right\},$$

where $\mathscr{N} = \{A \subset \Omega$: there is $B \in \mathscr{F}$ with $A \subset B$ and $P(B) = 0\}$.

Definition 3.12 An $\mathbb{R}^k$-valued stochastic process $\{v(t)\}_{0\le t\le T}$ on $(\Omega, \mathscr{F}, P)$ is said to be $\mathscr{F}_t$-**progressively measurable** if for every $t \in [0, T]$, the map $(s, \omega) \mapsto v(s, \omega)$ from $([0, t] \times \Omega, \mathscr{B}([0, t]) \otimes \mathscr{F}_t)$ to $(\mathbb{R}^k, \mathscr{B}(\mathbb{R}^k))$ is measurable.

Definition 3.13 Let $\mathscr{A}$ [resp., $\bar{\mathscr{A}}$] denotes the collection of all $\mathscr{G}_t$-progressively [resp., $\mathscr{F}_t$-progressively] measurable processes $\{v(t)\}_{0\le t\le T}$ that satisfy the integrability condition $E[\int_0^T \|v(t)\|^2 dt] < \infty$.

The following representation theorem for bounded measurable functionals of a Brownian motion is analogous to the one stated in Proposition 3.1 for functionals of an iid sequence. It is a special case of a representation that will be proved in Chap. 8 [Theorem 8.3]. In the representation, the controlled measures have been replaced by just a control process, and the relative entropy cost is the expected L^2-norm of this process. Recall that $\mathscr{C}([0, T] : \mathbb{R}^k)$ denotes the space of $\mathbb{R}^k$-valued continuous functions on $[0, T]$. This space is equipped with the uniform metric, which makes it a Polish space.

Theorem 3.14 *Let G be a bounded Borel measurable function mapping $\mathscr{C}([0, T] : \mathbb{R}^k)$ into $\mathbb{R}$. Then*

$$-\log E e^{-G(W)} = \inf_{v\in\mathscr{A}} E\left[G\left(W + \int_0^{\cdot} v(s)ds\right) + \frac{1}{2}\int_0^T \|v(s)\|^2 ds\right]. \tag{3.13}$$

Remark 3.15 The proof of this representation first appeared in [32]. The form of the representation closely parallels the corresponding discrete time result for product measure, reflecting the fact that Brownian motion is the integral of "white" noise, and progressive measurability is analogous to the fact that in the representation for iid noises, $\bar{\mu}_i^n$ is allowed to depend on all controlled noises up to time $i-1$. In fact, if one replaces W by the corresponding piecewise linear interpolation with interpolation interval $\delta > 0$ (which is equivalent to a collection of $1/\delta$ iid $N(0, \delta)$ random variables) and assumes that the minimizing measures are Gaussian with means $\delta\bar{v}_i^n$, then the L^2 cost in (3.13) corresponds to $R\left(N(\delta\bar{v}_i^n, \delta)\,\|\,N(0, \delta)\right) = \delta\|\bar{v}_i^n\|^2/2$. The assumption that one can restrict the discrete time measures to those of the form $N(\delta\bar{v}_i^n, \delta)$ is valid in the limit $\delta \to 0$, which is why the continuous time representation is in some ways simpler than the corresponding discrete time representation.

Remark 3.16 One can replace the class $\mathscr{A}$ with $\bar{\mathscr{A}}$ in (3.13) (see Chap. 8). Although in this chapter we use progressively measurable controls (as in [32]), in Chap. 8 these are replaced by predictable controls. For the case of Brownian motion, the two are interchangeable, since any $\mathscr{G}_t$ [resp., $\mathscr{F}_t$] predictable process satisfying the square integrability condition in Definition 3.13 is in $\mathscr{A}$ [resp., $\bar{\mathscr{A}}$], and conversely, to any v in $\mathscr{A}$ [resp., $\bar{\mathscr{A}}$] there is a predictable $\tilde{v}$ in $\mathscr{A}$ [resp., $\bar{\mathscr{A}}$] such that $v(t, \omega) = \tilde{v}(t, \omega)$ a.s. $dt \times P$; see [168, Remark 3.3.1]. However, predictability is needed for the case of processes with jumps, e.g., systems driven by a Poisson random measure.

We next state a version of the representation that restricts the class of controls to a compact set. For $M \in [0, \infty)$ let

$$S_M \doteq \left\{ \phi \in \mathscr{L}^2([0,T] : \mathbb{R}^k) : \int_0^T \|\phi(s)\|^2 \, ds \le M \right\},$$

where $\mathscr{L}^2([0,T] : \mathbb{R}^k)$ is the Hilbert space of square integrable functions from $[0,T]$ to $\mathbb{R}^k$, and define $\mathscr{A}_{b,M}$ to be the subset of $\mathscr{A}$ such that $v \in \mathscr{A}_{b,M}$ if $v(\omega) \in S_M$ for all $\omega \in \Omega$. Let $\mathscr{A}_b = \cup_{M=1}^{\infty} \mathscr{A}_{b,M}$. In the statement of the theorem, we introduce a scaling that will be appropriate for large deviation analysis of small noise diffusions.

Theorem 3.17 *Let G be a bounded Borel measurable function mapping $\mathscr{C}([0,T] : \mathbb{R}^k)$ into $\mathbb{R}$ and let $\delta > 0$. Then there exist $M < \infty$ depending on $\|G\|_\infty$ and δ such that for all $\varepsilon \in (0,1)$,*

$$\begin{aligned} &-\varepsilon \log E \exp\left\{ -\frac{1}{\varepsilon} G(\sqrt{\varepsilon} W) \right\} \\ &\ge \inf_{v \in \mathscr{A}_{b,M}} E\left[G\left(\sqrt{\varepsilon} W + \int_0^{\cdot} v(s) ds \right) + \frac{1}{2} \int_0^T \|v(s)\|^2 ds \right] - \delta. \end{aligned} \tag{3.14}$$

Proof To consolidate notation, for $v \in \mathscr{A}$ let $W^v \doteq W + \int_0^{\cdot} v(s) ds$. For the given $\varepsilon \in (0,1)$ and $\eta \in (0,1)$, choose $\tilde{v}^\varepsilon \in \mathscr{A}$ such that

$$\begin{aligned} &\inf_{v \in \mathscr{A}} E\left[G\left(\sqrt{\varepsilon} W^{v/\sqrt{\varepsilon}} \right) + \frac{1}{2} \int_0^T \|v\|^2 \, ds \right] \\ &\quad \ge E\left[G\left(\sqrt{\varepsilon} W^{\tilde{v}^\varepsilon/\sqrt{\varepsilon}} \right) + \frac{1}{2} \int_0^T \|\tilde{v}^\varepsilon\|^2 \, ds \right] - \eta. \end{aligned}$$

From the boundedness of G it follows that

$$\infty > C_G \doteq 2(2\|G\|_\infty + 1) \ge \sup_{\varepsilon \in (0,1)} E\left[\int_0^T \|\tilde{v}^\varepsilon(s)\|^2 \, ds \right].$$

We next show using an approximation argument that one can in fact assume an almost sure bound. For $M \in (0,\infty)$ let

$$\tau_M^\varepsilon \doteq \inf\left[t \in [0,T] : \int_0^t \|\tilde{v}^\varepsilon(s)\|^2 \, ds \ge M \right] \wedge T.$$

Note that v^ε defined by $v^\varepsilon(s) \doteq \tilde{v}^\varepsilon(s) 1_{[0,\tau_M^\varepsilon]}(s)$, $s \in [0,T]$ is an element of $\mathscr{A}$, and that $v^\varepsilon \in S_M$ a.s. Note also that

$$\begin{aligned} &E\left[G\left(\sqrt{\varepsilon} W^{\tilde{v}^\varepsilon/\sqrt{\varepsilon}} \right) + \frac{1}{2} \int_0^T \|\tilde{v}^\varepsilon(s)\|^2 \, ds \right] \\ &\quad \ge E\left[G\left(\sqrt{\varepsilon} W^{v^\varepsilon/\sqrt{\varepsilon}} \right) + \frac{1}{2} \int_0^T \|v^\varepsilon(s)\|^2 \, ds \right] \end{aligned}$$

$$+ E\left[G\left(\sqrt{\varepsilon}W^{\tilde{v}^\varepsilon/\sqrt{\varepsilon}}\right) - G\left(\sqrt{\varepsilon}W^{v^\varepsilon/\sqrt{\varepsilon}}\right)\right].$$

By Chebyshev's inequality,

$$E\left[\left|G\left(\sqrt{\varepsilon}W^{\tilde{v}^\varepsilon/\sqrt{\varepsilon}}\right) - G\left(\sqrt{\varepsilon}W^{v^\varepsilon/\sqrt{\varepsilon}}\right)\right|\right] \le 2\|G\|_\infty P\{\tau_M^\varepsilon < T\} \le 2\|G\|_\infty \frac{C_G}{M}.$$

For $\delta > 0$, let $M = (2\|G\|_\infty C_G + 1)/\delta$. Then for all $\varepsilon \in (0, 1)$,

$$\begin{aligned} &E\left[G\left(\sqrt{\varepsilon}W^{\tilde{v}^\varepsilon/\sqrt{\varepsilon}}\right) + \frac{1}{2}\int_0^T \|\tilde{v}^\varepsilon(s)\|^2\, ds\right] \\ &\quad \ge E\left[G\left(\sqrt{\varepsilon}W^{v^\varepsilon/\sqrt{\varepsilon}}\right) + \frac{1}{2}\int_0^T \|v^\varepsilon(s)\|^2\, ds\right] - \delta. \end{aligned}$$

Since $\eta > 0$ is arbitrary, the conclusion of the theorem follows from the last display and Theorem 3.14. □

3.2.1 *Large Deviation Theory of Small Noise Diffusions*

The representation (3.13) and its variant (3.14) are very convenient for weak convergence large deviation analysis, and in many ways they make the continuous time setting simpler than the corresponding discrete time setting. As an illustration of their use we prove the large deviation principle for a class of small noise diffusions. While fairly general, the assumptions on the coefficients are chosen to make the presentation simple, and they can be significantly relaxed.

Condition 3.18 *There is $C \in (0, \infty)$ such that $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times k}$ satisfy*

$$\|b(x) - b(y)\| + \|\sigma(x) - \sigma(y)\| \le C\|x - y\|, \quad \|b(x)\| + \|\sigma(x)\| \le C(1 + \|x\|)$$

for all $x, y \in \mathbb{R}^d$.

Fix $x \in \mathbb{R}^d$, and for $\varepsilon > 0$ let $X^\varepsilon = \{X^\varepsilon(t)\}_{0\le t\le T}$ be the strong solution of the stochastic differential equation (SDE) (cf. [172, Sect. 5.2])

$$dX^\varepsilon(t) = b(X^\varepsilon(t))dt + \sqrt{\varepsilon}\sigma(X^\varepsilon(t))dW(t), \ X^\varepsilon(0) = x. \tag{3.15}$$

Let $\mathscr{AC}_x([0, T] : \mathbb{R}^d)$ denote the space of $\mathbb{R}^d$-valued absolutely continuous functions φ on $[0, T]$ with $\varphi(0) = x$. Also, for $\varphi \in \mathscr{AC}_x([0, T] : \mathbb{R}^d)$, let

$$U_\varphi = \left\{ u \in \mathscr{L}^2([0,T] : \mathbb{R}^k) : \varphi(\cdot) = x + \int_0^\cdot b(\varphi(s))ds + \int_0^\cdot \sigma(\varphi(s))u(s)ds \right\}. \tag{3.16}$$

For all other $\varphi \in \mathscr{C}([0,T] : \mathbb{R}^d)$ let U_φ be the empty set. The following large deviation principle for such small noise diffusions is one of the classical results in the theory [140]. Following our standard convention, the infimum over the empty set is taken to be ∞.

Theorem 3.19 *Assume Condition 3.18. Then the collection* $\{X^\varepsilon\}_{\varepsilon\in(0,1)}$ *satisfies the LDP on* $\mathscr{C}([0,T] : \mathbb{R}^d)$ *with rate function*

$$I(\varphi) \doteq \inf_{u\in U_\varphi} \left[\frac{1}{2} \int_0^T \|u(t)\|^2 dt \right].$$

To prove the theorem, we must show that I is a rate function and for bounded and continuous $F : \mathscr{C}([0,T] : \mathbb{R}^d) \to \mathbb{R}$,

$$\lim_{\varepsilon\to 0} -\varepsilon \log E \exp\left\{ -\frac{1}{\varepsilon} F(X^\varepsilon) \right\} = \inf_{\varphi\in\mathscr{C}([0,T]:\mathbb{R}^d)} [F(\varphi) + I(\varphi)].$$

Following a convention that is used here for the first time, we present the proof just for the case $T = 1$, noting that the general case involves only notational differences. The first step is to interpret $F(X^\varepsilon)$ as a bounded measurable function of W. From unique pathwise solvability of the SDE in (3.15) [172, Definition 5.3.2 and Corollary 5.3.23] it follows that for each $\varepsilon > 0$, there is a measurable map $\mathscr{G}^\varepsilon : \mathscr{C}([0,1] : \mathbb{R}^k) \to \mathscr{C}([0,1] : \mathbb{R}^d)$ such that whenever $\tilde{W}$ is a k-dimensional standard Brownian motion given on some probability space $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{P})$, then $\tilde{X}^\varepsilon = \mathscr{G}^\varepsilon(\sqrt{\varepsilon}\tilde{W})$ is the unique solution of the SDE (3.15) with W replaced by $\tilde{W}$. Recalling the notation $W^v \doteq W + \int_0^\cdot v(s)ds$, this says that

$$\begin{aligned} -\varepsilon \log E \exp\left\{ -\frac{1}{\varepsilon} F(X^\varepsilon) \right\} &= -\varepsilon \log E \exp\left\{ -\frac{1}{\varepsilon} F \circ \mathscr{G}^\varepsilon(\sqrt{\varepsilon}W) \right\} \\ &= \inf_{v\in\mathscr{A}} E\left[F \circ \mathscr{G}^\varepsilon(\sqrt{\varepsilon}W^{v/\sqrt{\varepsilon}}) + \frac{1}{2}\int_0^1 \|v(s)\|^2 ds \right]. \end{aligned}$$

Assume that $v \in \mathscr{A}_{b,M}$ for some $M < \infty$, and consider the probability measure Q^ε on $(\Omega, \mathscr{F})$ defined by

$$\frac{dQ^\varepsilon}{dP} = \exp\left[-\frac{1}{\sqrt{\varepsilon}} \int_0^1 v(s)dW(s) - \frac{1}{2\varepsilon}\int_0^1 \|v(s)\|^2 ds \right].$$

From Girsanov's theorem (see Theorem D.1) it follows that $Q^\varepsilon\{\sqrt{\varepsilon}W^{v/\sqrt{\varepsilon}} \in \cdot\} = P\{\sqrt{\varepsilon}W \in \cdot\}$. Consequently $\bar{X}^\varepsilon = \mathscr{G}^\varepsilon(\sqrt{\varepsilon}W^{v/\sqrt{\varepsilon}})$ solves the SDE

$$d\bar{X}^\varepsilon(t) = b(\bar{X}^\varepsilon(t))dt + \sqrt{\varepsilon}\sigma(\bar{X}^\varepsilon(t))dW^{v/\sqrt{\varepsilon}}(t), \;\; \bar{X}^\varepsilon(0) = x$$

on the filtered probability space $(\Omega, \mathscr{F}, Q^\varepsilon, \{\mathscr{F}_t\})$. Since Q^ε is mutually absolutely continuous with respect to P, it follows that $\bar{X}^\varepsilon$ is the unique solution of the following SDE on $(\Omega, \mathscr{F}, P, \{\mathscr{F}_t\})$:

$$d\bar{X}^\varepsilon(t) = b(\bar{X}^\varepsilon(t))dt + \sqrt{\varepsilon}\sigma(\bar{X}^\varepsilon(t))dW(t) + \sigma(\bar{X}^\varepsilon(t))v(t)dt, \;\; \bar{X}^\varepsilon(0) = x. \tag{3.17}$$

Thus whenever $v \in \mathscr{A}_{b,M}$, we have that $\mathscr{G}^\varepsilon(\sqrt{\varepsilon}W^{v/\sqrt{\varepsilon}})$ and the solution to (3.17) coincide. A collection of controls $\{v^\varepsilon\} \subset \mathscr{A}_{b,M}$ for fixed $M < \infty$ will be regarded as a collection of S_M-valued random variables, where S_M is equipped with the weak topology on the Hilbert space $\mathscr{L}^2([0,1] : \mathbb{R}^d)$. Recall that in a Hilbert space $(\mathscr{H}, \langle\cdot,\cdot\rangle)$, $f_n \to f$ under **the weak topology** if for all $g \in \mathscr{H}$, $\langle f_n - f, g\rangle \to 0$. Since S_M is weakly compact in $\mathscr{L}^2([0,1] : \mathbb{R}^d)$, such a collection is automatically tight.

We now turn to the proof of the LDP, which will follow the same scheme of proof as in Sanov's theorem. Thus we first prove a tightness result and show how to relate the weak limits of controls and controlled processes. The proof of the variational lower bound (which corresponds to the Laplace upper bound) as well as the proof that I is a rate function follows, and we conclude with the proof of the variational upper bound (Laplace lower bound).

3.2.2 Tightness and Weak Convergence

As noted above, a collection of controls $\{v^\varepsilon\} \subset \mathscr{A}_{b,M}$ is trivially tight, since S_M is compact. The following lemma shows that the corresponding collection of solutions of controlled SDEs is also tight.

Lemma 3.20 *Assume Condition 3.18. Consider any collection of controls $\{v^\varepsilon\} \subset \mathscr{A}_{b,M}$ for fixed $M < \infty$, and define $\bar{X}^\varepsilon$ by (3.17) with $v = v^\varepsilon$. Then $\{(\bar{X}^\varepsilon, v^\varepsilon)\}_{\varepsilon\in(0,1)}$ is a tight collection of $\mathscr{C}([0,1] : \mathbb{R}^d) \times S_M$-valued random variables.*

Proof Tightness of $\{v^\varepsilon\}$ is immediate. Since for $\varepsilon \in (0,1)$, $\int_0^1 \|v^\varepsilon(s)\|^2 ds \le M$ a.s., it follows on using the linear growth properties of the coefficients and an application of Gronwall's lemma [Lemma E.2] that

$$\sup_{\varepsilon\in(0,1)} E \sup_{0\le t\le 1} \|\bar{X}^\varepsilon(t)\|^2 < \infty. \tag{3.18}$$

Also note that

$$\bar{X}^\varepsilon(t) - x = \int_0^t b(\bar{X}^\varepsilon(s))ds + \sqrt{\varepsilon}\int_0^t \sigma(\bar{X}^\varepsilon(s))dW(s) + \int_0^t \sigma(\bar{X}^\varepsilon(s))v^\varepsilon(s)ds. \tag{3.19}$$

The first and second terms on the right side are easily seen to be tight in $\mathscr{C}([0,1] : \mathbb{R}^d)$ using the moment bound (3.18). Tightness of the third follows on using the inequality

$$\left\| \int_s^t \sigma(\bar{X}^\varepsilon(r))v^\varepsilon(r)dr \right\| \le C(t-s)^{1/2}\left(1+\sup_{0\le t\le 1}\|\bar{X}^\varepsilon(t)\|\right)\left(\int_0^1 \|v^\varepsilon(r)\|^2\,dr\right)^{1/2}$$
$$\le C(t-s)^{1/2}M^{1/2}\left(1+\sup_{0\le t\le 1}\|\bar{X}^\varepsilon(t)\|\right)$$

for $0 \le s \le t \le 1$ and once more using the moment bound. □

The following lemma will be used to characterize the limit points of $\{(\bar{X}^\varepsilon, v^\varepsilon)\}$.

Lemma 3.21 *Assume Condition 3.18. Suppose for each $\varepsilon \in (0,1)$ that $(\bar{X}^\varepsilon, v^\varepsilon)$ solves (3.19), and that $(\bar{X}^\varepsilon, v^\varepsilon)$ converges weakly to $(\bar{X}, v)$ as $\varepsilon \to 0$. Then w.p.1,*

$$\bar{X}(t) - x = \int_0^t b(\bar{X}(s))ds + \int_0^t \sigma(\bar{X}(s))v(s)ds. \tag{3.20}$$

Proof By a standard martingale bound (see (D.3) and Sect. D.2.1),

$$E\sup_{0\le t\le T}\left\|\int_0^t \sigma(\bar{X}^\varepsilon(r))dW(r)\right\|^2 \le C\int_0^T E\left(1+\|\bar{X}^\varepsilon(r)\|^2\right)dr,$$

and thus using the moment bound in (3.18), the stochastic integral term in (3.19) converges to 0 as $\varepsilon \to 0$. By the continuous mapping theorem, it suffices to check that for each $t \in [0,1]$, the maps $\phi \mapsto \int_0^t b(\phi(s))ds$ and $(\phi, u) \mapsto \int_0^t \sigma(\phi(s))u(s)ds$, from $\mathscr{C}([0,1]:\mathbb{R}^d)$ to $\mathbb{R}^d$ and from $\mathscr{C}([0,1]:\mathbb{R}^d) \times S_M$ to $\mathbb{R}^d$, are continuous. The continuity of the first map is immediate from the Lipschitz property of b. Consider now the second map. Suppose $\phi_n \to \phi$ in $\mathscr{C}([0,1]:\mathbb{R}^d)$ and $u_n \to u$ in S_M as $n \to \infty$. We can write

$$\int_0^t \sigma(\phi_n(s))u_n(s)ds - \int_0^t \sigma(\phi(s))u(s)ds$$
$$= \int_0^t [\sigma(\phi_n(s)) - \sigma(\phi(s))]\,u_n(s)ds + \int_0^t \sigma(\phi(s))\,[u_n(s) - u(s)]\,ds.$$

The first term tends to zero by Hölder's inequality and since $u_n \in S_M$, and the second converges to zero since $s \mapsto \sigma(\phi(s))1_{[0,t]}(s)$ is in $\mathscr{L}^2([0,1]:\mathbb{R}^d)$ and $u_n \to u$ in S_M. □

3.2.3 Laplace Upper Bound

We now prove the Laplace upper bound by establishing the lower bound

$$\liminf_{\varepsilon\to 0} -\varepsilon \log E\exp\left\{-\frac{1}{\varepsilon}F(X^\varepsilon)\right\} \ge \inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)}[F(\varphi)+I(\varphi)]. \tag{3.21}$$

We prove (3.21) using the variational representation. It suffices to show that for every sequence $\varepsilon_k \to 0$ there is a further subsequence for which (3.21) holds when the limit inferior on the left side is taken along the particular subsequence. Let $\delta > 0$, and with $G = F \circ \mathscr{G}^\varepsilon$ choose M according to Theorem 3.17 (note that M does not depend on ε), and choose a sequence $\{v^\varepsilon\} \subset \mathscr{A}_{b,M}$ that is within δ of the infimum in (3.14). We now fix a sequence $\{\varepsilon_k\}$. From Lemma 3.20 we can find a subsequence along which $(\bar{X}^{\varepsilon_k}, v^{\varepsilon_k})$ converges in distribution. For notational convenience, we index this subsequence once more by ε. Denoting the weak limit of $(\bar{X}^\varepsilon, v^\varepsilon)$ by $(\bar{X}, v)$, we have from Lemma 3.21 that $\bar{X}$ is the unique solution of (3.20). Therefore

$$\begin{aligned}
&\liminf_{\varepsilon\to 0} -\varepsilon \log E \exp\left\{-\frac{1}{\varepsilon}F(X^\varepsilon)\right\} + 2\delta \\
&\quad \geq \liminf_{\varepsilon\to 0} E\left[F(\bar{X}^\varepsilon) + \frac{1}{2}\int_0^1 \|v^\varepsilon(s)\|^2\, ds\right] \\
&\quad \geq E\left[F(\bar{X}) + \frac{1}{2}\int_0^1 \|v(s)\|^2\, ds\right] \\
&\quad \geq \inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [F(\varphi) + I(\varphi)]\,.
\end{aligned}$$

Here the second inequality is a consequence of Fatou's lemma and the lower semicontinuity of the map $\phi \mapsto \int_0^1 \|\phi(s)\|^2\, ds$ from $\mathscr{L}^2([0,1]:\mathbb{R}^d)$ to $\mathbb{R}$ with the weak topology on $\mathscr{L}^2([0,1]:\mathbb{R}^d)$. Recalling the definition of U_φ in (3.16), the last inequality follows from the a.s. inequality

$$\begin{aligned}
F(\bar{X}) + \frac{1}{2}\int_0^1 \|v(s)\|^2\, ds &\geq F(\bar{X}) + \inf_{u\in U_{\bar{X}}}\left[\frac{1}{2}\int_0^1 \|u(s)\|^2\, ds\right] \\
&\geq \inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [F(\varphi) + I(\varphi)]\,.
\end{aligned}$$

Since $\delta > 0$ is arbitrary, (3.21) follows. □

3.2.4 *Compactness of Level Sets*

We now argue that I introduced in Theorem 3.19 is a rate function, which requires that we show that it has compact level sets. As we will see, it is essentially just a deterministic version of the argument used for the Laplace upper bound (variational lower bound). This is in fact generic in the weak convergence approach to large deviations and not at all surprising, in that the main difference between these two arguments is that the variational lower bound has the additional complication of a law of large numbers limit as the large deviation parameter tends to its limit, an item missing in the corresponding and purely deterministic analysis of the rate function.

Let $M \in (0,\infty)$ and let $\{\varphi_n\} \subset \mathscr{C}([0,1]:\mathbb{R}^d)$ be a sequence such that $I(\varphi_n) \le M$ for all $n \in \mathbb{N}$. Choose $u_n \in U_{\varphi_n}$ such that $\frac{1}{2}\int_0^1 \|u_n(s)\|^2\,ds \le M + 1/n$. Then the sequence $\{u_n\}$ is contained in the (weakly) compact set $S_{2(M+1)}$. Let u be a limit point of u_n along some subsequence. Then $\frac{1}{2}\int_0^1 \|u(s)\|^2\,ds \le M$. Also, a simpler version of an argument in the proof of Lemma 3.21 shows that along the same subsequence, $\varphi_n(\cdot)$ converges to $\varphi(\cdot)$, where φ is the unique solution of $\varphi(t) = x + \int_0^t (b(\varphi(s)) + \sigma(\varphi(s))u(s))\,ds$. In particular, $u \in U_\varphi$ and thus $I(\varphi) \le M$. This proves the compactness of level sets of I. □

3.2.5 *Laplace Lower Bound*

To prove the Laplace lower bound we use the variational representation to show that

$$\limsup_{\varepsilon\to 0} -\varepsilon \log E \exp\left\{-\frac{1}{\varepsilon}F(X^\varepsilon)\right\} \le \inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [F(\varphi) + I(\varphi)].$$

For $\delta > 0$ choose $\varphi^* \in \mathscr{C}([0,1]:\mathbb{R}^d)$ such that

$$F(\varphi^*) + I(\varphi^*) \le \inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [F(\varphi) + I(\varphi)] + \delta.$$

Let $u \in U_{\varphi^*}$ be such that $\frac{1}{2}\int_0^1 \|u(s)\|^2 ds \le I(\varphi^*) + \delta$, so that in particular, $u \in \mathscr{A}_{b,2(I(\varphi^*)+\delta)}$. Let $\bar{X}^\varepsilon$ be the unique solution of (3.17) when we replace v on the right side of the equation by u. By Lemmas 3.20 and 3.21 on tightness and weak convergence, $\bar{X}^\varepsilon$ converges in probability to φ^*. Thus

$$\begin{aligned}
&\limsup_{\varepsilon\to 0} -\varepsilon \log E e^{-\frac{1}{\varepsilon}F(X^\varepsilon)} \\
&\quad= \limsup_{\varepsilon\to 0} \inf_{v\in\mathscr{A}} E\left[F\circ\mathscr{G}^\varepsilon(\sqrt{\varepsilon}W^{v/\sqrt{\varepsilon}}) + \frac{1}{2}\int_0^1 \|v(s)\|^2 ds\right] \\
&\quad\le \limsup_{\varepsilon\to 0} E\left[F\left(\bar{X}^\varepsilon\right) + \frac{1}{2}\int_0^1 \|u(s)\|^2 ds\right] \\
&\quad= F\left(\varphi^*\right) + \frac{1}{2}\int_0^1 \|u(s)\|^2 ds \\
&\quad\le F\left(\varphi^*\right) + I\left(\varphi^*\right) + \delta \\
&\quad\le \inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [F(\varphi) + I(\varphi)] + 2\delta.
\end{aligned}$$

Since $\delta > 0$ is arbitrary, the upper bound follows. □

Remark 3.22 One can consider the form

$$I(\varphi) \doteq \inf_{u\in U_\varphi}\left[\frac{1}{2}\int_0^T \|u(t)\|^2 dt\right]$$

of the rate function, where U_φ are those u satisfying $\varphi(t) = x + \int_0^t b(\varphi(s))ds + \int_0^t \sigma(\varphi(s))u(s)ds$, as a "control" formulation. If $\sigma(x)$ is $d \times d$ and invertible for all $x \in \mathbb{R}^d$, then one can solve for u and obtain the calculus of variations form

$$I(\varphi) \doteq \int_0^T \frac{1}{2} \left\langle (\dot{\varphi}(t) - b(\varphi(t))), [\sigma\sigma^T(\varphi(t))]^{-1}(\dot{\varphi}(t) - b(\varphi(t))) \right\rangle dt,$$

where σ^T is the transpose of σ.

3.3 Representation for Functionals of a Poisson Process

Our final example in this chapter is the representation for positive functionals of a Poisson process. This example will be substantially generalized in Chap. 8, where we prove the representation for a Poisson random measure (PRM) on an arbitrary locally compact Polish space. The representation for a PRM allows the treatment of a much broader class of process models, and in particular when used as a driving noise, a PRM can easily accommodate both state-dependent jump rates and jumps sizes, while a Poisson process (which is essentially a PRM with only one "type" of point) is limited to state dependence of jump sizes. However, the purpose of this chapter is to illustrate the use of representations, and we prefer to postpone the notation and terminology required for the general case of a PRM.

Fix $T \in (0, \infty)$ and let $(\Omega, \mathscr{F}, P)$ be a probability space with filtration $\{\mathscr{F}_t\}_{0 \le t \le T}$ satisfying the usual conditions. Recall that $\mathscr{D}([0, T] : \mathbb{R})$ is the space of functions from $[0, T]$ to $\mathbb{R}$ that are right continuous and with limits from the left at each $t \in (0, T]$. As noted in Chap. 2, there is a metric that is consistent with the usual Skorohod topology that makes this a Polish space [24, Chap. 3, Sect. 12]. An **$\mathscr{F}_t$-Poisson process** is a measurable mapping N from Ω into $\mathscr{D}([0, T] : \mathbb{R})$ such that $N(t)$ is $\mathscr{F}_t$-measurable for every $t \in [0, T]$, and for all $0 \le t < s \le T$, $N(s) - N(t)$ is independent of $\mathscr{F}_t$ and has a Poisson distribution with parameter $s - t$: $P(N(s) - N(t) = j) = (s - t)^j e^{-(s-t)}/j!$. We say that such a standard Poisson process has **jump intensity** or **jump rate** 1, since the probability that $N(s) - N(t) = 1$ is approximately $s - t$ when this difference is small [and the probability of more than one jump is $o(s - t)$].

In contrast to the case of Brownian motion, in which the natural controlled version shifts the mean, here the controlled version will shift the jump rate and pay the appropriate cost suggested by Girsanov's theorem for Poisson processes (see, for example, Theorem 8.15). There are various ways to construct Poisson processes with general jump rates on a common probability space. The most convenient one requires the use of a PRM on the space $[0, T] \times [0, \infty)$ and with intensity measure equal to Lebesgue measure on this space (see Chap. 8 for definitions and associated terminology). In this framework the Poisson process on $[0, T]$ is considered a PRM

on $[0, T]$, and to accommodate general controls we suitably enlarge the space. We do not give the details here, but instead just state the outcome of this construction.

One can construct a probability space $\left(\bar{\Omega}, \bar{\mathscr{F}}, \bar{P}\right)$, and on this space a filtration $\{\bar{\mathscr{F}}_t\}_{0\le t\le T}$ satisfying the usual conditions, such that the following properties hold. Let $\theta \in (0, \infty)$ (later θ will play the role of a large deviation parameter). Denote by $\mathscr{A}$ the collection of predictable processes $\varphi : [0, T] \times \bar{\Omega} \to [0, \infty)$ (see Definition 8.2 for the definition of predictability in a general setting) such that $\int_0^T \varphi(s)ds < \infty$ a.s. Predictable processes are in a suitable way not allowed to anticipate the jumps of a Poisson process with respect to the same filtration, and hence are the appropriate analogue of the class of controls used for representations in discrete time. Associated with each $\varphi \in \mathscr{A}$ one can construct a "controlled" Poisson process $N^{\theta\varphi}$ with jump intensity $\theta\varphi$ and jump size 1. To be precise, $N^{\theta\varphi}$ is an $\bar{\mathscr{F}}_t$-adapted stochastic process with trajectories in $\mathscr{D}([0, T] : \mathbb{R})$ such that for every bounded function $f : [0, \infty) \to [0, \infty)$,

$$f(N^{\theta\varphi}(t)) - f(0) - \theta \int_0^t \varphi(s) \left[f(N^{\theta\varphi}(s) + 1) - f(N^{\theta\varphi}(s)) \right] ds$$

is an $\bar{\mathscr{F}}_t$-martingale, and $N^{\theta\varphi}(0) = 0$. Note that N^{θ} is an ordinary Poisson process with constant jump intensity θ and jump size 1.

In terms of these controls and controlled processes, we have the following representation. Recall the function ℓ introduced in (3.10): for $r \in [0, \infty)$, we have $\ell(r) \doteq r \log r - r + 1$, with the convention that $0 \log 0 = 0$. We consider all processes $N^{\theta\varphi}$ to be random variables with values in $\mathscr{D}([0, T] : \mathbb{R})$. We also introduce

$$S_M \doteq \left\{ \phi \in \mathscr{L}^0([0, T] : \mathbb{R}_+) : \int_0^T \ell(\phi(s))ds \le M \right\},$$

where $\mathscr{L}^0([0, T] : \mathbb{R}_+)$ denotes the space of Borel-measurable functions from $[0, T]$ to $[0, \infty)$, and given $M \in (0, \infty)$ define $\mathscr{A}_{b,M}$ to be the subset of $\mathscr{A}$ such that $\varphi \in \mathscr{A}_{b,M}$ implies $\varphi(\omega) \in S_M$ for all $\omega \in \bar{\Omega}$ and for some $K \in (0, \infty)$ (possibly depending on φ), $K^{-1} \le \varphi \le K$, a.s. Also, let $\mathscr{A}_b = \cup_{M=1}^{\infty} \mathscr{A}_{b,M}$. The spaces S_M, $\mathscr{A}_{b,M}$, $\mathscr{A}_b$ in this section play an analogous role for Poisson processes to that of the corresponding spaces introduced in Sect. 3.2 for the Brownian motion case.

Theorem 3.23 *Let G be a bounded Borel measurable function mapping $\mathscr{D}([0, T] : \mathbb{R})$ into $\mathbb{R}$ and let $\theta \in (0, \infty)$. Then*

$$-\log E \exp\{-G(N^{\theta})\} = \inf_{\varphi\in\mathscr{A}} E\left[G\left(N^{\theta\varphi}\right) + \theta \int_0^T \ell(\varphi(s))ds \right].$$

If $\delta > 0$, then there exists $M < \infty$ depending on $\|G\|_\infty$ and δ such that for all $\theta \in (0, \infty)$,

$$-\frac{1}{\theta}\log E\exp\left\{-\theta G(N^{\theta})\right\} \geq \inf_{\varphi\in\mathscr{A}_{b,M}} E\left[G\left(N^{\theta\varphi}\right)+\int_0^T \ell(\varphi(s))ds\right]-\delta. \quad (3.22)$$

The proof of Theorem 3.23 follows as a special case of more general results [Theorems 8.12 and 8.13] that will be proved in Chap. 8. In particular, the general result will show that $\mathscr{A}$ can be replaced by $\mathscr{A}_b$ in the first representation. We now show how this representation can be used to obtain a large deviation principle for SDEs driven by a Poisson process. We begin with a condition on the coefficients that can be relaxed substantially (see, for example, Chap. 10).

Condition 3.24 *There is $C \in (0,\infty)$ such that $b : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ satisfy*

$$|b(x)-b(y)|+|\sigma(x)-\sigma(y)| \leq C\,|x-y| \quad \text{and} \quad |b(x)|+|\sigma(x)| \leq C$$

for all $x, y \in \mathbb{R}$.

Fix $x \in \mathbb{R}$, and for $n \in \mathbb{N}$ let $X^n = \{X^n(t)\}_{0\leq t\leq T}$ be the pathwise solution of the SDE

$$dX^n(t) = b(X^n(t))dt + \frac{1}{n}\sigma(X^n(t-))dN^n(t),\ \ X^n(0) = x, \quad (3.23)$$

where $X^n(t-)$ denotes the limit from the left. One can explicitly construct the solution in terms of the jump times $\{t_i^n\}_{i\in\mathbb{N}}$ of $N^n(\cdot)$. With probability 1, these jump times satisfy $0 < t_1^n < t_2^n < \cdots$ and $t_i^n \to \infty$. Letting $t_0^n = 0$ and $X^n(t_0^n) = x$, we then recursively define $X^n(t)$ as follows. Assuming that $X^n(t_i^n)$ is given, let

$$\dot{X}^n(t) = b(X^n(t)) \text{ for } t \in (t_i^n, t_{i+1}^n)$$

and then set $X^n(t_{i+1}^n) \doteq X^n(t_{i+1}^n-) + \sigma(X^n(t_{i+1}^n-))/n$. With $X^n(t_{i+1}^n)$ now given, we repeat the procedure, and since $t_i^n \to \infty$, the construction on $[0,T]$ is well defined.

For $\psi \in \mathscr{AC}_x([0,T]:\mathbb{R})$, let

$$U_\psi = \left\{\gamma \in \mathscr{L}^1([0,T]:\mathbb{R}_+) : \psi(\cdot) = x + \int_0^{\cdot} b(\psi(s))ds + \int_0^{\cdot} \sigma(\psi(s))\gamma(s)ds\right\}, \quad (3.24)$$

where $\mathscr{L}^1([0,T]:\mathbb{R}_+)$ is the space of $\mathbb{R}_+$-valued integrable functions on $[0,T]$.

Theorem 3.25 *Assume Condition 3.24. Then the collection $\{X^n\}_{n\in\mathbb{N}}$ satisfies the LDP on $\mathscr{D}([0,T]:\mathbb{R})$ with rate function*

$$I(\psi) \doteq \inf_{\gamma\in U_\psi}\left[\int_0^T \ell(\gamma(t))dt\right].$$

The proof of this theorem is a close parallel to that of Brownian motion, and because of this we do not separate the proof into a series of statements (lemmas, propositions, etc.) and their proofs. We must show that I is a rate function and that for every bounded and continuous $F : \mathscr{D}([0,T]:\mathbb{R}) \to \mathbb{R}$,

$$\lim_{n\to\infty} -\frac{1}{n}\log E\exp\left\{-nF(X^n)\right\} = \inf_{\psi\in\mathscr{D}([0,T]:\mathbb{R})}\left[F(\psi)+I(\psi)\right].$$

Following our convention, we consider just the case $T = 1$. We have already explicitly identified the measurable map $\mathscr{G}^n : \mathscr{D}([0,1]:\mathbb{R}) \to \mathscr{D}([0,1]:\mathbb{R})$ such that whenever $\tilde{N}^n$ is a Poisson process with rate n on some probability space $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{P})$, then $\tilde{X}^n = \mathscr{G}^n(\tilde{N}^n)$ is the unique solution of the SDE (3.23) with N^n replaced by $\tilde{N}^n$. Hence by Theorem 3.23 with $\theta = n$,

$$\begin{aligned}-\frac{1}{n}\log E\exp\left\{-nF(X^n)\right\} &= -\frac{1}{n}\log E\exp\left\{-nF\circ\mathscr{G}^n(N^n)\right\}\\ &= \inf_{\varphi\in\mathscr{A}} E\left[F\circ\mathscr{G}^n(N^{n\varphi}) + \int_0^1 \ell(\varphi(t))dt\right].\end{aligned}$$

Analogous to the case of Brownian motion, if $\varphi \in \mathscr{A}_{b,M}$ for some $M < \infty$, then $\bar{X}^n = \mathscr{G}^n(N^{n\varphi})$ is the solution of the SDE

$$d\bar{X}^n(t) = b(\bar{X}^n(t))dt + \sigma(\bar{X}^n(t-))dN^{n\varphi}(t),\ \ \bar{X}^n(0) = x. \tag{3.25}$$

Here, the important property that follows from $\varphi \in \mathscr{A}_{b,M}$ is that it guarantees (as easily follows from Girsanov's formula) that the jump times of $N^{n\varphi}$ tend to ∞ w.p.1, and so the recursive construction of $\bar{X}^n$ is well defined on $[0,1]$.

A distinction with respect to the case of Brownian motion is that it is no longer appropriate to consider S_M as a subset of a Hilbert space. Instead, we will identify S_M with a compact space of measures. In particular, associated with each element γ of S_M is a measure ν^γ on $([0,1], \mathscr{B}([0,1]))$ defined by $\nu^\gamma(ds) \doteq \gamma(s)m(ds)$, where m denotes Lebesgue measure. As discussed in Lemma A.11, when considered with the natural generalization of the weak topology from probability measures to measures with finite total measure, S_M is a compact Polish space.

Next suppose that Condition 3.24 holds. Consider any collection of controls $\{\varphi^n\} \subset \mathscr{A}_{b,M}$ for fixed $M < \infty$, and define $\bar{X}^n$ by (3.25) with $\varphi = \varphi^n$. We claim that $\{(\bar{X}^n, \varphi^n)\}_{n\in\mathbb{N}}$ is a tight collection of $\mathscr{D}([0,1]:\mathbb{R}) \times S_M$-valued random variables. Tightness of $\{\varphi^n\}$ follows from the compactness of S_M. For the tightness of $\{\bar{X}^n\}$, we consider the Doob decomposition

$$\begin{aligned}\bar{X}^n(t) - x &= \int_0^t b(\bar{X}^n(s))ds + \frac{1}{n}\int_0^t \sigma(\bar{X}^n(s-))dN^{n\varphi^n}(s)\\ &= \int_0^t b(\bar{X}^n(s))ds + \int_0^t \sigma(\bar{X}^n(s))\varphi^n(s)ds\\ &\quad + \int_0^t \sigma(\bar{X}^n(s-))[dN^{n\varphi^n}(s)/n - \varphi^n(s)]ds.\end{aligned} \tag{3.26}$$

Since the restriction of the Skorohod metric to $\mathscr{C}([0,1]:\mathbb{R})$ is equivalent to the standard uniform metric, it suffices, for the first two terms, to show tightness in

$\mathscr{C}([0,1]:\mathbb{R})$. Tightness of the first follows from $\|b\|_\infty \le C$. For the second term we use the bound $ab \le e^{ca} + \ell(b)/c$, valid for $a \ge 0$, $b \ge 0$ and $c \ge 1$ [see (2.9)]. For all $0 \le s \le t \le 1$,

$$\int_s^t \sigma(\bar{X}^n(r))\varphi^n(r)dr \le \int_s^t [e^{c\|\sigma\|_\infty} + \ell(\varphi^n(r))/c]dr \le (t-s)e^{c\|\sigma\|_\infty} + \frac{1}{c}M.$$

This shows equicontinuity of the second term in (3.26) that is uniform in ω, and tightness of that term follows. Let $Q^n(t)$ denote the third term. This term is a martingale with quadratic variation (see Sects. D.1 and D.2.2) $[Q^n]_t$ bounded above by

$$\frac{1}{n^2}\|\sigma\|_\infty^2 E N^{n\varphi^n}(1) = \frac{1}{n}\|\sigma\|_\infty^2 E\int_0^1 \varphi^n(s)ds \le \frac{1}{n}\|\sigma\|_\infty^2 (e+M),$$

where $b \le e + \ell(b)$ is used for the last inequality. By the Burkholder–Gundy–Davis inequality (see (D.3) in Sect. D.1), $E\sup_{t\in[0,1]}|Q^n(t)| \le C_1 E[Q^n]_1^{1/2} \to 0$ for some $C_1 \in (0,\infty)$. Thus by Chebyshev's inequality, Q^n converges weakly to zero uniformly in t, which both shows tightness and identifies the limit. Since all three terms on the right-hand side of (3.26) are tight (and limit points are continuous a.s.), so is $\{\bar{X}^n\}$.

To identify weak limits along any convergent subsequence, we need to know that if $\gamma_n \to \gamma$ in S_M and $\psi_n \to \psi$ uniformly, then

$$\int_0^t \sigma(\psi_n(s))\gamma_n(s)ds \to \int_0^t \sigma(\psi(s))\gamma(s)ds. \tag{3.27}$$

Again using $b \le e + \ell(b)$, we have

$$\begin{aligned}\left|\int_0^t [\sigma(\psi_n(s)) - \sigma(\psi(s))]\gamma_n(s)ds\right| &\le \sup_{s\in[0,1]}|\sigma(\psi_n(s)) - \sigma(\psi(s))|\int_0^t \gamma_n(s)ds \\ &\le \sup_{s\in[0,1]}|\sigma(\psi_n(s)) - \sigma(\psi(s))|\,(e+M) \\ &\to 0\end{aligned}$$

as $n \to \infty$. To show that

$$\int_0^t \sigma(\psi(s))[\gamma_n(s) - \gamma(s)]ds \to 0,$$

we use that $\nu^{\gamma_n}(ds) \doteq \gamma_n(s)m(ds)$ converges in the weak topology to $\nu^\gamma(ds)$. Since $s \mapsto 1_{[0,t]}(s)\sigma(\psi(s))$ is bounded and discontinuous only at $s = t$ and $\nu^\gamma(\{t\}) = 0$, the last display is valid, and this completes the proof of (3.27).

Consider any subsequence of $\{(\bar{X}^n, \varphi^n)\}_{n\in\mathbb{N}}$ that converges in distribution with limit $(\bar{X}, \varphi)$. Sending $n \to \infty$ in (3.26) and using (3.27) establishes the w.p.1 relation

$$\bar{X}(t) - x = \int_0^t b(\bar{X}(s))ds + \int_0^t \sigma(\bar{X}(s))\varphi(s)ds. \tag{3.28}$$

The rest of the proof is now essentially identical to that for Brownian motion. For the Laplace upper bound, we need to show that

$$\liminf_{n\to\infty} -\frac{1}{n} \log E \exp\left\{-nF(X^n)\right\} \geq \inf_{\psi\in\mathscr{D}([0,1]:\mathbb{R})} [F(\psi) + I(\psi)].$$

Let $\delta > 0$, choose M according to Theorem 3.23, and choose a sequence $\{\varphi^n\} \subset \mathscr{A}_{b,M}$ that is within δ of the infimum in the representation (3.22) (with G replaced by $F \circ \mathscr{G}^n$ and θ replaced by n). Fix any subsequence of n and choose a further subsequence (again denoted by n) along which $(\bar{X}^n, \varphi^n)$ converges in distribution to $(\bar{X}, \varphi)$. Then

$$\begin{aligned}
&\liminf_{n\to\infty} -\frac{1}{n} \log E \exp\left\{-nF(X^n)\right\} + 2\delta \\
&\quad \geq \liminf_{n\to\infty} E\left[F(\bar{X}^n) + \int_0^1 \ell(\varphi^n(s))ds\right] \\
&\quad \geq E\left[F(\bar{X}) + \int_0^1 \ell(\varphi(s))ds\right] \\
&\quad \geq \inf_{\psi\in\mathscr{D}([0,1]:\mathbb{R})} [F(\psi) + I(\psi)],
\end{aligned}$$

where the second inequality uses Fatou's lemma and the lower semicontinuity of the map $\varphi \mapsto \int_0^1 \ell(\varphi(s))ds$ from S_M to $[0, \infty)$. Recalling the definition of U_ψ in (3.24), the last inequality is a consequence of the a.s. inequality

$$\begin{aligned}
F(\bar{X}) + \int_0^1 \ell(\varphi(s))ds &\geq F(\bar{X}) + \inf_{\varphi\in U_{\bar{X}}} \left[\int_0^1 \ell(\varphi(s))ds\right] \\
&\geq \inf_{\psi\in\mathscr{D}([0,1]:\mathbb{R})} [F(\psi) + I(\psi)].
\end{aligned}$$

Since $\delta > 0$ is arbitrary, the Laplace upper bound follows.

As in Sect. 3.2.4, a deterministic version of the argument used for the Laplace upper bound gives the compactness of level sets for the rate function, and so this argument is omitted. To complete the proof, all that remains is the Laplace lower bound, which requires that for bounded and continuous F,

$$\limsup_{n\to\infty} -\frac{1}{n} \log E \exp\left\{-nF(X^n)\right\} \leq \inf_{\psi\in\mathscr{D}([0,1]:\mathbb{R})} [F(\psi) + I(\psi)].$$

For $\delta > 0$ choose $\psi^* \in \mathscr{D}([0, 1] : \mathbb{R})$ such that

$$F(\psi^*) + I(\psi^*) \leq \inf_{\psi\in\mathscr{D}([0,1]:\mathbb{R})} [F(\psi) + I(\psi)] + \delta.$$

Let $\varphi \in U_{\psi^*}$ be such that $\int_0^1 \ell(\varphi(s))ds \le I(\psi^*) + \delta$. We now approximate φ with an element in $\mathscr{A}_{b,M}$, where $M = I(\psi^*) + \delta$. For $q \in \mathbb{N}$ let

$$\varphi_q(t) = \left(\varphi(t) \vee \frac{1}{q}\right) \wedge q.$$

Then $\varphi_q \in \mathscr{A}_{b,M}$ and $\int_0^1 \ell(\varphi_q(s))ds \uparrow \int_0^1 \ell(\varphi(s))ds$ as $q \to \infty$. Let ψ_q^* be the solution of (3.28) with φ replaced by φ_q. It is easily seen that $\psi_q^* \to \psi^*$ in $\mathscr{C}([0,1] : \mathbb{R})$ as $q \to \infty$. Let $\bar{X}^n$ be the unique solution of (3.25) with φ replaced by φ_q. The tightness of $(\bar{X}^n, \varphi_q)$ and identification of limits is exactly as in the proof of the Laplace upper bound, since $\varphi_q \in \mathscr{A}_{b,M}$. Using the uniqueness of solutions to the limit ordinary differential equation (ODE), $\bar{X}^n$ converges in probability to ψ_q^*. Thus

$$\begin{aligned}\limsup_{n\to\infty} -\frac{1}{n} \log E e^{-nF(X^n)} &\le \limsup_{n\to\infty} E\left[F\left(\bar{X}^n\right) + \int_0^1 \ell(\varphi_q(s))ds\right] \\ &= F\left(\psi_q^*\right) + \int_0^1 \ell(\varphi_q(s))ds.\end{aligned}$$

Sending $q \to \infty$, we now have

$$\begin{aligned}\limsup_{n\to\infty} -\frac{1}{n} \log E e^{-nF(X^n)} &\le F\left(\psi^*\right) + \int_0^1 \ell(\varphi(s))ds \\ &\le F\left(\psi^*\right) + I\left(\psi^*\right) + \delta \\ &\le \inf_{\psi \in \mathscr{D}([0,1]:\mathbb{R})} [F(\psi) + I(\psi)] + 2\delta.\end{aligned}$$

Since $\delta > 0$ is arbitrary, the upper bound follows, thus completing the proof of Theorem 3.25. □

3.4 Notes

Our treatment of Sanov's theorem [228] follows very closely the one in [97], though as noted in the introduction we use the representation based on the chain rule rather than that based on dynamic programming. The proof of Cramér's theorem [68] differs from that of [97] and follows a line of argument that will be used elsewhere, which is that to analyze discrete time "small noise" problems, we first establish an empirical-measure-type large deviation result for the driving noises, and then (in combination with integrability properties of the noise) obtain the large deviation properties of a process driven by these noises through a continuous-mapping-type argument.

The proof of large deviation estimates for small noise diffusions is taken from [32], and it suggests why the more highly structured setting of continuous time Markov processes is, given the appropriate representations, easier than for discrete time. The

solution mapping to the SDE (3.15) is not continuous, since if it were, we would just use the contraction principle and the large deviation theory for scaled Brownian motion (Schilder's theorem [229]). However, it is in some sense almost continuous on the support of the measure induced by $\sqrt{\varepsilon}W$, in that the mapping $u \mapsto \varphi$ when $\varphi(t) = x + \int_0^t (b(\varphi(s)) + \sigma(\varphi(s))u(s))\,ds$ is continuous on S_M for all $M < \infty$, a fact that was key in proving the convergence of the variational representation. An analogous continuity applies only to particular models in the setting of discrete time. In particular, the reader will note that the arguments of Chap. 4 are considerably more involved than those for SDEs in continuous time.

The idea of viewing a diffusion as a nearly continuous mapping on Brownian motion (in the small noise limit) originates with Azencott [7]. The first proofs of an LDP for diffusions appear in the papers of Wentzell [245–248], where they appear as just a special case of a more general treatment. Fleming [133] considers certain problems of large deviations involving diffusion process and computes the desired limits using ideas from stochastic control. His approach is closely related to the approach of this book and in many ways inspired it.

In the final example of an SDE driven by a Poisson process we have attempted to emphasize the similarity with the case of Brownian motion, and indeed, the arguments are very close, with the main differences due to the weaker control one obtains from bounded costs and the need to place the controls in a space more complicated than $\mathscr{L}^2([0, T] : \mathbb{R}^k)$ with the weak topology. This example is a simplified form of the problem considered in [45].

Part II
Discrete Time Processes

In the last chapter we considered two basic examples of large deviation theory for discrete time processes. One example was the empirical measure for iid random variables in a Polish space S, and the other was the sample mean for the case $S = \mathbb{R}^d$. In the next part of the book we will generalize these two examples in various directions. In all cases, the starting point will be a representation obtained using the chain rule for relative entropy, though the underlying noise models will be more complex than the simple product measures of Chap. 3. One generalization will be relatively straightforward, which is the extension from the empirical measure for iid to the empirical measure for a Markov chain in Chap. 6.

A second generalization will be a "process level" and "state dependent" generalization of Cramér's theorem in Chap. 4. Here we will consider the large deviation principle for a very general stochastic recursive model. A scaling is introduced that makes explicit that the system can be thought of as a small random perturbation of an ordinary differential equation, and indeed included would be models such as the Euler–Maruyama approximation to an SDE with small noise. Our perspective here will be very similar to the one used in Chap. 3 to obtain Cramér's theorem from Sanov's theorem, and in fact, we will prove large deviation properties of a "time-dependent" empirical measure for the driving noises and then view the state of the stochastic recursive system as a mapping on this empirical measure. Under suitable integrability assumptions, the passage from the large deviation of the empirical measure to that of the process will parallel that taking us from Sanov's theorem to Cramér's theorem.

Chapter 7 focuses on models with various special features. One class comprises discrete time dynamical models for occupancy-type problems. These are classical models from combinatorial probability (e.g., the coupon collector's problem), and the particular feature that makes their analysis different is that the probabilities of certain types of jumps tend to zero as a boundary is approached. This "diminishing rates" feature puts them outside the models of Chap. 4, and in fact, a careful construction is needed to establish the large deviation lower bound for trajectories that touch the boundary. On the other hand, as also discussed in this chapter, these models and many related generalizations have the feature that the variational problems one needs to solve to extract information from the rate function on path space have nearly explicit solutions, i.e., solutions that can be identified by solving a low-dimensional

constrained convex optimization problem. Also included in this chapter is the formulation of a two-time-scale model and the statement of the corresponding LDP. The proof of the LDP for this model, which combines arguments used in Chaps. 4 and 6, is omitted.

Moderate deviations for the same class of small noise Markov processes as in Chap. 4 are the topic of Chap. 5. What is meant by "moderate deviations" is approximations for events that are closer to the LLN limit then those approximated via standard large deviations. While the starting point is the same relative entropy representation as in Chap. 4, moderate deviations (which can be phrased as large deviations for a suitably centered and rescaled system) are in some ways simpler but in other ways more difficult than the corresponding large deviations. As discussed at some length in Chap. 5, a particular motivation for the moderate deviation approximation is the development of accelerated Monte Carlo schemes for this same class of events. An example of such will be given in Sect. 17.5.

Chapter 4
Recursive Markov Systems with Small Noise

In Chap. 3 we presented several examples of representations and how they could be used for large deviation analysis. A simplifying feature of all the examples of Chap. 3 is that the process models (e.g., empirical measure, solution to an SDE) could be thought of as a "nice" functional of a process that is "white" in the time variable, by which we mean independent in the setting of discrete time, and with independent increments in the setting of continuous time (see Sect. 3.5 for what is meant by a nice functional in the case of small noise SDEs).

In this chapter we study a model for which there is, in general, no convenient representation as a functional of white noise. Note that we do not claim that such a representation is impossible, but rather that it will not (in general) be useful, e.g., in proving law of large numbers limits. Because of this feature, a more complex representation and weak convergence analysis cannot be avoided. In particular, the "base" measure in the representation will be a Markov measure rather than a product measure, and the process model will be a general "small noise" Markov process. The model provides a substantial generalization of the random walk considered in Cramér's theorem. It occurs frequently in stochastic systems theory, e.g., stochastic approximation and related recursive algorithms [18, 182, 193], where the rate function can be used to define a rate of convergence [102]. The model also arises as a discrete time approximation to various continuous time models, such as the small noise SDE in Sect. 3.2.1 of Chap. 3, and indeed provides an alternative approach to proving large deviation estimates for such models (though we much prefer the direct approach of Chap. 3).

4.1 Process Model

We begin with a description of the process model. Suppose that $\theta(dy|x)$ is a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d$. One can construct a probability space that supports **iid random vector fields** $\{v_i(x), i \in \mathbb{N}_0, x \in \mathbb{R}^d\}$, with the property that for all

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_4

$x \in \mathbb{R}^d$, $v_i(x)$ has distribution $\theta(\cdot|x)$. To be precise, there exists a probability space $(\Omega, \mathscr{F}, P)$ such that for each $i \in \mathbb{N}_0$, v_i is a measurable map from $\mathbb{R}^d \times \Omega$ to $\mathbb{R}^d$; for $k \in \mathbb{N}$ and distinct $i_1, \ldots, i_k \in \mathbb{N}_0$ and $x_{i_1}, \ldots, x_{i_k} \in \mathbb{R}^d$, the random vectors $v_{i_1}(x_{i_1}), \ldots, v_{i_k}(x_{i_k})$ are mutually independent; and for each $i \in \mathbb{N}_0$, $v_i(x)$ has distribution $\theta(\cdot|x)$. We then define for each $n \in \mathbb{N}$ a Markov process $\left\{X_i^n\right\}_{i=0,\ldots,n}$ by setting

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n), \quad X_0^n = x_0. \tag{4.1}$$

This discrete time process is interpolated into continuous time according to

$$X^n(t) = X_i^n + \left[X_{i+1}^n - X_i^n\right](nt - i), \quad t \in [i/n, i/n + 1/n]. \tag{4.2}$$

The goal of this chapter is to study a large deviation principle for the sequence $\{X^n\}_{n\in\mathbb{N}}$ of $\mathscr{C}([0, T] : \mathbb{R}^d)$-valued random variables.

Example 4.1 Suppose that for each $x \in \mathbb{R}^d$, $v_i(x)$ has a normal distribution with continuous mean $b(x)$ and covariance $\sigma(x)\sigma^T(x)$. Then $X^n(t)$ is the Euler approximation with step size $1/n$ to the SDE (3.15) with drift coefficient b, diffusion coefficient σ, and $\varepsilon = 1/n$.

Example 4.2 For an example in the form of a stochastic approximation algorithm, take $v_i(x) = -\nabla V(x) + w_i$, where the w_i are iid with $Ew_i = 0$ and V is a smooth function. In this case, $1/n$ is the "gain" of the algorithm [18, 182].

Of course, to prove an LDP for $\{X^n\}_{n\in\mathbb{N}}$, additional assumptions must be made. For $x \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}^d$, define

$$H(x, \alpha) \doteq \log E e^{\langle \alpha, v_i(x)\rangle}.$$

Condition 4.3 (a) For each $\alpha \in \mathbb{R}^d$, $\sup_{x\in\mathbb{R}^d} H(x, \alpha) < \infty$.

(b) The mapping $x \mapsto \theta(\cdot|x)$ from $\mathbb{R}^d$ to $\mathscr{P}(\mathbb{R}^d)$ is continuous in the topology of weak convergence.

The first condition is not needed for an LDP to hold. However, if $H(x, \alpha) = \infty$ for some values of x and α, then $v_i(x)$ has relatively heavy tails in certain directions. Paths with jumps may be important from the perspective of large deviations, and the setting used here is no longer appropriate. The second condition can also be weakened. However, this often leads to a qualitatively different form of the rate function, and the process models that violate this condition are said to have "discontinuous statistics" [95, 98]. For an example of such a process but in continuous time, see Chap. 13.

4.2 The Representation

The first issue to resolve is the formulation of a representation that reflects the natural structure of the process model. As noted at the beginning of the chapter, it is possible to represent $\{X^n\}$ in terms of iid random variables, e.g., in the form $X^n_{i+1} = X^n_i + \frac{1}{n} g(X^n_i, U_i)$, where g is measurable and the $\{U_i, i \in \mathbb{N}_0\}$ are iid random variables with uniform distribution on $[0, 1]$. Although this form would allow a representation in terms of an iid base measure, it would not be useful. This is because the map g is not in general continuous in x, and hence this formulation is poorly suited for even a law of large numbers analysis.

An alternative and more useful representation follows from the form (4.1) and the continuity of $x \mapsto \theta(\cdot|x)$. Following our convention, we present only the representation needed to prove an LDP on $\mathscr{C}([0, 1] : \mathbb{R}^d)$, but the analogous representation holds with $[0, 1]$ replaced by any interval $[0, T]$, $T < \infty$. The line of argument used to prove the LDP will adapt the arguments used for Sanov's theorem and Cramér's theorem to this functional setting. However, obtaining "process-level" information requires a more complicated empirical measure than the one used for Sanov's theorem. Define L^n by

$$L^n(A \times B) \doteq \int_B L^n(A|t)dt, \quad L^n(A|t) \doteq \delta_{v_i(X^n_i)}(A) \text{ if } t \in [i/n, i/n + 1/n) \tag{4.3}$$

for Borel sets $A \subset \mathbb{R}^d$ and $B \subset [0, 1]$. This measure and controlled analogues to be introduced below record the joint empirical distribution of velocity and time. Owing to conflicting but standard usage, in this chapter L is used for both an empirical measure (as defined above) and the local rate function. The intended use should always be clear, since the former appears only as L^n, and the latter as L.

The following construction identifies quantities that will appear in the representation as well as others to be used in the convergence analysis. As first discussed in Sect. 3.1, we can consider $[\mu^n]_{i|0,\dots,i-1}(dv_i|\bar{v}^n_0, \dots, \bar{v}^n_{i-1})$ to be simply a random measure on $\mathbb{R}^d$ that is measurable with respect to $\sigma(\bar{v}^n_j, j = 0, \dots, i-1) = \sigma(\bar{X}^n_j, j = 1, \dots, i)$, and this ω-dependent measure is denoted by $\bar{\mu}^n_i(dv_i)$. Also as in Sect. 3.1, for notational convenience we assume that the original processes as well as controlled analogues are all defined on the same probability space. Note that the role of the "driving noises" played by X_i in Sect. 3.1 is here played by $v_i(X^n_i)$. The measure μ^n picks new distributions for these driving noises, as reflected by the notation. Another minor notational difference is that the noise index is from 0 to $n-1$ rather than 1 to n.

Construction 4.4 *Suppose we are given a probability measure $\mu^n \in \mathscr{P}((\mathbb{R}^d)^n)$ and decompose it in terms of conditional distributions $[\mu^n]_{i|1,\dots,i-1}$ on the ith variable given variables 0 through $i-1$:*

$$\mu^n(dv_0 \times \cdots \times dv_{n-1}) = [\mu^n]_0(dv_0)[\mu^n]_{1|0}(dv_1|v_0) \\ \times \cdots \times [\mu^n]_{n-1|1,\ldots,n-2}(dv_{n-1}|v_0, \ldots, v_{n-2}).$$

Let $\{\bar{v}_i^n\}_{i=0,\ldots,n-1}$ be random variables defined on a probability space $(\Omega, \mathscr{F}, P)$ and with joint distribution μ^n. Thus conditioned on $\bar{\mathscr{F}}_i^n \doteq \sigma(\bar{v}_j^n, j = 0, \ldots, i-1)$, $\bar{v}_i^n$ has distribution $\bar{\mu}_i^n(dv_i) \doteq [\mu^n]_{i|0,\ldots,i-1}(dv_i|\bar{v}_0^n, \ldots, \bar{v}_{i-1}^n)$. The collection $\{\bar{\mu}_i^n\}_{i=0,\ldots,n-1}$ will be called a control. Then controlled processes $\bar{X}^n$ and measures $\bar{L}^n$ are recursively constructed as follows. Let $\bar{X}_0^n = x_0$, and for $i = 1, \ldots, n$ define $\bar{X}_i^n$ recursively by

$$\bar{X}_{i+1}^n = \bar{X}_i^n + \frac{1}{n}\bar{v}_i^n.$$

When $\{\bar{X}_i^n\}_{i=1,\ldots,n}$ has been constructed, $\bar{X}^n(t)$ is defined as in (4.2) as piecewise linear interpolation, and

$$\bar{L}^n(A \times B) \doteq \int_B \bar{L}^n(A|t)dt, \quad \bar{L}^n(A|t) \doteq \delta_{\bar{v}_i^n}(A) \text{ if } t \in [i/n, i/n + 1/n).$$

We also define

$$\bar{\mu}^n(A \times B) \doteq \int_B \bar{\mu}^n(A|t)dt, \quad \bar{\mu}^n(A|t) \doteq \bar{\mu}_i^n(A) \text{ if } t \in [i/n, i/n + 1/n)$$

and

$$\lambda^n(A \times B) \doteq \int_B \lambda^n(A|t)dt, \quad \lambda^n(A|t) \doteq \theta(A|\bar{X}_i^n) \text{ if } t \in [i/n, i/n + 1/n).$$

The measures $\bar{\mu}^n(dx \times dt)$ record the time dependence of the $\bar{\mu}_i^n$. When taking limits, we will also want to keep track of the corresponding $\theta(\cdot|\bar{X}_i^n)$, since the two appear together in the relative entropy representation. This information is recorded in $\lambda^n \in \mathscr{P}(\mathbb{R}^d \times [0, 1])$. Note also that, as remarked previously, $\bar{\mathscr{F}}_i^n = \sigma(\bar{X}_j^n, j = 1, \ldots, i)$.

Theorem 4.5 *Let $G : \mathscr{P}(\mathbb{R}^d \times [0, 1]) \to \mathbb{R}$ be bounded from below and measurable. Let L^n be defined as in (4.3), and given a control $\{\bar{\mu}_i^n\}$, let $\{\bar{X}_i^n\}$ and $\{\bar{L}^n\}$ be defined as in Construction 4.4. Then*

$$-\frac{1}{n}\log E e^{-nG(L^n)} = \inf_{\{\bar{\mu}_i^n\}} E\left[G(\bar{L}^n) + \frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot) \,\|\, \theta(\cdot|\bar{X}_i^n)\right)\right].$$

Proof The representation follows directly from the high-level variational representation for exponential integrals [part (a) of Proposition 2.3] and the chain rule [Theorem 2.6], and the argument is almost the same as that used to derive the representation

used to prove Sanov's theorem [Proposition 3.1]. The only difference is that the base measure in that case was product measure, reflecting the iid noise structure. Here the base measure is

$$\theta(dv_0|x_0^n)\theta(dv_1|x_1^n) \times \cdots \times \theta(dv_{n-1}|x_{n-1}^n),$$

where

$$x_i^n = x_0 + \frac{1}{n}\sum_{j=0}^{i-1} v_j.$$

One applies the chain rule exactly as was done in Proposition 3.1. The change in the base measure is reflected by a change in the measures appearing in the relative entropy cost, i.e., $\theta(\cdot|\bar{X}_i^n)$ rather than $\theta(\cdot)$ as in the iid case. □

Note that the definition of $\bar{L}^n$ allows us to write

$$\bar{X}^n(t) = \int_{\mathbb{R}^d \times [0,t]} y\bar{L}^n(dy \times ds) + x_0.$$

Thus a special case of the representation in Theorem 4.5 occurs for F that is a bounded and measurable map from $\mathscr{C}([0,1] : \mathbb{R}^d)$ to $\mathbb{R}$:

$$-\frac{1}{n}\log Ee^{-nF(X^n)} = \inf_{\{\bar{\mu}_i^n\}} E\left[F(\bar{X}^n) + \frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\middle\|\,\theta(\cdot|\bar{X}_i^n)\right)\right]. \tag{4.4}$$

This representation will be used in the proof of the LDP for $\{X^n\}$. As in passing from Sanov's theorem to Cramér's theorem, convergence of $\bar{L}^n$ plus some uniform integrability will imply convergence of $\bar{X}^n$.

Remark 4.6 Although the proof of the LDP requires only bounded F (and hence bounded G), we state Theorem 4.5 so as to allow its use in the analysis of importance sampling in Chap. 15, where unbounded functionals cannot be avoided.

4.3 Form of the Rate Function

Before going further, we pause to comment on the expected form of the rate function. We give a completely heuristic calculation, based on a time scale separation due to the $1/n$ scaling of the noise and the weak continuity of $x \to \theta(\cdot|x)$, which suggests the correct form of the rate function. Over an interval $[s, s+\delta]$, with $\delta > 0$ small and $1/\delta \in \mathbb{N}$, the noise terms in the definition of $X^n(s+\delta) - X^n(s)$ are approximately iid with distribution $\theta(\cdot|X^n(s))$. Therefore,

$$\frac{X^n(s+\delta) - X^n(s)}{\delta} \approx \frac{1}{n\delta} \sum_{i=\lfloor ns \rfloor}^{\lfloor ns+n\delta \rfloor} v_i(X^n(s)),$$

and by Cramér's theorem, the right-hand side satisfies an LDP with the rate function $\delta L(X^n(s), \beta)$, where

$$L(x, \beta) = \inf \left[R\left(\mu(\cdot) \,\|\theta(\cdot|x)\right) : \int_{\mathbb{R}^d} y\mu(dy) = \beta \right]. \tag{4.5}$$

Suppose that $\sigma > 0$ is small, and that in the following display, $B(y, \sigma)$ denotes a (context-dependent) open ball of radius σ. Using the Markov property to combine estimates over small intervals, for a smooth trajectory $\phi \in \mathscr{C}([0, 1] : \mathbb{R}^d)$ that starts at x_0, we have

$$\begin{aligned}
&P\left\{X^n \in B(\phi, \sigma)\right\} \\
&\quad \approx P\left\{X^n(j\delta) \in B(\phi(j\delta), \sigma) \text{ all } 1 \le j \le 1/\delta\right\} \\
&\quad \approx P\left\{ \frac{X^n(j\delta+\delta) - X^n(j\delta)}{\delta} \in B\left(\frac{\phi(j\delta+\delta) - \phi(j\delta)}{\delta}, \frac{2\sigma}{\delta} \right),\ 0 \le j < \frac{1}{\delta} \right\} \\
&\quad \approx \prod_{j=0}^{\frac{1}{\delta}-1} \exp\left\{ -n\delta L\left(\phi(j\delta), \frac{\phi(j\delta+\delta) - \phi(j\delta)}{\delta} \right) \right\} \\
&\quad \approx \exp\left\{ -n \int_0^1 L(\phi(s), \dot{\phi}(s)) ds \right\}.
\end{aligned}$$

Therefore, one may expect the rate function $I(\phi) = \int_0^1 L(\phi(s), \dot{\phi}(s)) ds$ for such ϕ. Owing to this interpretation, $\beta \mapsto L(x, \beta)$ is often called a **local rate function** in this context.

4.4 Statement of the LDP

We now turn to the rigorous analysis. As was the case in Chap. 3 with Sanov's theorem and small noise diffusions, we first establish tightness, and then prove a result that links the limits of weakly converging controls and controlled processes. With these results in hand, the Laplace principle is proved by establishing upper and lower bounds. The conditions we assume and some of the arguments are close to those used in [97]. However, the perspective is somewhat different, with the main argument being a functional version of the one used to obtain Cramér's theorem from Sanov's theorem, and we also set the arguments up so they can easily be adapted to the problems of importance sampling considered later in the book.

We show that Condition 4.3 by itself suffices for the Laplace principle and large deviation upper bound. For the lower bound we need additional conditions. Two types of conditions will be used, and are formulated as Conditions 4.7 and 4.8 below. The Laplace principle lower bound under Conditions 4.3 and 4.7 will be proved in Sect. 4.7, and under Conditions 4.3 and 4.8 it will be proved in Sect. 4.8. The convex hull of the support of $\mu \in \mathscr{P}(\mathbb{R}^d)$ is the smallest closed and convex set $A \subset \mathbb{R}^d$ such that $\mu(A) = 1$.

Condition 4.7 For each $x \in \mathbb{R}^d$, the convex hull of the support of $\theta(\cdot|x)$ is $\mathbb{R}^d$.

Condition 4.8 For every compact $K \subset \mathbb{R}^d$ and $\varepsilon \in (0, 1)$, there exist $\eta = \eta(K, \varepsilon) \in (0, 1)$ and $m = m(K, \varepsilon) \in (0, \infty)$, such that whenever $\xi, \chi \in K$ satisfy $\|\xi - \chi\| \leq \eta$, we can find for each $\gamma \in \mathbb{R}^d$ a $\beta \in \mathbb{R}^d$ such that

$$L(\xi, \beta) - L(\chi, \gamma) \leq \varepsilon(1 + L(\chi, \gamma)), \quad \|\beta - \gamma\| \leq m(1 + L(\chi, \gamma))\|\xi - \chi\|.$$

Condition 4.7 can be weakened to the requirement that the relative interior of the convex hull of the support of $\theta(\cdot|x)$ be independent of x and contain 0 (see Sect. 6.3 of [97]). Condition 4.8 is very important in that it allows the noise to push the process in only a subset of all possible directions. For example, if the model of Example 4.1 corresponds to a degenerate diffusion, which means that $\sigma(x)\sigma^T(x)$ is only positive semidefinite, then Condition 4.7 is not valid, but under the assumption that b and σ are Lipschitz continuous, Condition 4.8 holds. Under similar Lipschitz-type assumptions, Condition 4.8 is satisfied for a broad range of models, and we refer the reader to Sect. 6.3 in [97] for additional illustrative examples.

Recall that $\mathscr{AC}_{x_0}([0, T] : \mathbb{R}^d)$ denotes the subset of $\mathscr{C}([0, T] : \mathbb{R}^d)$ consisting of all absolutely continuous functions satisfying $\phi(0) = x_0$.

Theorem 4.9 *Assume Condition 4.3 and define X^n by (4.2) and $L : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ by (4.5). Let*

$$I(\phi) = \int_0^T L(\phi(s), \dot{\phi}(s))ds \quad \text{if } \phi \in \mathscr{AC}_{x_0}([0, T] : \mathbb{R}^d),$$

and in all other cases set $I(\phi) = \infty$. Then the following conclusions hold.

(a) *I is a rate function and $\{X^n\}_{n\in\mathbb{N}}$ satisfies the Laplace principle upper bound with rate function I.*

(b) *Suppose that in addition, either Condition 4.7 or Condition 4.8 holds. Then $\{X^n\}_{n\in\mathbb{N}}$ satisfies the Laplace principle with rate function I.*

Remark 4.10 In the proofs to follow, the initial condition is fixed at x_0. However, the arguments apply with only notational changes if instead we consider a sequence of initial conditions $\{x_0^n\}_{n\in\mathbb{N}}$ with $x_0^n \to x_0$, and establish

$$\frac{1}{n} \log E_{x_0^n} e^{-nF(X^n)} + \inf_{\phi:\phi(0)=x_0^n} [F(\phi) + I(\phi)] \to 0. \tag{4.6}$$

Using an elementary argument by contradiction, this implies that the Laplace and large deviation principles hold uniformly for initial conditions in compact sets, as defined in Chap. 1. To be specific, if the uniform Laplace principle is not valid, then there exists a compact set $K \subset \mathbb{R}^d$, $\delta > 0$, and for each $n \in \mathbb{N}$, an initial condition $x_0^n \in K$ such that

$$\left| \frac{1}{n} \log E_{x_0^n} e^{-nF(X^n)} + \inf_{\phi:\phi(0)=x_0^n} [F(\phi) + I(\phi)] \right| \geq \delta. \tag{4.7}$$

However, since K is compact, there exist a subsequence $x_0^{n_k}$ and $x_0 \in K$ such that $x_0^{n_k} \to x_0$. Then (4.6) contradicts (4.7), and thus the uniform Laplace principle holds.

The rest of the chapter is organized as follows. In Sect. 4.5 we prove part (a) of Theorem 4.9. In preparation for the (two) proofs of the Laplace lower bound, Sect. 4.6 studies some basic properties of the function $L(x, \beta)$. The last two sections of the chapter, Sects. 4.7 and 4.8, contain the proof of the lower bound under Condition 4.7 and Condition 4.8, respectively. Throughout the chapter we assume Condition 4.3, and to simplify notation, proofs are given for $T = 1$.

4.5 Laplace Upper Bound

We begin with preliminary results on tightness and uniform integrability of the controlled processes from Sect. 4.2.

4.5.1 Tightness and Uniform Integrability

Lemma 4.11 *Assume Condition 4.3 and consider any sequence of controls* $\{\bar{\mu}_i^n\}$ *for which the relative entropy costs satisfy*

$$\sup_{n\in\mathbb{N}} E\left[\frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot) \,\|\, \theta(\cdot|\bar{X}_i^n)\right)\right] \leq K < \infty.$$

Let $\{\bar{L}^n\}_{n\in\mathbb{N}}$, $\{\bar{X}^n\}_{n\in\mathbb{N}}$, $\{\bar{\mu}^n\}_{n\in\mathbb{N}}$, *and* $\{\lambda^n\}_{n\in\mathbb{N}}$ *be defined as in Construction 4.4. Then the empirical measures* $\{\bar{L}^n\}$ *are tight and in fact uniformly integrable in the sense that*

$$\lim_{M\to\infty} \limsup_{n\to\infty} E\left[\int_{\mathbb{R}^d\times[0,1]} \|y\| \, 1_{\{\|y\|\geq M\}} \bar{L}^n(dy \times dt)\right] = 0. \tag{4.8}$$

The measures $\{\bar{\mu}^n\}_{n\in\mathbb{N}}$ *are also uniformly integrable in the sense of* (4.8), *and* $\{\bar{X}^n\}$, $\{\bar{\mu}^n\}$, *and* $\{\lambda^n\}$ *are all tight.*

Proof Except for more complicated notation, the proof is almost the same as the analogous result needed for Cramér's theorem. From the inequality (2.9), it follows that if $\mu \in \mathscr{P}(\mathbb{R}^d \times [0,1])$ satisfies $\mu \ll \lambda^n$, then for all $\sigma \in [1, \infty)$,

$$\int_{\mathbb{R}^d \times [0,1]} \|y\| \, 1_{\{\|y\| \geq M\}} \mu(dy \times dt) \leq \int_{\mathbb{R}^d \times [0,1]} e^{\sigma \|y\|} 1_{\{\|y\| \geq M\}} \lambda^n(dy \times dt) + \frac{1}{\sigma} R\left(\mu \,\|\, \lambda^n\right).$$

By a conditioning argument it follows that $E \int f d\bar{L}^n = E \int f d\bar{\mu}^n$ for every bounded and measurable function f. Using the definitions of $\bar{\mu}^n$ and λ^n and the chain rule to get the first equality, we have

$$\begin{aligned} E\left[R\left(\bar{\mu}^n \,\|\, \lambda^n\right)\right] &= E\left[\int_0^1 R\left(\bar{\mu}^n(\cdot|t) \,\|\, \lambda^n(\cdot|t)\right) dt\right] \\ &= E\left[\frac{1}{n} \sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot) \,\|\, \theta(\cdot|\bar{X}_i^n)\right)\right] \\ &\leq K. \end{aligned} \tag{4.9}$$

Therefore,

$$\begin{aligned} &E\left[\int_{\mathbb{R}^d \times [0,1]} \|y\| \, 1_{\{\|y\| \geq M\}} \bar{L}^n(dy \times dt)\right] \\ &= E\left[\int_{\mathbb{R}^d \times [0,1]} \|y\| \, 1_{\{\|y\| \geq M\}} \bar{\mu}^n(dy \times dt)\right] \\ &\leq \sup_{x \in \mathbb{R}^d} \int_{\mathbb{R}^d} e^{\sigma \|y\|} 1_{\{\|y\| \geq M\}} \theta(dy|x) + \frac{1}{\sigma} K. \end{aligned} \tag{4.10}$$

From part (a) of Condition 4.3 it follows that for $\sigma \in \mathbb{R}$,

$$\sup_{x \in \mathbb{R}^d} \int_{\mathbb{R}^d} e^{2\sigma \|y\|} \theta(dy|x) < \infty \tag{4.11}$$

(for details see the analogous claim in the proof of Lemma 3.9). Since

$$\int_{\mathbb{R}^d} e^{\sigma \|y\|} 1_{\{\|y\| \geq M\}} \theta(dy|x) \leq e^{-\sigma M} \int_{\mathbb{R}^d} e^{2\sigma \|y\|} \theta(dy|x), \tag{4.12}$$

sending first $n \to \infty$, then $M \to \infty$, and finally $\sigma \to \infty$ in (4.10), the limit (4.8) holds for both $\{\bar{L}^n\}$ and $\{\bar{\mu}^n\}$. Tightness of $\{\bar{L}^n\}$ and $\{\bar{\mu}^n\}$ follows directly, and the tightness of $\{\lambda^n\}$ follows from part (a) of Condition 4.3.

To establish tightness of $\{\bar{X}^n\}$ we use the fact that

$$\bar{X}^n(t) = \int_{\mathbb{R}^d \times [0,t]} y \bar{L}^n(dy \times ds) + x_0. \tag{4.13}$$

Tightness will follow if given $\varepsilon > 0$ and $\eta > 0$, there is $\delta > 0$ such that

$$\limsup_{n\to\infty} P\left\{w^n(\delta) \geq \varepsilon\right\} \leq \eta, \tag{4.14}$$

where $w^n(\delta) \doteq \sup_{0\leq s<t\leq 1: t-s\leq\delta} \left\|\bar{X}^n(t) - \bar{X}^n(s)\right\|$. Using (4.8), choose $M < \infty$ such that

$$\limsup_{n\to\infty} E\left[\int_{\mathbb{R}^d\times[0,1]} \|y\| \, 1_{\{\|y\|\geq M\}} \bar{L}^n(dy \times dt)\right] \leq \frac{\varepsilon\eta}{2}.$$

Let $\delta \doteq (\varepsilon/2M) \wedge 1$. Then since $M\delta \leq \varepsilon/2$, we have

$$\sup_{0\leq s<u\leq 1: u-s\leq\delta} \int_{\mathbb{R}^d\times[s,u]} \|y\| \, 1_{\{\|y\|\leq M\}} \bar{L}^n(dy \times dt) \leq M\delta \leq \frac{\varepsilon}{2}.$$

Hence

$$\begin{aligned} P\left\{w^n(\delta) \geq \varepsilon\right\} &\leq P\left\{\int_{\mathbb{R}^d\times[0,1]} \|y\| \, 1_{\{\|y\|\geq M\}} \bar{L}^n(dy \times dt) \geq \frac{\varepsilon}{2}\right\} \\ &\leq \frac{2}{\varepsilon} E\left[\int_{\mathbb{R}^d\times[0,1]} \|y\| \, 1_{\{\|y\|\geq M\}} \bar{L}^n(dy \times dt)\right] \\ &\leq \eta. \end{aligned}$$

This proves (4.14), and tightness of $\{\bar{X}^n\}$ follows. □

4.5.2 Weak Convergence

Lemma 4.11 proved tightness of $\{(\bar{L}^n, \bar{\mu}^n, \lambda^n, \bar{X}^n)\}_{n\in\mathbb{N}}$. The following lemma characterizes the weak limits of this collection.

Lemma 4.12 *Consider any sequence of controls $\left\{\bar{\mu}_i^n\right\}$ as in Construction 4.4 for which the relative entropy costs satisfy*

$$E\left[\frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot) \,\middle\|\, \theta(\cdot|\bar{X}_i^n)\right)\right] \leq K < \infty.$$

Let $\{(\bar{X}^n, \bar{L}^n, \bar{\mu}^n)\}$ denote a weakly converging subsequence, which for notational convenience we again label by n, with limit $(\bar{X}, \bar{L}, \bar{\mu})$. Then w.p.1, $\bar{L} = \bar{\mu}$, and

$\bar{\mu}(dy \times dt)$ can be decomposed as $\bar{\mu}(dy|t)dt$, where $\bar{\mu}(dy|t)$ is a stochastic kernel on $\mathbb{R}^d$ given $[0, 1]$*, and w.p.1 for all* $t \in [0, 1]$*,*

$$\bar{X}(t) = \int_{\mathbb{R}^d \times [0,t]} y\bar{\mu}(dy \times ds) + x_0 = \int_{\mathbb{R}^d \times [0,t]} y\bar{\mu}(dy|s)ds + x_0. \tag{4.15}$$

In addition, λ^n converges weakly to a limit λ of the form

$$\lambda(A \times B) = \int_B \theta(A|\bar{X}(t))dt. \tag{4.16}$$

Proof Recall that $\bar{\mu}_i^n$ picks the conditional distribution of $\bar{\nu}_i^n$. Hence a minor modification of the martingale argument used to prove the analogous result needed for Sanov's theorem (Lemma 3.5) can be used to show that $\bar{L} = \bar{\mu}$ w.p.1. The changes are mainly notational, and are needed, since in the present setting the measures must record time information. For completeness we give the details.

Now, $\mathbb{R}^d \times [0, 1]$ is a Polish space, and on such a space there exists a countable separating class of bounded uniformly continuous functions (see Appendix A). Thus to verify $\bar{L} = \bar{\mu}$ w.p.1, it suffices to show that for every bounded uniformly continuous f,

$$P\left\{\int_{\mathbb{R}^d \times [0,1]} f(v,t)\bar{L}\,(dv \times dt) = \int_{\mathbb{R}^d \times [0,1]} f(v,t)\bar{\mu}\,(dv \times dt)\right\} = 1. \tag{4.17}$$

Define $K \doteq \|f\|_\infty$ and $\Delta_i^n \doteq f\left(\bar{\nu}_i^n, i/n\right) - \int_{\mathbb{R}^d} f\,(v, i/n)\,\bar{\mu}_i^n\,(dv)$. For all $\varepsilon > 0$,

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^n f\left(\bar{\nu}_i^n, i/n\right) - \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^d} f\,(v, i/n)\,\bar{\mu}_i^n\,(dv)\right| > \varepsilon\right\}$$
$$\leq \frac{1}{\varepsilon^2} E\left[\frac{1}{n^2}\sum_{i,j=1}^n \Delta_i^n \Delta_j^n\right].$$

Recall that $\bar{\mathscr{F}}_i^n \doteq \sigma(\bar{\nu}_j^n, j = 0, \ldots, i-1)$. By a standard argument, for $i \neq j$, conditioning on $\bar{\mathscr{F}}_{j\wedge i-1}^n$ gives $E[\Delta_i^n \Delta_j^n] = 0$. Since $|\Delta_i^n| \leq 2K$,

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^n f(\bar{\nu}_i^n, i/n) - \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^d} f(v, i/n)\bar{\mu}_i^n(dv)\right| > \varepsilon\right\} \leq \frac{4K^2}{n\varepsilon^2}.$$

Let $\gamma(\delta)$ denote the modulus $\sup_{v\in\mathbb{R}^d, 0\leq s\leq t\leq 1: t-s\leq\delta} \{|f(v,t) - f(v,s)|\}$. Since f is uniformly continuous, $\gamma(\delta) \downarrow 0$ as $\delta \downarrow 0$, and the definition of $\gamma(\delta)$ implies

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(\bar{v}_i^n, i/n) - \int_{\mathbb{R}^d \times [0,1]} f(v,t) \bar{L}^n(dv \times dt) \right| \leq \gamma(1/n)$$

$$\left| \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}^d} f(v, i/n) \bar{\mu}_i^n(dv) - \int_{\mathbb{R}^d \times [0,1]} f(v,t) \bar{\mu}^n(dv \times dt) \right| \leq \gamma(1/n).$$

Letting first $n \to \infty$ and then $\varepsilon \to 0$, we obtain (4.17), which proves $\bar{L} = \bar{\mu}$ w.p.1.

Note that both $\bar{L}^n$ and $\bar{\mu}^n$ have second marginals equal to Lebesgue measure. Since this property is inherited by the weak limits, $\bar{L}(\mathbb{R}^d \times \{t\}) = 0$ w.p.1. This property and the uniform integrability allow us to pass to the limit in (4.13) and obtain

$$\bar{X}(t) = \int_{\mathbb{R}^d \times [0,t]} y \bar{L}(dy \times ds) + x_0.$$

Now use that $\bar{L} = \bar{\mu}$ w.p.1 to get the first part of (4.15). Since each $\bar{\mu}$ has Lebesgue measure as its second marginal, both the decomposition and the second part of (4.15) follow. Finally, the weak convergence of λ^n and the form of the limit follow from the weak convergence of $\bar{X}^n$ to $\bar{X}$ and the assumption that $x \to \theta(\cdot|x)$ is continuous in the weak topology. □

4.5.3 Completion of the Laplace Upper Bound

The large deviation and Laplace principle upper bounds correspond to the variational lower bound. To prove such a lower bound, we again follow the line of argument used for Cramér's theorem in Sect. 3.1.6. Fix a continuous and bounded $F : \mathscr{C}([0,1] : \mathbb{R}^d) \to \mathbb{R}$ and $\varepsilon > 0$. Using (4.4), let $\left\{\bar{\mu}_i^n\right\}_{i=1,\ldots,n}$ satisfy

$$-\frac{1}{n} \log E e^{-nF(X^n)} + \varepsilon \geq E\left[F(\bar{X}^n) + \frac{1}{n} \sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot) \,\|\, \theta(\cdot|\bar{X}_i^n)\right) \right].$$

Then since F is bounded, we have $\sup_n \frac{1}{n} E \sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot) \,\|\, \theta(\cdot|\bar{X}_i^n)\right) < \infty$, and therefore by Lemma 4.11, it follows that $\left\{(\bar{L}^n, \bar{X}^n, \bar{\mu}^n, \lambda^n)\right\}$ is tight. Consider any subsequence that converges to a weak limit $(\bar{L}, \bar{X}, \bar{\mu}, \lambda)$, and denote the convergent subsequence by n. If the lower bound is demonstrated for this subsequence, then the standard argument by contradiction establishes the lower bound for the original sequence. Details of the following calculation are given after the display:

$$\liminf_{n \to \infty} -\frac{1}{n} \log E \exp\{-nF(X^n)\} + \varepsilon$$

$$
\begin{aligned}
&\geq \liminf_{n\to\infty} E\left[F(\bar{X}^n) + \frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\theta(\cdot|\bar{X}_i^n)\right)\right] \\
&= \liminf_{n\to\infty} E\left[F(\bar{X}^n) + R\left(\bar{\mu}^n(dy\times dt)\,\|\lambda^n(dy\times dt)\right)\right] \\
&\geq E\left[F\left(\bar{X}\right) + R\left(\bar{\mu}(dy\times dt)\,\|\lambda(dy\times dt)\right)\right] \\
&= E\left[F\left(\bar{X}\right) + \int_{[0,1]} R\left(\bar{\mu}(\cdot|t)\,\|\theta(\cdot|\bar{X}(t))\right)dt\right] \\
&\geq E\left[F(\bar{X}) + \int_{[0,1]} L(\bar{X}(t),\dot{\bar{X}}(t))dt\right] \\
&\geq \inf_{\phi}\left[F(\phi) + \int_{[0,1]} L(\phi(t),\dot{\phi}(t))dt\right]. \qquad (4.18)
\end{aligned}
$$

The first equality uses the rewriting of the relative entropy in (4.9); the next inequality is due to the weak convergence, the lower semicontinuity of $R(\cdot\,\|\cdot)$, continuity of F, and Fatou's lemma; the next equality uses the decompositions $\bar{\mu}(dy\times dt) = \bar{\mu}(dy|t)dt$ and $\lambda(dy\times dt) = \theta(dy|\bar{X}(t))dt$ and the chain rule; the third inequality follows from (4.5) and (4.15); and the infimum in the last line is over all $\phi \in \mathscr{AC}_{x_0}([0,1]:\mathbb{R}^d)$. Since $\varepsilon > 0$ is arbitrary, we have proved the Laplace upper bound for $\{X^n\}$:

$$
\limsup_{n\to\infty} \frac{1}{n}\log E\exp\{-nF(X^n)\} \leq -\inf_{\phi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [F(\phi) + I(\phi)]\,.
$$

□

4.5.4 *I is a Rate Function*

As first noted in Chap. 3, in the weak convergence approach, a deterministic version of the argument used to prove the Laplace upper bound will usually show that the proposed rate function is indeed a rate function, i.e., that it has compact level sets.

Theorem 4.13 *Assume Condition 4.3, define $L(x,\beta)$ by (4.5), and let I be the function defined in Theorem 4.9. Then I has compact level sets in $\mathscr{C}([0,T]:\mathbb{R}^d)$.*

Proof As usual, the proof is given for $T = 1$. Suppose $\{\phi_j\}_{j\in\mathbb{N}}$ is given such that $I(\phi_j) \leq K < \infty$ for all $j \in \mathbb{N}$. Then we need to show that $\{\phi_j\}$ is precompact, and that if $\phi_j \to \phi$, then

$$
\liminf_{j\to\infty} I(\phi_j) \geq I(\phi).
$$

Since $I(\phi_j) < \infty$, we know ϕ_j is absolutely continuous. Define probability measures μ^j on $\mathbb{R}^d \times [0,1]$ by

$$\mu^j(A \times B) = \int_B \delta_{\dot{\phi}_j(t)}(A)dt, \ A \in \mathscr{B}(\mathbb{R}^d), B \in \mathscr{B}([0,1]).$$

Note that

$$\phi_j(t) = x_0 + \int_{\mathbb{R}^d \times [0,t]} y\mu^j(dy \times ds).$$

Using $I(\phi_j) \le K < \infty$, exactly the same argument as in Lemma 4.11 shows that $\{\mu^j\}_{j\in\mathbb{N}}$ is tight and uniformly integrable. By the usual subsequential argument, we can assume that μ^j converges along the full sequence, and a deterministic version of Lemma 4.12 shows that the limit μ can be factored in the form $\mu(dy \times dt) = \mu(dy|t)dt$, and that $\phi_j \to \phi$, with $\int_{\mathbb{R}^d} y\mu(dy|t) = \dot{\phi}(t)$. Thus $\{\phi_j\}$ is precompact. We now argue that $I(\phi) \le K$. In Lemma 4.14 it will be shown that L is a lower semicontinuous function that is convex in the second variable. Using these properties, we obtain

$$\begin{aligned}
K &\ge \liminf_{j\to\infty} I(\phi_j) \\
&= \liminf_{j\to\infty} \int_{\mathbb{R}^d \times [0,1]} L(\phi_j(t), y)\mu^j(dy \times dt) \\
&\ge \int_{\mathbb{R}^d \times [0,1]} L(\phi(t), y)\mu(dy \times dt) \\
&= \int_0^1 \int_{\mathbb{R}^d} L(\phi(t), y)\mu(dy|t)dt \\
&\ge \int_0^1 L(\phi(t), \dot{\phi}(t))dt = I(\phi),
\end{aligned}$$

where the second inequality is a consequence of Fatou's lemma, the lower semicontinuity of L, and the convergence of (ϕ_j, μ^j) to (ϕ, μ), while the third inequality uses the convexity of L and Jensen's inequality. Thus I has compact level sets, and hence is a rate function. □

4.6 Properties of $L(x, \beta)$

To prove a Laplace lower bound, we must take a trajectory ϕ that nearly minimizes in $\inf_{\phi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [F(\phi) + I(\phi)]$ and show how to construct a control that can be applied in the representation that will give asymptotically the same cost. For the continuous-time models in Chap. 3 this was not very difficult, in part because the implementation of the control was straightforward. For example, in the case of the diffusion model, the construction of a solution to (3.17) is possible when v is measurable in t and has appropriate integrability properties; in particular, piecewise continuity or some similar form of regularity is not required. The situation is different in discrete time.

In general, $I(\phi) < \infty$ implies only that ϕ is absolutely continuous. As we will see, it is natural to define controls for the prelimit in terms of $\dot{\phi}(t)$, where ϕ is nearly minimizing. Since the derivative is well defined only up to a set of Lebesgue measure zero, this causes a number of problems. The solution is to show that one can always construct a "nice" nearly minimizing trajectory, e.g., one whose derivative is continuous from the left with right-hand limits. Such a construction requires some regularity properties of $L(x, \beta)$, which we now present.

Recall that for $x \in \mathbb{R}^d$, the Legendre–Fenchel transform of $H(x, \cdot)$ is defined by

$$H^*(x, \beta) = \sup_{\alpha \in \mathbb{R}^d} [\langle \alpha, \beta \rangle - H(x, \alpha)], \ \beta \in \mathbb{R}^d,$$

and $L(x, \beta)$ is defined as in (4.5). As noted in Remark 3.10 and shown in Lemma 4.16, for each fixed x these are dual representations of the same function.

Lemma 4.14 *Assume Condition 4.3. Then the following are valid.*

(a) *For each $x \in \mathbb{R}^d$, $\alpha \mapsto H(x, \alpha)$ is a finite convex function on $\mathbb{R}^d$ that is differentiable for all $\alpha \in \mathbb{R}^d$. Also, $(x, \alpha) \mapsto H(x, \alpha)$ is continuous on $\mathbb{R}^d \times \mathbb{R}^d$.*

(b) *For each $x \in \mathbb{R}^d$, $\beta \mapsto H^*(x, \beta)$ is a convex function on $\mathbb{R}^d$. Furthermore, $(x, \beta) \mapsto H^*(x, \beta)$ is a nonnegative lower semicontinuous function on $\mathbb{R}^d \times \mathbb{R}^d$.*

(c) *The map $\beta \mapsto H^*(x, \beta)$ is superlinear, uniformly in x, which means that*

$$\lim_{N \to \infty} \inf_{x \in \mathbb{R}^d} \inf_{\beta \in \mathbb{R}^d : \|\beta\| = N} \frac{H^*(x, \beta)}{\|\beta\|} = \infty.$$

(d) *The map $(x, \beta) \mapsto L(x, \beta)$ is a lower semicontinuous function on $\mathbb{R}^d \times \mathbb{R}^d$, and for each $x \in \mathbb{R}^d$, the map $\beta \mapsto L(x, \beta)$ is convex.*

Proof (a) Part (a) of Condition 4.3 ensures that $H(x, \alpha) \in (-\infty, \infty)$ for all $(x, \alpha) \in \mathbb{R}^d \times \mathbb{R}^d$. The convexity of $\alpha \mapsto H(x, \alpha)$ then follows from Hölder's inequality: if $\alpha_1, \alpha_2 \in \mathbb{R}^d$ and $\rho \in [0, 1]$, then

$$\int_{\mathbb{R}^d} e^{\langle \rho\alpha_1 + (1-\rho)\alpha_2, y \rangle} \theta(dy|x) \le \left(\int_{\mathbb{R}^d} e^{\langle \alpha_1, y \rangle} \theta(dy|x) \right)^{\rho} \left(\int_{\mathbb{R}^d} e^{\langle \alpha_2, y \rangle} \theta(dy|x) \right)^{1-\rho}.$$

Taking the logarithm of both sides demonstrates convexity. Under part (a) of Condition 4.3 one can easily construct an appropriate dominating function, and hence show that $\alpha \mapsto H(x, \alpha)$ is differentiable, where for all $(x, \alpha) \in \mathbb{R}^d \times \mathbb{R}^d$, the gradient $\nabla_\alpha H(x, \alpha)$ is given by

$$\nabla_\alpha H(x, \alpha) = \frac{\int_{\mathbb{R}^d} y e^{\langle \alpha, y \rangle} \theta(dy|x)}{\int_{\mathbb{R}^d} e^{\langle \alpha, y \rangle} \theta(dy|x)}. \tag{4.19}$$

To see the continuity of $(x, \alpha) \mapsto H(x, \alpha)$, let $(x_n, \alpha_n) \to (x, \alpha)$ in $\mathbb{R}^d \times \mathbb{R}^d$. We write

$$e^{H(x_n,\alpha_n)} - e^{H(x,\alpha)} = \int_{\mathbb{R}^d} \left[e^{\langle \alpha_n, y\rangle} - e^{\langle \alpha, y\rangle}\right] \theta(dy|x_n) + \left[\int_{\mathbb{R}^d} e^{\langle \alpha, y\rangle} \theta(dy|x_n) - \int_{\mathbb{R}^d} e^{\langle \alpha, y\rangle} \theta(dy|x)\right],$$

and recall the bounds (4.11) and (4.12). The second term on the right converges to zero by the Feller property assumed in Condition 4.3(b) and the uniform integrability implied by Condition 4.3(a). For every $M < \infty$, we have $\sup_{\|y\| \le M} |\exp\langle \alpha_n, y\rangle - \exp\langle \alpha, y\rangle| \to 0$ as $n \to \infty$, and by part (a) of Condition 4.3,

$$\sup_{n\in\mathbb{N}} \sup_{x\in\mathbb{R}^d} \int_{\{\|y\|\ge M\}} e^{\langle \alpha_n, y\rangle} \theta(dy|x) \to 0$$

as $M \to \infty$. Hence the first term on the right converges to zero, which completes the proof of the continuity of $(x, \alpha) \mapsto H(x, \alpha)$.

(b) By duality for Legendre–Fenchel transforms [217, Theorem 23.5],

$$H(x, \alpha) = \sup_{\beta\in\mathbb{R}^d} [\langle \alpha, \beta\rangle - H^*(x, \beta)], \ \alpha \in \mathbb{R}^d.$$

Taking $\alpha = 0$ in the last display shows that $\inf_{\beta\in\mathbb{R}^d} H^*(x, \beta) = -H(x, 0) = 0$, and thus H^* is nonnegative. For each $x \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}^d$, the mapping $\beta \mapsto \langle \alpha, \beta\rangle - H(x, \alpha)$ is convex (in fact affine), and for each $\alpha \in \mathbb{R}^d$, the mapping $(x, \beta) \mapsto \langle \alpha, \beta\rangle - H(x, \alpha)$ is continuous. Recalling that the pointwise supremum of convex functions is convex and that the pointwise supremum of continuous functions is lower semicontinuous, we see that $H^*(x, \cdot)$ is convex on $\mathbb{R}^d$ and H^* is lower semicontinuous on $\mathbb{R}^d \times \mathbb{R}^d$.

(c) From part (a) of Condition 4.3, for every $M < \infty$, we have

$$C_M \doteq \sup_{x\in\mathbb{R}^d} \sup_{\alpha\in\mathbb{R}^d:\|\alpha\|=M} H(x, \alpha) < \infty.$$

Also, for every $\beta \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$, the definition of H^* implies

$$H^*(x, \beta) \ge \langle M\beta/\|\beta\|, \beta\rangle - H(x, M\beta/\|\beta\|) \ge M\|\beta\| - C_M.$$

Thus

$$\inf_{x\in\mathbb{R}^d} \inf_{\beta\in\mathbb{R}^d:\|\beta\|=N} \frac{H^*(x, \beta)}{\|\beta\|} \ge M - \frac{C_M}{N}.$$

The asserted superlinearity follows by sending first $N \to \infty$ and then $M \to \infty$ in the last display.

(d) The claimed properties of $L(x, \beta)$ follow from its definition in (4.5) and the corresponding properties of $R(\cdot\,\|\,\cdot)$ [part (b) of Lemma 2.4]. We first consider convexity. Fix x, let $\beta_1, \beta_2 \in \mathbb{R}^d$, $\delta > 0$, $\rho \in [0, 1]$ be given, and suppose that μ_i

are within δ of the infimum in (4.5) for β_i, $i = 1, 2$. Then since the mean under $\rho\mu_1 + (1-\rho)\mu_2$ is $\rho\beta_1 + (1-\rho)\beta_2$, we have

$$\begin{aligned} L(x, \rho\beta_1 + (1-\rho)\beta_2) &\le R(\rho\mu_1(\cdot) + (1-\rho)\mu_2(\cdot) \,\|\theta(\cdot|x)) \\ &\le \rho R(\mu_1(\cdot) \,\|\theta(\cdot|x)) + (1-\rho) R(\mu_2(\cdot) \,\|\theta(\cdot|x)) \\ &\le \rho L(x, \beta_1) + (1-\rho) L(x, \beta_2) + \delta. \end{aligned}$$

Convexity follows, since $\delta > 0$ is arbitrary. Next suppose that $x_j \to x$ and $\beta_j \to \beta$ as $j \to \infty$. By an argument by contradiction based on subsequences, we can assume without loss that $L(x_j, \beta_j)$ converges in $[0, \infty]$, and we need to prove that $L(x, \beta) \le \lim_{j\to\infty} L(x_j, \beta_j)$. If the limit is ∞, there is nothing to prove, and hence we assume $L(x_j, \beta_j) \le M < \infty$ for all j. Choose μ_j that is within $\delta > 0$ of the infimum in the definition of $L(x_j, \beta_j)$. Then by part (d) of Lemma 2.4, $\{\mu_j\}$ is tight and uniformly integrable. Thus if μ^* is the limit of any convergence subsequence, then the mean of μ^* is β, and so (along this subsequence)

$$L(x, \beta) \le R(\mu^*(\cdot) \,\|\theta(\cdot|x)) \le \liminf_{j\to\infty} R(\mu_j(\cdot) \,\|\theta(\cdot|x_j)) \le \liminf_{j\to\infty} L(x_j, \beta_j) + \delta.$$

This establishes the lower semicontinuity. □

Remark 4.15 For the next result we will assume, in addition to Condition 4.3, that the support of $\theta(\cdot|x)$ is all of $\mathbb{R}^d$. Part (d) of the lemma was mentioned in Remark 3.10, which noted that the rate function for Cramér's theorem (which plays the role of the local rate function here) has two variational representations. These correspond here to H^* (as a supremum involving a moment-generating function) and L (as an infimum involving relative entropy). Although for our needs it suffices to prove this assuming Conditions 4.3 and 4.7, the functions H^* and L coincide when $\sup_{x\in\mathbb{R}^d} H(x, \alpha) < \infty$ for α in some open neighborhood of the origin. This will be proved in Lemma 5.4. Several statements in the lemma below hold assuming only Condition 4.3 (cf. [97, Lemma 6.2.3]). However, for simplicity we assume here that both Conditions 4.3 and 4.7 are satisfied.

Lemma 4.16 *Assume Conditions 4.3 and 4.7. Then the following conclusions hold.*

(a) H^* *is finite on* $\mathbb{R}^d \times \mathbb{R}^d$.

(b) *For every* $x \in \mathbb{R}^d$, $\alpha \mapsto H(x, \alpha)$ *is strictly convex on* $\mathbb{R}^d$.

(c) *For every* $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$, *there is a unique* $\alpha = \alpha(x, \beta) \in \mathbb{R}^d$ *such that* $\nabla_\alpha H(x, \alpha(x, \beta)) = \beta$.

(d) $H^* = L$.

(e) *For every* $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$, *with* $\alpha(x, \beta)$ *as in part (c),*

$$L(x, \beta) = \langle \alpha(x, \beta), \beta \rangle - H(x, \alpha(x, \beta)).$$

(f) $(x, \beta) \mapsto L(x, \beta)$ *is continuous on* $\mathbb{R}^d \times \mathbb{R}^d$.

(g) *There exists a measurable* $\alpha : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ *such that the stochastic kernel* $\gamma(dy|x, \beta)$ *on* $\mathbb{R}^d$ *given* $\mathbb{R}^d \times \mathbb{R}^d$ *defined by*

$$\gamma(dy|x, \beta) \doteq e^{\langle \alpha(x,\beta), y\rangle - H(x,\alpha(x,\beta))}\theta(dy|x),$$

satisfies

$$R(\gamma(\cdot|x, \beta)\|\theta(\cdot|x)) = L(x, \beta) \text{ and } \int_{\mathbb{R}^d} y\gamma(dy|x, \beta) = \beta \text{ for all } x \in \mathbb{R}^d, \beta \in \mathbb{R}^d. \tag{4.20}$$

Proof Since x plays no role other than as a parameter in parts (a)–(e), it is dropped from the notation in the proofs of these parts.

(a) Let Y be a random variable with distribution θ, and let Y_i denote the ith component. We first claim that the map $\alpha \to H(\alpha)$ is superlinear. By Condition 4.7, the support of θ is all of $\mathbb{R}^d$. Thus for each $M < \infty$,

$$\Theta_M \doteq \min_{\mathscr{I} \subset \{1,\ldots,d\}} P\left\{[\cap_{i\in\mathscr{I}}\{Y_i \geq M\}] \bigcap [\cap_{i\in\mathscr{I}^c}\{Y_i \leq -M\}]\right\} > 0.$$

Therefore, for every $\alpha \in \mathbb{R}^d$,

$$\frac{1}{\|\alpha\|+1}\log Ee^{\langle\alpha,Y\rangle} \geq \frac{1}{\|\alpha\|+1}\log\left[\Theta_M e^{M\sum_{i=1}^d |\alpha_i|}\right] = M\frac{\sum_{i=1}^d |\alpha_i|}{\|\alpha\|+1} + \frac{\log\Theta_M}{\|\alpha\|+1}.$$

The superlinearity of H now follows by sending $\|\alpha\| \to \infty$ and then $M \to \infty$. The superlinearity implies that the Legendre–Fenchel transform H^* of H is finite everywhere, since for each $\beta \in \mathbb{R}^d$, one can find a compact set $K \subset \mathbb{R}^d$ such that

$$H^*(\beta) = \sup_{\alpha\in\mathbb{R}^d}[\langle\alpha, \beta\rangle - H(\alpha)] = \sup_{\alpha\in K}[\langle\alpha, \beta\rangle - H(\alpha)].$$

Since H is continuous, the last expression is finite, and thus (a) follows.

(b) As shown in Lemma 4.14, H is convex. Suppose that for some $\alpha_1, \alpha_2 \in \mathbb{R}^d$, $\alpha_1 \neq \alpha_2$, and $\rho \in (0, 1)$, we have

$$H(\rho\alpha_1 + (1-\rho)\alpha_2) = \rho H(\alpha_1) + (1-\rho)H(\alpha_2).$$

Then the condition for equality in Hölder's inequality requires that for $\theta(dy)$ a.e. y,

$$\frac{\exp\langle\alpha_1, y\rangle}{\int_{\mathbb{R}^d}\exp\langle\alpha_1, z\rangle\theta(dz)} = \frac{\exp\langle\alpha_2, y\rangle}{\int_{\mathbb{R}^d}\exp\langle\alpha_2, z\rangle\theta(dz)},$$

which implies that

$$\langle\alpha_1 - \alpha_2, y\rangle = H(\alpha_1) - H(\alpha_2) \text{ a.s. } \theta(dy).$$

In other words, θ is supported on a hyperplane of dimension $d-1$. But this contradicts the fact that the support of $\theta(dy)$ is all of $\mathbb{R}^d$, which proves the strict convexity of $\alpha \mapsto H(\alpha)$.

(c) From Corollary 26.4.1 in [217] it now follows that the gradient $\nabla_\alpha H(\alpha)$ is onto $\mathbb{R}^d$. Thus given β, there exists a vector $\alpha(\beta)$ such that $\nabla_\alpha H(\alpha(\beta)) = \beta$. We claim that $\alpha(\beta)$ is unique. Suppose $\alpha_1 \neq \alpha_2$ are such that

$$\nabla_\alpha H(\alpha_1) = \nabla_\alpha H(\alpha_2) = \beta. \tag{4.21}$$

Define $\zeta : \mathbb{R} \to \mathbb{R}$ by $\zeta(\lambda) \doteq H(\alpha_1 + \lambda(\alpha_2 - \alpha_1))$, $\lambda \in \mathbb{R}$. From part (b), ζ is strictly convex on $\mathbb{R}$, and so

$$\zeta'(0) = \langle \nabla_\alpha H(\alpha_1), \alpha_2 - \alpha_1 \rangle < \zeta'(1) = \langle \nabla_\alpha H(\alpha_2), \alpha_2 - \alpha_1 \rangle.$$

This contradicts (4.21), and thus there is only one α that satisfies $\nabla_\alpha H(\alpha) = \beta$.

(d,e) Setting $\gamma(dy) = e^{\langle \alpha(\beta), y \rangle} \theta(dy) / e^{H(\alpha(\beta))}$, we have from (4.19),

$$\int_{\mathbb{R}^d} y\gamma(dy) = \frac{1}{e^{H(\alpha(\beta))}} \int_{\mathbb{R}^d} y e^{\langle \alpha(\beta), y \rangle} \theta(dy) = \nabla_\alpha H(\alpha(\beta)) = \beta. \tag{4.22}$$

A direct calculation using the form of $\gamma(dy)$, the definition of relative entropy, and the definition of L in (4.5) then gives

$$L(\beta) \leq R(\gamma \| \theta) = \langle \alpha(\beta), \beta \rangle - H(\alpha(\beta)) \leq H^*(\beta). \tag{4.23}$$

Using the definition of H and part (c) of Proposition 2.3, we have

$$H(\alpha) = \sup_{\mu \in \mathscr{P}(\mathbb{R}^d) : R(\mu \| \theta) < \infty} \left[\int_{\mathbb{R}^d} \langle \alpha, y \rangle \, \mu(dy) - R(\mu \| \theta) \right].$$

Therefore, for all $\alpha \in \mathbb{R}^d$ and $\mu \in \mathscr{P}(\mathbb{R}^d)$,

$$R(\mu \| \theta) \geq \left\langle \alpha, \int_{\mathbb{R}^d} y\mu(dy) \right\rangle - H(\alpha),$$

and consequently

$$L(\beta) \geq \langle \alpha, \beta \rangle - H(\alpha).$$

Since $\alpha \in \mathbb{R}^d$ is arbitrary, we have $L(\beta) \geq H^*(\beta)$. By (4.23), the reverse inequality holds, which shows that $L(\beta) = H^*(\beta)$ and also proves part (e).

(f) For the last two parts of the lemma we include the x dependence. From Lemma 4.14, $(x, \alpha) \mapsto H(x, \alpha)$ is continuous. We now show that joint continuity of $L(x, \beta)$ follows from this. If a sequence of differentiable convex functions g_i with Legendre transforms g_i^* converges pointwise to another differentiable convex function g with transform g^*, and if β is any point such that $g^*(\beta) < \infty$, then when-

ever $\beta_i \to \beta$, we have $g_i^*(\beta_i) \to g^*(\beta)$ [97, Lemma C.8.1]. We apply this result with $g_i(\alpha) = H(x_i, \alpha)$ and $g(\alpha) = H(x, \alpha)$ to conclude that if $x_i \to x$ and $\beta_i \to \beta$, then $L(x_i, \beta_i) \to L(x, \beta)$.

(g) To see the measurability of $(x, \beta) \mapsto \alpha(x, \beta)$, note that for each x, the strict convexity of $\alpha \mapsto H(x, \alpha)$ and the fact $L(x, \beta) = H^*(x, \beta) < \infty$ imply that $\beta \mapsto L(x, \beta)$ is differentiable for all $\beta \in \mathbb{R}^d$ [217, Theorem 26.3]. The characterization of $H(x, \alpha)$ as the Legendre–Fenchel transform of $L(x, \beta)$ then gives $\alpha(x, \beta) = \nabla_\beta L(x, \beta)$, from which measurability follows.

The second equality in (4.20) follows from (4.22), while the first equality follows on noting that the first inequality in (4.23) was shown to be an equality. □

4.7 Laplace Lower Bound Under Condition 4.7

In this section we prove the Laplace principle lower bound under Conditions 4.3 and 4.7. For the proof of this lower bound we construct a nearly optimal trajectory ϕ^* for $\inf_{\phi \in \mathscr{C}([0,1]:\mathbb{R}^d)} [F(\phi) + I(\phi)]$ that has a simple form. Based on ϕ^*, a control is constructed for use in the representation, so that the running cost is close to $I(\phi^*)$ and the associated controlled process converges to the nearly optimal trajectory ϕ^* as $n \to \infty$.

Fix $\varepsilon > 0$. Then there is $\zeta \in \mathscr{C}([0, 1] : \mathbb{R}^d)$ such that

$$[F(\zeta) + I(\zeta)] \le \inf_{\phi \in \mathscr{C}([0,1]:\mathbb{R}^d)} [F(\phi) + I(\phi)] + \varepsilon. \tag{4.24}$$

While $\{\zeta(t) : 0 \le t \le 1\}$ is bounded by continuity, we also claim that without loss of generality, we can assume that

$$\{\dot{\zeta}(t) : 0 \le t \le 1\} \text{ is bounded.}$$

This claim will be established in Sect. 4.7.3.

Recall from part (f) of Lemma 4.16 that L is continuous. Let $M < \infty$ and $K < \infty$ be such that

$$\sup_{t\in[0,1]} \|\zeta(t)\| \vee \sup_{t\in[0,1]} \|\dot{\zeta}(t)\| \le M, \qquad \sup_{\{(x,\beta):\|x\|\le M+1, \|\beta\|\le M+1\}} L(x, \beta) \le K.$$

For $\delta > 0$, let ζ^δ be the piecewise linear interpolation of ζ, with interpolation points $t = k\delta$. Since ζ is absolutely continuous, $\dot{\zeta}^\delta(t)$ converges to $\dot{\zeta}(t)$ for a.e. $t \in [0, 1]$. Also, since $\sup_{t\in[0,1]} \|\zeta^\delta(t)\| \le M$ and $\sup_{t\in[0,1]} \|\dot{\zeta}^\delta(t)\| \le M$, the continuity of L and the dominated convergence theorem imply that there is $\delta > 0$ such that $\left[F(\zeta^\delta) + I(\zeta^\delta)\right] \le [F(\zeta) + I(\zeta)] + \varepsilon$. We set $\phi^* = \zeta^\delta$ for such a δ.

4.7.1 Construction of a Nearly Optimal Control

The construction of a control to apply in the representation is now straightforward. Let $\gamma(dy|x,\beta)$ be as in part (g) of Lemma 4.16. Recall that $\bar{\mu}_i^n$ is allowed to be any measurable function of $\bar{X}_j^n$, $j = 0,\dots,i$. Define $N^n \doteq \inf\{j : \|\bar{X}_j^n - \phi^*(j/n)\| > 1\} \wedge n$. Then we set

$$\bar{\mu}_i^n(\cdot) = \begin{cases} \gamma(\cdot|\bar{X}_i^n, \dot{\phi}^*(i/n)) & \text{if } i < N^n, \\ \theta(\cdot|\bar{X}_i^n) & \text{if } i \geq N^n. \end{cases} \tag{4.25}$$

The cost under this control satisfies

$$E\left[\frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\middle\|\,\theta(\cdot|\bar{X}_i^n)\right)\right] = E\left[\frac{1}{n}\sum_{i=0}^{N^n-1} L(\bar{X}_i^n, \dot{\phi}^*(i/n))\right] \leq K,$$

and therefore Lemma 4.11 applies. Since $\tau^n \doteq N^n/n$ takes values in a compact set, given any subsequence of $\mathbb{N}$ we can find a further subsequence (again denoted by n) such that $(\bar{X}^n, \bar{\mu}^n, \tau^n)$ converges in distribution to a limit $(\bar{X}, \bar{\mu}, \tau)$. Also, it follows from the fact that the mean of $\bar{\mu}_i^n$ is $\dot{\phi}^*(i/n)$ for $i < N^n$ that

$$\begin{aligned} x_0 + \int_{\mathbb{R}^d\times[0,t\wedge\tau_n]} y\bar{\mu}^n(dy\times ds) &= x_0 + \frac{1}{n}\sum_{i=0}^{\lfloor nt\rfloor\wedge N^n-1} \dot{\phi}^*(i/n) + O(1/n) \\ &= \phi^*(t\wedge\tau^n) + O(1/n). \end{aligned}$$

Using the uniform integrability (4.8), we can send $n\to\infty$ and obtain

$$x_0 + \int_{\mathbb{R}^d\times[0,t]} y\bar{\mu}(dy|s)ds = \phi^*(t)$$

for all $t \in [0,\tau]$. It then follows from Lemma 4.12 [see (4.15)] that $\bar{X}(t) = \phi^*(t)$ for all $t \in [0,\tau]$, w.p.1. However, since $\|\bar{X}^n(\tau^n) - \phi^*(\tau^n)\|$ converges in distribution to $\|\bar{X}(\tau) - \phi^*(\tau)\|$, the definition of N^n implies that on the set $\tau < 1$, we have $\left\|\bar{X}(\tau) - \phi^*(\tau)\right\| \geq 1$. Thus $P(\tau < 1) = 0$, and so $\bar{X}(t) = \phi^*(t)$ for $t \in [0,1]$. We conclude that along the full sequence $\mathbb{N}$, $\bar{X}^n$ converges in distribution to ϕ^*.

4.7.2 Completion of the Proof of the Laplace Lower Bound

We now put the pieces together to prove the Laplace lower bound. For the particular control $\{\bar{\mu}_i^n\}$ just constructed, we have

$$
\begin{aligned}
&\limsup_{n\to\infty} -\frac{1}{n}\log E\exp\{-nF(X^n)\} \\
&\le \limsup_{n\to\infty} E\left[F(\bar{X}^n) + \frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n)\right)\right] \\
&= \limsup_{n\to\infty} E\left[F(\bar{X}^n) + \frac{1}{n}\sum_{i=0}^{N^n-1} L(\bar{X}_i^n, \dot{\phi}^*(i/n))\right] \\
&= \left[F(\phi^*) + \int_0^1 L(\phi^*(t), \dot{\phi}^*(t))dt\right] \\
&\le \inf_{\phi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [F(\phi) + I(\phi)] + 2\varepsilon.
\end{aligned}
$$

The first inequality follows since the representation considers the infimum over all controls, and the first equality is due to the definition of $\bar{\mu}_i^n$ in (4.25). The second equality follows from the weak convergence $\bar{X}^n \Rightarrow \phi^*$ [Lemma 4.12], the uniform bound on $L(\bar{X}_i^n, \dot{\phi}^*(i/n))$ for $i \le N^n - 1$, and the dominated convergence theorem. The last inequality uses the fact that ϕ^* as constructed is within 2ε of the infimum. Since $\varepsilon > 0$ is arbitrary, the Laplace lower bound

$$
\liminf_{n\to\infty} \frac{1}{n}\log E\exp\{-nF(X^n)\} \ge -\inf_{\phi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [F(\phi) + I(\phi)]
$$

follows. □

4.7.3 Approximation by Bounded Velocity Paths

We now prove the claim that for ζ satisfying (4.24), $\dot{\zeta}$ can also be assumed bounded.

Lemma 4.17 *Consider $\zeta \in \mathscr{C}([0,1]:\mathbb{R}^d)$ such that $[F(\zeta) + I(\zeta)] < \infty$. Then given $\varepsilon > 0$, there is ζ^* such that $\{\dot{\zeta}^*(t) : 0 \le t \le 1\}$ is bounded and*

$$
\sup_{0\le t\le 1} \|\zeta(t) - \zeta^*(t)\| \le \varepsilon, \quad I(\zeta^*) \le I(\zeta) + \varepsilon.
$$

Proof Since F is bounded, $I(\zeta) < \infty$. For $\lambda \in (0,1)$ let $D_\lambda \doteq \{t : \|\dot{\zeta}(t)\| \ge 1/\lambda\}$, and define a time rescaling $S_\lambda : [0,1] \to [0,\infty)$ by $S_\lambda(0) = 0$ and

$$
\dot{S}_\lambda(t) = \begin{cases} \|\dot{\zeta}(t)\|/(1-\lambda), & t \in D_\lambda, \\ 1 & \text{otherwise.} \end{cases}
$$

Then $S_\lambda(t)$ is continuous and strictly increasing. Let T_λ be the inverse of S_λ, which means that T_λ satisfies $T_\lambda(S_\lambda(t)) = t$ for all $t \in [0,1]$ and $S_\lambda(T_\lambda(t)) = t$ for all $t \in [0, S_\lambda(1)] \supset [0,1]$. Also define ζ_λ on $[0, S_\lambda(1)]$ by

$$\zeta_\lambda(t) \doteq \zeta(T_\lambda(t)),$$

which is a "slowed" version of ζ. By the ordinary chain rule, $\dot{\zeta}_\lambda(S_\lambda(t)) = \dot{\zeta}(t)/\dot{S}_\lambda(t)$, and therefore $\dot{\zeta}_\lambda(t)$ has uniformly bounded derivative for $t \in [0,1]$. The ζ^* in the lemma will be ζ_λ for a positive λ.

From part (c) of Lemma 4.14 and part (d) of Lemma 4.16, $L(x,\beta)$ is uniformly superlinear in β: $L(x,\beta)/\|\beta\| \to \infty$ uniformly in x as $\|\beta\| \to \infty$. This property and $I(\zeta) < \infty$ imply $\int_0^1 \|\dot{\zeta}(t)\| dt < \infty$, and consequently

$$\lim_{\lambda \to 0} \int_0^1 1_{D_\lambda}(t) \|\dot{\zeta}(t)\| dt = 0. \tag{4.26}$$

It follows that $\lim_{\lambda\to 0} S_\lambda(s) = s$ uniformly for $s \in [0,1]$. Since

$$\sup_{t\in[0,1]} \|\zeta_\lambda(t) - \zeta(t)\| = \sup_{t\in[0,1]} \|\zeta(T_\lambda(t)) - \zeta(t)\| = \sup_{s\in[0,T_\lambda(1)]} \|\zeta(s) - \zeta(S_\lambda(s))\|,$$

it follows that $\sup_{t\in[0,1]} \|\zeta_\lambda(t) - \zeta(t)\| \to 0$ as $\lambda \to 0$.

Thus we need only show that $I(\zeta_\lambda)$ is close to $I(\zeta)$. Let

$$\Gamma \doteq \sup_{t\in[0,1]} \sup_{\beta:\|\beta\|\le 1} L(\zeta(t),\beta) < \infty.$$

For $t \in D_\lambda$, the nonnegativity of L implies

$$\begin{aligned} L\left(\zeta(t), \frac{\dot{\zeta}(t)}{\dot{S}_\lambda(t)}\right)\dot{S}_\lambda(t) - L(\zeta(t),\dot{\zeta}(t)) &\le L\left(\zeta(t), \frac{(1-\lambda)\dot{\zeta}(t)}{\|\dot{\zeta}(t)\|}\right)\frac{\|\dot{\zeta}(t)\|}{1-\lambda} \\ &\le \frac{\Gamma}{1-\lambda}\|\dot{\zeta}(t)\|, \end{aligned}$$

and therefore

$$\begin{aligned} I(\zeta_\lambda) - I(\zeta) &\le \int_0^{S_\lambda(1)} L(\zeta_\lambda(t),\dot{\zeta}_\lambda(t))dt - \int_0^1 L(\zeta(t),\dot{\zeta}(t))dt \\ &= \int_0^1 L(\zeta_\lambda(S_\lambda(t)),\dot{\zeta}_\lambda(S_\lambda(t)))\dot{S}_\lambda(t)dt - \int_0^1 L(\zeta(t),\dot{\zeta}(t))dt \\ &= \int_0^1 L\left(\zeta(t), \frac{\dot{\zeta}(t)}{\dot{S}_\lambda(t)}\right)\dot{S}_\lambda(t)dt - \int_0^1 L(\zeta(t),\dot{\zeta}(t))dt \\ &\le \frac{\Gamma}{1-\lambda}\int_0^1 1_{D_\lambda}(t)\|\dot{\zeta}(t)\|dt. \end{aligned}$$

From (4.26), the last expression converges to 0 as $\lambda \to 0$. Thus $\limsup_{\lambda\to 0} I(\zeta_\lambda) \le I(\zeta)$ and $\zeta_\lambda \to \zeta$ as $\lambda \to 0$, and the claim is proved. □

With the proof that we can assume that $\dot{\zeta}$ is bounded for ζ appearing in (4.24), the proof of the Laplace lower bound under Condition 4.7 is complete.

4.8 Laplace Lower Bound Under Condition 4.8

In this section we prove the Laplace principle lower bound without assuming the support condition on $\theta(\cdot|x)$. Instead, besides Condition 4.3, we use the Lipschitz-type assumption of Condition 4.8. The main difficulty in the proof for this case is that the construction of a piecewise linear nearly minimizing trajectory, which simplified the proof in Sect. 4.7, is not directly available here. The arguments of Sect. 4.7 relied on the finiteness and continuity of the function $(x, \beta) \mapsto L(x, \beta)$, properties that in general will not hold for the setting considered in this section.

In order to overcome this difficulty, we introduce a mollification that in a sense reduces the problem to the form studied in Sect. 4.7. The mollification introduces an error that needs to be carefully controlled. This is the main technical challenge in the proof. We will also make use of Lemma 1.10, which shows that the Laplace principle lower bound holds if and only if it holds for F that are Lipschitz continuous. Mollification techniques are often used in large deviation analysis, and are especially useful in proving lower bounds.

The section is organized as follows. We begin in Sect. 4.8.1 by introducing the mollification of the state dynamics. This takes the form of a small additive Gaussian perturbation, parametrized by $\sigma > 0$, to the noise sequence $\{v_i(X_i^n)\}$ in (4.1). We then estimate the asymptotics of $-\frac{1}{n} \log E \exp\{-nF(X^n)\}$ through an analogous expression when X^n is replaced by the perturbed state process Z_σ^n. Next in Sect. 4.8.2 we give a variational upper bound for functionals of the perturbed process in terms of a convenient family of controls. The limits of cost functions in this representation are given in terms of a perturbation L_σ of the function L introduced in (4.5). Section 4.8.3 studies properties of L_σ. In particular, we show that L_σ is a finite continuous function, is bounded above by L, and satisfies properties analogous to those assumed of L in Condition 4.8. Using these results, in Sect. 4.8.4 we construct a piecewise linear nearly optimal trajectory for $\inf_{\phi \in \mathscr{C}([0,1]:\mathbb{R}^d)} [F(\phi) + I_\sigma(\phi)]$, where I_σ is the rate function associated with the local rate function L_σ, which is then used to construct an asymptotically nearly optimal control sequence for the representation. Section 4.8.5 studies tightness and convergence properties of the associated controlled processes. Finally, Sect. 4.8.6 uses these convergence results and estimates from Sect. 4.8.1 to complete the proof of the variational upper bound. Throughout this section, F will be a real-valued bounded Lipschitz continuous function on $\mathscr{C}([0, 1] : \mathbb{R}^d)$.

4.8.1 Mollification

For $\sigma > 0$, let $\{w_{i,\sigma}\}_{i \in \mathbb{N}_0}$ be an iid sequence of Gaussian random variables with mean 0 and covariance σI that is independent of the random vector fields $\{v_i(\cdot)\}_{i \in \mathbb{N}_0}$.

For $n \in \mathbb{N}$ and $\sigma > 0$, consider along with the sequence $\{X_i^n\}_{i=0,\dots,n}$, the sequence $\{U_{i,\sigma}^n\}_{i=0,\dots,n}$ defined by

$$U_{i+1,\sigma}^n \doteq U_{i,\sigma}^n + \frac{1}{n} w_{i,\sigma}, \quad U_{0,\sigma}^n = 0.$$

Define the piecewise linear process $\{U_\sigma^n(t)\}_{t\in[0,1]}$ by

$$U_\sigma^n(t) \doteq U_{i,\sigma}^n + \left[U_{i+1,\sigma}^n - U_{i,\sigma}^n\right](nt - i), \quad t \in [i/n, i/n + 1/n].$$

Let $Z_\sigma^n = X^n + U_\sigma^n$, where X^n is as in (4.2). Note that Z_σ^n is the piecewise linear interpolation of the sequence $\{Z_{i,\sigma}^n\}$, where $Z_{i,\sigma}^n = X_i^n + U_{i,\sigma}^n$.

The following result shows that the Laplace lower bound properties of $\{X^n\}$ can be bounded in terms of the lower bound properties of $\{Z_\sigma^n\}$. For $\phi \in \mathscr{C}([0,1] : \mathbb{R}^d)$, recall that $\|\phi\|_\infty \doteq \sup_{0\le t\le 1} \|\phi(t)\|$.

Lemma 4.18 *For every $\sigma > 0$,*

$$\limsup_{n\to\infty} -\frac{1}{n} \log E e^{-nF(X^n)} \le \limsup_{n\to\infty} -\frac{1}{n} \log E e^{-nF(Z_\sigma^n)} + \frac{M^2\sigma^2}{2},$$

where

$$M \doteq \sup_{\phi,\eta\in\mathscr{C}([0,1]:\mathbb{R}^d),\phi\neq\eta} \frac{|F(\phi) - F(\eta)|}{\|\phi - \eta\|_\infty}.$$

Proof Let $B = 2\|F\|_\infty$. Then since

$$F(Z_\sigma^n) \ge F(X^n) - (M\|U_\sigma^n\|_\infty) \wedge B,$$

we see that

$$-\frac{1}{n} \log E e^{-nF(X^n)} \le -\frac{1}{n} \log E e^{-nF(Z_\sigma^n)} + \frac{1}{n} \log E e^{n[(M\|U_\sigma^n\|_\infty)\wedge B]}. \tag{4.27}$$

We now estimate the second term on the right side of (4.27) using the Laplace principle upper bound (which was proved in Sect. 4.5) with $\theta(\cdot|x) = \rho_\sigma(\cdot)$, where ρ_σ is the law of a d-dimensional normal random variable with mean 0 and covariance σI. Let

$$H_\sigma(\alpha) \doteq \log \int_{\mathbb{R}^d} \exp\langle \alpha, y\rangle \rho_\sigma(dy) = \frac{\sigma^2}{2}\|\alpha\|^2, \ \alpha \in \mathbb{R}^d.$$

The Legendre–Fenchel transform of H_σ is given by

$$L_\sigma(\beta) \doteq \sup_{\alpha\in\mathbb{R}^d} [\langle \alpha, \beta\rangle - H_\sigma(\alpha)] = \frac{1}{2\sigma^2}\|\beta\|^2.$$

Then $\{U_\sigma^n\}_{n\in\mathbb{N}}$ satisfies the Laplace upper bound with rate function

$$I_{0,\sigma}(\varphi) \doteq \frac{1}{2\sigma^2}\int_0^1 \|\dot\varphi(s)\|^2 ds$$

if $\varphi \in \mathscr{C}([0,1]:\mathbb{R}^d)$ is absolutely continuous and $\varphi(0)=0$, and $I_{0,\sigma}(\varphi) \doteq \infty$ otherwise. This upper bound yields

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\log Ee^{n[(M\|U_\sigma^n\|_\infty)\wedge B]} \\
&\qquad \le -\inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)}\left[I_{0,\sigma}(\varphi) - (M\|\varphi\|_\infty)\wedge B\right] \\
&\qquad \le -\inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)}\left[I_{0,\sigma}(\varphi) - M\|\varphi\|_\infty\right].
\end{aligned} \tag{4.28}$$

For all $\varphi \in \mathscr{C}([0,1]:\mathbb{R}^d)$ with $I_{0,\sigma}(\varphi) < \infty$, we have

$$\|\varphi\|_\infty^2 = \sup_{t\in[0,1]}\left\|\int_0^t \dot\varphi(s)ds\right\|^2 \le \int_0^1 \|\dot\varphi(s)\|^2 ds,$$

and thus

$$\begin{aligned}
\inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)}\left[I_{0,\sigma}(\varphi) - M\|\varphi\|_\infty\right] &\ge \inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)}\left[\frac{\|\varphi\|_\infty^2}{2\sigma^2} - M\|\varphi\|_\infty\right] \\
&= \inf_{r\ge 0}\left[\frac{r^2}{2\sigma^2} - Mr\right] \\
&= -\frac{M^2\sigma^2}{2}.
\end{aligned}$$

The claim of the lemma now follows from the last display, (4.27), and (4.28). □

4.8.2 *Variational Bound for the Mollified Process*

In this section we present a variational bound for $Ee^{-nF(Z_\sigma^n)}$. The basic idea is to apply Theorem 4.5 with the Markov chain $\{X_i^n\}_{i\in\mathbb{N}_0}$ replaced by the $\mathbb{R}^{2d}$-valued Markov chain $\{(X_i^n, U_{i,\sigma}^n)\}_{i\in\mathbb{N}_0}$. Let $Y_{i,\sigma}^n \doteq (X_i^n, U_{i,\sigma}^n)$. The following construction is analogous to Construction 4.4, but for the doubled set of noises appearing in the mollification. In addition, we will build in the restriction on controls just mentioned.

Construction 4.19 *Suppose we are given a probability measure $\mu^n \in P((\mathbb{R}^d\times\mathbb{R}^d)^n)$, and decompose it into a collection of stochastic kernels. With a point in $(\mathbb{R}^d\times\mathbb{R}^d)^n$ denoted by $(v_1,w_1,v_2,w_2,\dots,v_n,w_n)$, $[\mu^n]_{i|0,\dots,i-1}^1$ will denote the*

marginal distribution of μ^n on v_i given (v_j, w_j), $j < i$, and $[\mu^n]^2_{i|0,\dots,i-1}$ will denote the marginal distribution of μ^n on w_i given (v_j, w_j), $j < i$ and v_i.

We now assume that μ^n also has the property that $[\mu^n]^2_{i|0,\dots,i-1}$ does not depend on v_i, which implies that the distributions on v_i and w_i are conditionally independent given (v_j, w_j), $j < i$. Let $\{(\bar{v}^n_i, \bar{w}^n_i)\}_{i=0,\dots,n-1}$ be random variables defined on a probability space $(\Omega, \mathscr{F}, P)$ and with joint distribution μ^n. Let $\bar{\mathscr{F}}^n_i \doteq \sigma((\bar{v}^n_j, \bar{w}^n_j), j = 0, \dots, i-1)$, and define

$$\bar{\mu}^{1,n}_i(dv_i) \doteq [\mu^n]^1_{i|0,\dots,i-1}(dv_i|\bar{v}^n_0, \bar{w}^n_0, \dots, \bar{v}^n_{i-1}, \bar{w}^n_{i-1})$$
$$\bar{\mu}^{2,n}_i(dw_i) \doteq [\mu^n]^2_{i|0,\dots,i-1}(dw_i|\bar{v}^n_0, \bar{w}^n_0, \dots, \bar{v}^n_{i-1}, \bar{w}^n_{i-1}),$$

so that these controls pick the distributions of $\bar{v}^n_i$ and $\bar{w}^n_i$ conditioned on $\bar{\mathscr{F}}^n_i$. Controlled processes $\bar{X}^n$, $\bar{U}^n_\sigma$ and measures $\bar{L}^n$ are then recursively constructed as follows. Let $(\bar{X}^n_0, \bar{U}^n_{0,\sigma}) = (x_0, 0)$ and define

$$\bar{X}^n_{i+1} = \bar{X}^n_i + \frac{1}{n}\bar{v}^n_i, \quad \bar{U}^n_{i+1,\sigma} = \bar{U}^n_{i,\sigma} + \frac{1}{n}\bar{w}^n_i. \tag{4.29}$$

Note that $\bar{\mathscr{F}}^n_i = \sigma((\bar{X}^n_j, \bar{U}^n_{j,\sigma}), j = 1, \dots, i)$. When $\{(\bar{X}^n_i, \bar{U}^n_{i,\sigma})\}_{i=1,\dots,n}$ has been constructed, $\bar{X}^n(t)$ and $\bar{U}^n_\sigma(t)$ are defined as in (4.2) as the piecewise linear interpolations, and we set $\bar{Z}^n_\sigma(t) \doteq \bar{X}^n(t) + \bar{U}^n_\sigma(t)$ for $t \in [0, 1]$. In addition, define

$$\bar{L}^n(A \times B) \doteq \int_B \bar{L}^n(A|t)dt, \quad \bar{L}^n(A|t) = \delta_{(\bar{v}^n_i, \bar{w}^n_i)}(A) \text{ if } t \in [i/n, i/n + 1/n).$$

The following is the main result of this section. Owing to the restriction placed on the controls in Construction 4.19, we obtain only an inequality, but the inequality is in the right direction to establish a Laplace lower bound.

Proposition 4.20 *Let $F : \mathscr{C}([0, 1] : \mathbb{R}^d) \to \mathbb{R}$ be Lipschitz continuous. Given a control $\{(\bar{\mu}^{1,n}_i, \bar{\mu}^{2,n}_i)\}_{i=0,\dots,n-1}$, let $\{\bar{X}^n_i\}$ and $\{\bar{Z}^n_\sigma\}$ be defined as in Construction 4.19. Then for all $n \in \mathbb{N}$ and $\sigma > 0$,*

$$-\frac{1}{n}\log E e^{-nF(Z^n_\sigma)}$$
$$\leq \inf_{\{\bar{\mu}^{1,n}_i, \bar{\mu}^{2,n}_i\}} E\left[F(\bar{Z}^n_\sigma) + \frac{1}{n}\sum_{i=0}^{n-1}\left[R\left(\bar{\mu}^{1,n}_i(\cdot)\|\theta(\cdot|\bar{X}^n_i)\right) + R\left(\bar{\mu}^{2,n}_i\|\rho_\sigma\right)\right]\right].$$

Proof We apply Theorem 4.5 with d replaced by $2d$, $\{X^n_i\}$ replaced by $\{(X^n_i, U^n_{i,\sigma})\}$, and $G : \mathscr{P}(\mathbb{R}^{2d} \times [0, 1]) \to \mathbb{R}$ defined by

$$G(\gamma) \doteq F(\varphi_\gamma), \ \gamma \in \mathscr{P}(\mathbb{R}^{2d} \times [0, 1]),$$

where $\varphi_\gamma \in \mathscr{C}([0, 1] : \mathbb{R}^d)$ is defined by

$$\varphi_\gamma(t) \doteq \int_{\mathbb{R}^{2d}\times[0,t]} (y+z)\gamma(dy\times dz\times ds) + x_0,\ t\in[0,1]$$

if $\|y+z\|$ has finite integral under γ, and φ_γ identically zero otherwise. Let $Y^n_{i,\sigma} = (X^n_i, U^n_{i,\sigma})$ and let $\bar{Y}^n_{i,\sigma}$ be its controlled analogue according to (4.29), with the appropriate replacements, and in particular $\{\bar{v}^n_i\}$ replaced by $\{(\bar{v}^n_i, \bar{w}^n_i)\}$. Let

$$\bar{\mu}^n_i(A\times B) \doteq \bar{\mu}^{1,n}_i(A)\bar{\mu}^{2,n}_i(B),\ A,B\in\mathscr{B}(\mathbb{R}^d).$$

Since we have placed restrictions on the measures μ^n (equivalently on the controls $\{(\bar{\mu}^{1,n}_i, \bar{\mu}^{2,n}_i)\}$), Theorem 4.5 yields the inequality

$$\begin{aligned} &-\frac{1}{n}\log Ee^{-nF(Z^n_\sigma)} \\ &\le \inf_{\{\bar{\mu}^{1,n}_i,\bar{\mu}^{2,n}_i\}} E\left[F(\varphi_{\bar{L}^n}) + \frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}^n_i(dv_i\times dw_i)\|\theta(dv_i|\bar{X}^n_i)\rho_\sigma(dw_i)\right)\right]. \end{aligned} \tag{4.30}$$

Here we have used that the distribution of the original process on (v_i, w_i) depends on $Y^n_{i,\sigma}$ only through X^n_i. The chain rule implies

$$R\left(\bar{\mu}^n_i(dv_i\times dw_i)\|\theta(dv_i|\bar{X}^n_i)\rho_\sigma(dw_i)\right) = R\left(\bar{\mu}^{1,n}_i(\cdot)\|\theta(\cdot|\bar{X}^n_i)\right) + R\left(\bar{\mu}^{2,n}_i\|\rho_\sigma\right).$$

Finally, from the definition of φ_γ it follows that if the relative entropy cost is finite, then $F(\varphi_{\bar{L}^n}) = F(\bar{Z}^n_\sigma)$ w.p.1. Inserting these into (4.30) completes the proof of the lemma. □

4.8.3 Perturbation of L and Its Properties

In order to characterize the limits of the relative entropy terms in (4.30), we use a perturbation of the function L introduced in (4.5). For $\sigma > 0$, let

$$L_\sigma(x,\beta) \doteq \sup_{\alpha\in\mathbb{R}^d}\left[\langle\alpha,\beta\rangle - H(x,\alpha) - \frac{\sigma^2}{2}\|\alpha\|^2\right].$$

Note that for each $x\in\mathbb{R}^d$, $\beta\mapsto L_\sigma(x,\beta)$ is the Legendre–Fenchel transform of

$$H_\sigma(x,\alpha) \doteq \log\int_{\mathbb{R}^d} e^{\langle\alpha,y\rangle}\theta_\sigma(dy|x),$$

where $\theta_\sigma(\cdot|x)$ is the distribution of $v_0(x) + w_{0,\sigma}$. The following lemma records some important properties of L_σ.

Lemma 4.21 *Assume Conditions 4.3 and 4.8 and fix $\sigma > 0$. Then the following conclusions hold.*

(a) *For all $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$,*

$$L_\sigma(x, \beta) = \inf_{b \in \mathbb{R}^d} \left[L(x, \beta - b) + \frac{\|b\|^2}{2\sigma^2} \right].$$

(b) *For all $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$, $L_\sigma(x, \beta) \le L(x, \beta)$.*

(c) *$(x, \beta) \mapsto L_\sigma(x, \beta)$ is a finite nonnegative continuous function on $\mathbb{R}^d \times \mathbb{R}^d$.*

(d) *Condition 4.8 is satisfied with L replaced by L_σ uniformly in σ in the following sense: for every compact $K \subset \mathbb{R}^d$ and $\varepsilon \in (0, 1)$ there exist $\bar\eta = \bar\eta(K, \varepsilon) \in (0, 1)$ and $\bar m = \bar m(K, \varepsilon) \in (0, \infty)$ such that whenever $\xi, \chi \in K$ satisfy $\|\xi - \chi\| \le \bar\eta$, for every $\bar\gamma \in \mathbb{R}^d$ and $\sigma > 0$ we can find $\bar\beta \in \mathbb{R}^d$ such that*

$$L_\sigma(\xi, \bar\beta) - L_\sigma(\chi, \bar\gamma) \le \varepsilon(1 + L_\sigma(\chi, \bar\gamma)), \quad \|\bar\beta - \bar\gamma\| \le \bar m(1 + L_\sigma(\chi, \bar\gamma))\|\xi - \chi\|. \tag{4.31}$$

(e) *Given $\zeta \in \mathscr{C}([0, 1] : \mathbb{R}^d)$ satisfying $I(\zeta) < \infty$, there is a $\zeta^* \in \mathscr{C}([0, 1] : \mathbb{R}^d)$ that is piecewise linear with finitely many pieces such that $\|\zeta^* - \zeta\|_\infty < \sigma$ and*

$$\int_0^1 L_\sigma(\zeta^*(t), \dot\zeta^*(t))dt \le \int_0^1 L_\sigma(\zeta(t), \dot\zeta(t))dt + \sigma \le I(\zeta) + \sigma. \tag{4.32}$$

Proof (a) For each $x \in \mathbb{R}^d$, $\beta \mapsto L_\sigma(x, \beta)$ is the Legendre–Fenchel transform of the sum of the convex functions $H(x, \cdot)$ and $\bar H_\sigma(\cdot)$, where $\bar H_\sigma(\alpha) = \sigma^2\|\alpha\|^2/2$, $\alpha \in \mathbb{R}^d$. The Legendre transform of the first function is $L(x, \cdot)$, and that of the second function is $L_\sigma(\beta) = \|\beta\|^2/2\sigma^2$, $\beta \in \mathbb{R}^d$. From Theorem 16.4 of [217] it follows that

$$L_\sigma(x, \beta) = \inf \left[L(x, \beta_1) + \frac{\|\beta_2\|^2}{2\sigma^2} : \beta_i \in \mathbb{R}^d, i = 1, 2, \beta_1 + \beta_2 = \beta \right],$$

as claimed.

(b) This is an immediate consequence of part (a).

(c) Note that Conditions 4.3 and 4.7 are satisfied with $H(x, \alpha)$ replaced by $H_\sigma(x, \alpha) \doteq H(x, \alpha) + H_\sigma(\alpha)$ and $\theta(\cdot|x)$ replaced by $\theta_\sigma(\cdot|x)$. Part (c) now follows from parts (a), (d), and (f) of Lemma 4.16.

(d) Fix a compact $K \subset \mathbb{R}^d$ and $\varepsilon \in (0, 1)$. From Condition 4.8, we can find $\eta \in (0, 1)$ and $m \in (0, \infty)$ such that whenever $\xi, \chi \in K$ satisfy $\|\xi - \chi\| \le \eta$, we can find for every $\gamma \in \mathbb{R}^d$ a $\beta \in \mathbb{R}^d$ such that

$$L(\xi, \beta) - L(\chi, \gamma) \le \frac{\varepsilon}{2}(1 + L(\chi, \gamma)), \quad \|\beta - \gamma\| \le m(1 + L(\chi, \gamma))\|\xi - \chi\|. \tag{4.33}$$

We claim that the statement in (d) holds for $\bar\eta = \eta$ and $\bar m = 2m$. To see this, suppose $\xi, \chi \in K$ satisfy $\|\xi - \chi\| \le \eta$ and $\bar\gamma \in \mathbb{R}^d$. Using part (a), we can find $\bar b \in \mathbb{R}^d$ such that

$$L_\sigma(\chi,\bar\gamma) \ge L(\chi,\bar\gamma-\bar b) + \frac{\|\bar b\|^2}{2\sigma^2} - \frac{\varepsilon}{4}. \tag{4.34}$$

Using Condition 4.8 and taking $\gamma = \bar\gamma - \bar b$, we can find $\beta \in \mathbb{R}^d$ such that (4.33) holds. Letting $\bar\beta = \beta + \bar b$, we have

$$\begin{aligned}\|\bar\beta - \bar\gamma\| &= \|\beta - \gamma\| \\ &\le m(1 + L(\chi,\bar\gamma - \bar b))\|\xi - \chi\| \\ &\le m(1 + L_\sigma(\chi,\bar\gamma) + \varepsilon/4)\|\xi - \chi\| \\ &\le 2m(1 + L_\sigma(\chi,\bar\gamma))\|\xi - \chi\|\,,\end{aligned}$$

where the second inequality follows from (4.34). This proves the second inequality in (4.31). Also, from part (a),

$$\begin{aligned}L_\sigma(\xi,\bar\beta) - L_\sigma(\chi,\bar\gamma) &\le L(\xi,\bar\beta-\bar b) + \frac{\|\bar b\|^2}{2\sigma^2} - L(\chi,\bar\gamma-\bar b) - \frac{\|\bar b\|^2}{2\sigma^2} + \frac{\varepsilon}{4} \\ &\le \frac{\varepsilon}{2}\left(1 + L(\chi,\bar\gamma-\bar b) + \frac{\|\bar b\|^2}{2\sigma^2}\right) + \frac{\varepsilon}{4} \\ &\le \varepsilon\,(1 + L_\sigma(\chi,\bar\gamma))\,,\end{aligned}$$

where the second inequality is from (4.33) and the third is from (4.34). This proves the first inequality in (4.31) and completes the proof of part (d).

(e) Recall that for all $\sigma > 0$, $H_\sigma(\cdot,\cdot)$ and θ_σ satisfy Conditions 4.3 and 4.7. Fix $\zeta \in \mathscr{C}([0,1]:\mathbb{R}^d)$ satisfying $I(\zeta) < \infty$. Note that $I_\sigma(\zeta) \doteq \int_0^1 L_\sigma(\zeta(t),\dot\zeta(t))dt \le I(\zeta)$. Then applying Lemma 4.17 to I_σ, we can find $\zeta_1^* \in \mathscr{C}([0,1]:\mathbb{R}^d)$ such that $\{\dot\zeta_1^*(t) : 0 \le t \le 1\}$ is bounded, $\|\zeta_1^* - \zeta\|_\infty < \frac{\sigma}{2}$, and (4.32) holds with ζ^* replaced by ζ_1^* and σ replaced by $\sigma/2$. The statement in part (e) now follows by taking ζ^* to be a piecewise linear approximation of ζ_1^* and using the continuity of $(x,\beta) \mapsto L_\sigma(x,\beta)$. □

4.8.4 *A Nearly Optimal Trajectory and Associated Control Sequence*

For $\varepsilon > 0$ let $\zeta \in \mathscr{C}([0,1]:\mathbb{R}^d)$ be such that

$$F(\zeta) + I(\zeta) \le \inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [F(\varphi) + I(\varphi)] + \varepsilon. \tag{4.35}$$

Here F, as in Sect. 4.8.1, is a Lipschitz continuous function from $\mathscr{C}([0,1]:\mathbb{R}^d)$ to $\mathbb{R}$, and I is the expected rate function for the system without mollification. From part (e) of Lemma 4.21, for each fixed $\sigma \in (0,1)$, we can find a $\zeta^* \in \mathscr{C}([0,1]:\mathbb{R}^d)$ that

is piecewise linear with finitely many pieces such that $\|\zeta^* - \zeta\|_\infty \le \sigma$ and (4.32) holds, i.e., $\int_0^1 L_\sigma(\zeta^*(t), \dot\zeta^*(t))dt \le I(\zeta) + \sigma$.

We now construct a control sequence to be applied to the mollified process for which the running cost is asymptotically close to the left side in (4.32) and the associated controlled process $\bar Z_\sigma^n$ tracks the nearly optimal trajectory ζ^* closely as $n \to \infty$. Let

$$K \doteq \bigcup_{t\in[0,1]} \left\{y \in \mathbb{R}^d : \|y - \zeta(t)\| \le 2\right\}.$$

Since ζ is continuous, K is compact. We apply part (d) of Lemma 4.21 with K defined as in the last display and ε as in (4.35). Thus there exist $\bar\eta \in (0, 1)$ and $\bar m \in (0, \infty)$ such that whenever $\xi, \chi \in K$ satisfy $\|\xi - \chi\| \le \bar\eta$, we can find for every $\bar\gamma \in \mathbb{R}^d$ and $\sigma \in (0, 1)$, a $\bar\beta \in \mathbb{R}^d$ such that (4.31) holds. The following lemma says that the selection of $\bar\beta$ can be done in a measurable way. Note that the choice of $\bar\eta$ and $\bar m$ depends on ε and K, but is independent of σ.

Lemma 4.22 (a) *Fix $\chi, \bar\gamma \in \mathbb{R}^d$ and $\sigma > 0$. Given $\xi \in K(\chi) \doteq \{x \in K : \|x - \chi\| \le \bar\eta\}$, define Γ_ξ to be the set of all $\bar\beta \in \mathbb{R}^d$ such that (4.31) holds. Then there is a measurable map $B : K(\chi) \to \mathbb{R}^d$ such that $B(\xi) \in \Gamma_\xi$ for all $\xi \in K(\chi)$ and $\chi \in K$.*

(b) *Let $K_{\bar M} = \{\beta \in \mathbb{R}^d : \|\beta\| \le \bar M\}$, where*

$$\bar M = \sup_{s\in[0,1],\xi\in K} \left[\bar m\|\xi - \zeta^*(s)\| \left(1 + L_\sigma(\zeta^*(s), \dot\zeta^*(s))\right) + \|\dot\zeta^*(s)\|\right].$$

Given $(\xi, \beta) \in K \times K_{\bar M}$, define $\tilde\Gamma_{(\xi,\beta)}$ to be the set of all $(\beta^1, \beta^2) \in \mathbb{R}^d \times \mathbb{R}^d$ such that

$$L(\xi, \beta^1) + \frac{1}{2\sigma^2}\|\beta^2\|^2 \le L_\sigma(\xi, \beta) + \sigma, \ \ \beta^1 + \beta^2 = \beta.$$

Then there are measurable maps $B^i : K \times K_{\bar M} \to \mathbb{R}^d$, $i = 1, 2$, such that $(B^1(\xi, \beta), B^2(\xi, \beta)) \in \tilde\Gamma_{(\xi,\beta)}$ for all $(\xi, \beta) \in K \times K_{\bar M}$.

Proof Corollary E.3 in the appendix is concerned with measurable selections. The proof of part (a) is immediate from this corollary and the continuity of $L_\sigma(\cdot, \cdot)$. The second part also follows from Corollary E.3 and the lower semicontinuity of L proved in part (b) of Lemma 4.14. Indeed, suppose $(\xi_n, \beta_n) \in K \times K_{\bar M}$ are such that $\xi_n \to \xi$ and $\beta_n \to \beta$, and let $(\beta_n^1, \beta_n^2) \in \tilde\Gamma_{(\xi_n,\beta_n)}$. Since $K \times K_{\bar M}$ is compact and L_σ is continuous, $\sup_{n\in\mathbb{N}} L_\sigma(\xi_n, \beta_n) < \infty$. Using the inequality

$$L(\xi_n, \beta_n^1) + \frac{1}{2\sigma^2}\|\beta_n^2\|^2 \le L_\sigma(\xi_n, \beta_n) + \sigma\,, \tag{4.36}$$

we see that $\{\beta_n^2\}$ is bounded, and since $\beta_n^1 + \beta_n^2 = \beta_n$, $\{\beta_n^1\}$ is bounded as well. Suppose now that $\beta_n^1 \to \beta^1$ and $\beta_n^2 \to \beta^2$ along a subsequence. Clearly $\beta^1 + \beta^2 = \beta$, and from the lower semicontinuity of L and continuity of L_σ, (4.36) holds

with $(\xi_n, \beta_n, \beta_n^1, \beta_n^2)$ replaced by $(\xi, \beta, \beta^1, \beta^2)$. Thus $(\beta^1, \beta^2) \in \tilde{\Gamma}_{(\xi,\beta)}$. Hence the assumptions of Corollary E.3 are satisfied, and the result follows. □

As shown in Appendix B, it follows from part (g) of Lemma 4.16 that there are stochastic kernels γ^i, $i = 1, 2$, from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathscr{P}(\mathbb{R}^d)$ and $\mathbb{R}^d$ to $\mathscr{P}(\mathbb{R}^d)$, respectively, such that for all $(\xi, \beta^1) \in \mathbb{R}^d \times \mathbb{R}^d$ and $\beta^2 \in \mathbb{R}^d$,

$$R\left(\gamma^1(\cdot|\xi, \beta^1) \,\|\, \theta(\cdot|\xi)\right) = L(\xi, \beta^1) \quad \text{and} \quad \int_{\mathbb{R}^d} y\gamma^1(dy|\xi, \beta^1) = \beta^1$$

and

$$R\left(\gamma^2(\cdot|\beta^2) \,\|\, \rho_\sigma(\cdot)\right) = \frac{1}{2\sigma^2}\|\beta^2\|^2 \quad \text{and} \quad \int_{\mathbb{R}^d} y\gamma^2(dy|\beta^2) = \beta^2.$$

Using these kernels, we recursively define a control sequence $\{(\bar{\mu}_i^{1,n}, \bar{\mu}_i^{2,n})\}$, controlled processes $\{(\bar{X}_i^n, \bar{U}_{i,\sigma}^n)\}$, and a stopping time κ^n as follows. We initialize with $(\bar{X}_0^n, \bar{U}_{0,\sigma}^n) = (x_0, 0)$, and set

$$\kappa^n \doteq \inf\left\{i \in \mathbb{N}_0 : \|\bar{X}_i^n - \zeta^*(i/n)\| > \bar{\eta}\right\} \wedge n.$$

At each discrete time $j = 0, \ldots, \kappa^n - 1$ we apply part (a) of Lemma 4.22 with $(\chi, \bar{\gamma}) = (\zeta^*(j/n), \dot{\zeta}^*(j/n))$. Noting that $\bar{X}_j^n \in K(\zeta^*(j/n))$, we define $\bar{\beta}_j^n = B(\bar{X}_j^n)$. Note that $\bar{\beta}_j^n \in K_{\bar{M}}$. With B^i as in part (b) of Lemma 4.22, let $\beta_j^{i,n} = B^i(\bar{X}_j^n, \bar{\beta}_j^n)$, $i = 1, 2$. Define

$$\bar{\mu}_j^{1,n}(\cdot) = 1_{\{j<\kappa^n\}}\gamma^1(\cdot|\bar{X}_j^n, \beta_j^{1,n}) + 1_{\{j\geq\kappa^n\}}\theta(\cdot|\bar{X}_j^n),$$

$$\bar{\mu}_j^{2,n}(\cdot) = 1_{\{j<\kappa^n\}}\gamma^2(\cdot|\beta_j^{2,n}) + 1_{\{j\geq\kappa^n\}}\rho_\sigma(\cdot)$$

and define $\bar{v}_j^n$, $\bar{w}_{j,\sigma}^n$, $\bar{X}_{j+1}^n$, $\bar{U}_{j+1,\sigma}^n$, $\bar{X}^n$, $\bar{U}_\sigma^n$, and $\bar{Z}_\sigma^n$ according to Construction 4.19. As in the previous proof of a Laplace lower bound, we revert to the original distributions when $\bar{X}_i^n$ wanders farther than $\bar{\eta}$ from $\zeta^*(i/n)$ to keep the relative entropy costs uniformly bounded. This will be needed when we study the convergence of the controlled processes.

Note that the choice of the control sequence ensures that for $j \in \{0, 1, \ldots, \kappa^n - 1\}$,

$$R\left(\bar{\mu}_j^{1,n}(\cdot) \,\|\, \theta(\cdot|\bar{X}_j^n)\right) = L(\bar{X}_j^n, \beta_j^{1,n}), \quad \int_{\mathbb{R}^d} y\bar{\mu}_j^{1,n}(dy) = \beta_j^{1,n} \tag{4.37}$$

and

$$R\left(\bar{\mu}_j^{2,n} \,\|\, \rho_\sigma\right) = \frac{1}{2\sigma^2}\|\beta_j^{2,n}\|^2, \quad \int_{\mathbb{R}^d} y\bar{\mu}_j^{2,n}(dy) = \beta_j^{2,n}. \tag{4.38}$$

Also for $j \in \{\kappa^n, \ldots, n-1\}$, $R(\bar{\mu}_j^{1,n} || \theta(\cdot|\bar{X}_j^n)) = R(\bar{\mu}_j^{2,n} \| \rho_\sigma) = 0$. From the choice of $(\chi, \bar{\gamma}) = (\zeta^*(j/n), \dot{\zeta}^*(j/n))$ in the definition of $\bar{\beta}_j^n$,

$$L_\sigma(\bar{X}_j^n, \bar{\beta}_j^n) - L_\sigma(\zeta^*(j/n), \dot{\zeta}^*(j/n)) \le \varepsilon \left[1 + L_\sigma(\zeta^*(j/n), \dot{\zeta}^*(j/n))\right], \quad (4.39)$$

$$\|\bar{\beta}_j^n - \dot{\zeta}^*(j/n)\| \le \bar{m} \left[1 + L_\sigma(\zeta^*(j/n), \dot{\zeta}^*(j/n))\right] \|\bar{X}_j^n - \zeta^*(j/n)\| \quad (4.40)$$

and

$$L(\bar{X}_j^n, \bar{\beta}_j^{1,n}) + \frac{1}{2\sigma^2} \|\beta_j^{2,n}\|^2 \le L_\sigma(\bar{X}_j^n, \bar{\beta}_j^n) + \sigma. \quad (4.41)$$

It follows that

$$\begin{aligned}
&E\left[\frac{1}{n}\sum_{j=0}^{n-1}\left(R\left(\bar{\mu}_j^{1,n}(\cdot)\|\theta(\cdot|\bar{X}_j^n)\right) + R\left(\bar{\mu}_j^{2,n}(\cdot)\|\rho_\sigma(\cdot)\right)\right)\right] \\
&= E\left[\frac{1}{n}\sum_{j=0}^{\kappa^n-1}\left(R\left(\bar{\mu}_j^{1,n}(\cdot)\|\theta(\cdot|\bar{X}_j^n)\right) + R\left(\bar{\mu}_j^{2,n}(\cdot)\|\rho_\sigma(\cdot)\right)\right)\right] \\
&= E\left[\frac{1}{n}\sum_{j=0}^{\kappa^n-1}\left(L(\bar{X}_j^n, \beta_j^{1,n}) + \frac{1}{2\sigma^2}\|\beta_j^{2,n}\|^2\right)\right] \\
&\le E\left[\frac{1}{n}\sum_{j=0}^{\kappa^n-1} L_\sigma(\bar{X}_j^n, \bar{\beta}_j^n)\right] + \sigma \\
&\le \frac{1}{n}\sum_{j=0}^{n-1}\left[L_\sigma(\zeta^*(j/n), \dot{\zeta}^*(j/n)) + \varepsilon\left[1 + L_\sigma(\zeta^*(j/n), \dot{\zeta}^*(j/n))\right]\right] + \sigma,
\end{aligned}$$

where the first equality follows from observing that for $j \ge \kappa^n$, the relative entropy terms in the first line are zero, the second from (4.37) and (4.38), the inequality on the third line from (4.41), and the last line from (4.39). Taking limits as $n \to \infty$ in the last display and using the continuity of L_σ and that ζ^* is piecewise linear, it follows that

$$\begin{aligned}
&\limsup_{n\to\infty} E\left[\frac{1}{n}\sum_{j=0}^{n-1}\left(R\left(\bar{\mu}_j^{1,n}(\cdot)\|\theta(\cdot|\bar{X}_j^n)\right) + R\left(\bar{\mu}_j^{2,n}(\cdot)\|\rho_\sigma(\cdot)\right)\right)\right] \\
&\le (1+\varepsilon)\int_0^1 L_\sigma(\zeta^*(t), \dot{\zeta}^*(t))dt + (\sigma + \varepsilon) \\
&\le (1+\varepsilon) I(\zeta) + 2\sigma + \varepsilon(1+\sigma), \quad (4.42)
\end{aligned}$$

where the last inequality follows on recalling that ζ^* was chosen so that (4.32) is satisfied.

We recall that Lemma 4.18 gave a bound for $Ee^{-nF(X^n)}$ in terms of $Ee^{-nF(Z^n_\sigma)}$, and that Proposition 4.20 gave a variational bound for $-\frac{1}{n}\log Ee^{-nF(Z^n_\sigma)}$. If we combine these with the last display and (4.35), then

$$\begin{aligned}\limsup_{n\to\infty} -\frac{1}{n}\log Ee^{-nF(X^n)} &\le F(\zeta)+I(\zeta)+2\sigma+\varepsilon(1+\sigma+I(\zeta))\\ &\quad+\limsup_{n\to\infty} E\left[F(\bar{Z}^n_\sigma)-F(\zeta)\right]+M^2\sigma^2/2\\ &\le \inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)}[F(\varphi)+I(\varphi)]+2\sigma+\varepsilon(2+\sigma+I(\zeta))\\ &\quad+\limsup_{n\to\infty} E\left[F(\bar{Z}^n_\sigma)-F(\zeta)\right]+M^2\sigma^2/2,\end{aligned}$$

where M, as in the statement of Lemma 4.18, is the Lipschitz constant of F. Taking the limit as $\sigma\to 0$ and then $\varepsilon\to 0$ gives

$$\begin{aligned}&\limsup_{n\to\infty} -\frac{1}{n}\log Ee^{-nF(X^n)}\\ &\quad\le \inf_{\varphi\in\mathscr{C}([0,1]:\mathbb{R}^d)}[F(\varphi)+I(\varphi)]+\limsup_{\varepsilon\to 0}\limsup_{\sigma\to 0}\limsup_{n\to\infty} E\left[F(\bar{Z}^n_\sigma)-F(\zeta)\right].\end{aligned}$$

Hence in order to complete the proof of the Laplace principle lower bound, it now suffices to argue that

$$\limsup_{\varepsilon\to 0}\limsup_{\sigma\to 0}\limsup_{n\to\infty} E\left[F(\bar{Z}^n_\sigma)-F(\zeta)\right]=0. \tag{4.43}$$

To do this we must analyze the asymptotic properties of the controls and controlled processes.

4.8.5 *Tightness and Convergence of Controlled Processes*

To prove (4.43) we will need to establish tightness and characterize the limits of $\{\bar{Z}^n_\sigma\}$. The main results that are needed have already been established in Lemmas 4.11 and 4.12. To apply these results, we first must identify correspondences between objects here [on $\mathbb{R}^d\times\mathbb{R}^d$] and those of the lemmas [on $\mathbb{R}^d$]. We recall the definitions

$$\bar{L}^n(A\times C)\doteq\int_C \bar{L}^n(A|t)dt,\quad \bar{L}^n(A|t)\doteq\delta_{(\bar{v}^n_i,\bar{w}^n_{i,\sigma})}(A)\text{ if } t\in[i/n,i/n+1/n)$$

and $\bar{\mu}^n_i(A\times B)\doteq\bar{\mu}^{1,n}_i(A)\bar{\mu}^{2,n}_i(B)$. Define random probability measures by

$$\bar{\mu}^{1,n}(A \times C) \doteq \int_C \bar{\mu}^{1,n}(A|t)dt, \quad \bar{\mu}^{1,n}(A|t) \doteq \bar{\mu}_i^{1,n}(A) \text{ if } t \in [i/n, i/n + 1/n),$$
$$\bar{\mu}^{2,n}(B \times C) \doteq \int_C \bar{\mu}^{2,n}(B|t)dt, \quad \bar{\mu}^{2,n}(B|t) \doteq \bar{\mu}_i^{2,n}(B) \text{ if } t \in [i/n, i/n + 1/n),$$

and

$$\bar{\mu}^n(A \times B \times C) \doteq \int_C \bar{\mu}^{1,n}(A|t)\bar{\mu}^{2,n}(B|t)dt, \lambda^n(A \times B \times C) \doteq \int_C \lambda^n(A \times B|t)dt,$$
$$\lambda^n(A \times B|t) \doteq \theta(A|\bar{X}_i^n)\rho_\sigma(B) \text{ if } t \in [i/n, i/n + 1/n).$$

Here $A, B \in \mathscr{B}(\mathbb{R}^d)$ and $C \in \mathscr{B}([0, 1])$. Also, let $\bar{\kappa}^n = \kappa^n/n$.

Lemma 4.23 (a) *The collection* $\{(\bar{\mu}^n, \lambda^n, \bar{\mu}^{1,n}, \bar{\mu}^{2,n}, \bar{X}^n, \bar{U}_\sigma^n, \bar{\kappa}^n)\}_{n\in\mathbb{N}}$ *is a tight family of random variables with values in*

$$\mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d \times [0, 1])^2 \times \mathscr{P}(\mathbb{R}^d \times [0, 1])^2 \times \mathscr{C}([0, 1] : \mathbb{R}^d)^2 \times [0, 1].$$

(b) *Suppose* $(\bar{\mu}^n, \lambda^n, \bar{\mu}^{1,n}, \bar{\mu}^{2,n}, \bar{X}^n, \bar{U}_\sigma^n, \bar{\kappa}^n)$ *converges along a subsequence in distribution to* $(\bar{\mu}, \lambda, \bar{\mu}^1, \bar{\mu}^2, \bar{X}, \bar{U}_\sigma, \bar{\kappa})$. *Then a.s., for every* $t \in [0, 1]$,

$$\bar{X}(t) = x_0 + \int_{\mathbb{R}^d\times[0,t]} y\bar{\mu}^1(dy \times ds) \text{ and } \bar{U}_\sigma(t) = \int_{\mathbb{R}^d\times[0,t]} z\bar{\mu}^2(dz \times ds).$$

Proof It follows from (4.42) that

$$\sup_{n\in\mathbb{N}} E\left[R\left(\bar{\mu}^n \| \lambda^n\right)\right] < \infty.$$

Also, from part (a) of Condition 4.3, for every $\alpha = (\alpha^1, \alpha^2) \in \mathbb{R}^{2d}$,

$$\sup_{x\in\mathbb{R}^d} \log \int_{\mathbb{R}^{2d}} \exp\left(\langle\alpha^1, y\rangle + \langle\alpha^1, z\rangle\right)\theta(dy|x)\rho_\sigma(dz) < \infty.$$

We can therefore apply Lemma 4.11 with $\theta(dy|x)$ replaced with $\theta(dy|x)\rho_\sigma(dz)$ and $\bar{\mu}_i^n$ now given by $\bar{\mu}_i^{1,n} \times \bar{\mu}_i^{2,n}$. The lemma implies that the families $\{\bar{\mu}^n\}$ and $\{\bar{L}^n\}$ are tight and satisfy the uniform integrability property

$$\begin{aligned}
&\lim_{M\to\infty} \limsup_{n\to\infty} E\left[\int_{\mathbb{R}^{2d}\times[0,1]} \|(y, z)\| \, 1_{\{\|(y,z)\|\geq M\}}\bar{\mu}^n(dy \times dz \times dt)\right] \\
&= \lim_{M\to\infty} \limsup_{n\to\infty} E\left[\int_{\mathbb{R}^{2d}\times[0,1]} \|(y, z)\| \, 1_{\{\|(y,z)\|\geq M\}}\bar{L}^n(dy \times dz \times dt)\right] \\
&= 0. \qquad (4.44)
\end{aligned}$$

Tightness of $\bar{X}^n$ and $\bar{U}^n_\sigma$ follows from Lemma 4.11 with the identities

$$(\bar{X}^n(t), \bar{U}^n_\sigma(t)) = (x_0, 0) + \int_{\mathbb{R}^{2d}\times[0,t]} (y, z)\bar{L}^n(dy \times dz \times ds).$$

Finally, tightness of $\{(\bar{\mu}^{1,n}, \bar{\mu}^{2,n})\}$ is immediate from that of $\{\bar{\mu}^n\}$, and the tightness of $\{\bar{\kappa}^n\}$ holds trivially due to the compactness of $[0, 1]$.

It follows from Lemma 4.12 that

$$(\bar{X}(t), \bar{U}_\sigma(t)) = (x_0, 0) + \int_{\mathbb{R}^{2d}\times[0,t]} (y, z)\bar{\mu}(dy \times dz \times ds).$$

We then use that w.p.1 $\bar{\mu}^1(dy \times ds) = \bar{\mu}(dy \times \mathbb{R}^d \times ds)$ and $\bar{\mu}^2(dz \times ds) = \bar{\mu}(\mathbb{R}^d \times dz \times ds)$ to get part (b). □

4.8.6 Completion of the Proof of the Laplace Lower Bound

We now return to the proof of (4.43). We will argue that

$$\limsup_{n\to\infty} E\left[F(\bar{Z}^n_\sigma) - F(\zeta)\right] \le h(\sigma, \varepsilon), \tag{4.45}$$

where $h : (0, \infty) \times (0, \infty) \to [0, \infty)$ satisfies $\lim_{\sigma\to 0} h(\sigma, \varepsilon) = 0$ for all $\varepsilon \in (0, 1)$. For this, by a usual subsequential argument it is enough to argue that (4.45) holds along any subsequence as in part (b) of Lemma 4.23 with a function h that is independent of the choice of the subsequence. Using the Skorohod representation theorem [Appendix A, Theorem A.8], we can assume that the convergence in part (b) of Lemma 4.23 is a.s., and without loss we can also assume that it holds along the full sequence. Then for all $t \in [0, 1]$,

$$\bar{Z}_\sigma(t) \doteq \lim_{n\to\infty} \bar{Z}^n_\sigma(t) = \lim_{n\to\infty} \bar{X}^n(t) + \lim_{n\to\infty} \bar{U}^n_\sigma(t) = \bar{X}(t) + \bar{U}_\sigma(t).$$

The following lemma estimates the difference between $\bar{Z}_\sigma$ and $\bar{X}$.

Lemma 4.24 *Let* $m(\sigma, \varepsilon) \doteq 2\sigma^2\left((1 + \varepsilon)(2\|F\|_\infty + \sigma + \varepsilon) + \sigma + \varepsilon\right)$. *For every* $\sigma > 0$,

$$E\|\bar{Z}_\sigma - \bar{X}\|^2_\infty = E\|\bar{U}_\sigma\|^2_\infty \le m(\sigma, \varepsilon).$$

Proof We use the convergence of $\bar{\mu}^{2,n}$ to $\bar{\mu}^2$ and the uniform integrability stated in (4.44). The identities in part (b) of Lemma 4.23 and an application of Fatou's lemma then imply

$$
\begin{aligned}
E\|\bar{U}_\sigma\|_\infty^2 &\le \liminf_{n\to\infty} E\left\| \int_{\mathbb{R}^d\times[0,\cdot]} z\bar{\mu}^{2,n}(dz\times ds)\right\|_\infty^2 \\
&\le \liminf_{n\to\infty} E\frac{1}{n}\sum_{j=0}^{n-1}\left\| \int_{\mathbb{R}^d} z\bar{\mu}_j^{2,n}(dz)\right\|^2 \\
&= \liminf_{n\to\infty} E\frac{1}{n}\sum_{j=0}^{\kappa^n-1}\left\|\beta_j^{2,n}\right\|^2 \\
&\le 2\sigma^2 \liminf_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}\big[L_\sigma(\zeta^*(j/n),\dot\zeta^*(j/n)) \\
&\qquad + \varepsilon\big[1+L_\sigma(\zeta^*(j/n),\dot\zeta^*(j/n))\big]\big] + 2\sigma^3,
\end{aligned}
$$

where the second inequality follows from Jensen's inequality and the third uses (4.41) and (4.39). Using the continuity of L_σ leads to

$$
\begin{aligned}
E\|\bar{U}_\sigma\|_\infty^2 &\le 2\sigma^2\left((1+\varepsilon)\int_0^1 L_\sigma(\zeta^*(t),\dot\zeta^*(t))dt\right) + 2\sigma^2(\sigma+\varepsilon) \\
&\le 2\sigma^2\left((1+\varepsilon)(I(\zeta)+\sigma)\right) + 2\sigma^2(\sigma+\varepsilon) \\
&\le 2\sigma^2\left((1+\varepsilon)(2\|F\|_\infty+\sigma+\varepsilon)\right) + 2\sigma^2(\sigma+\varepsilon),
\end{aligned}
$$

where the second inequality follows from (4.32) and the third from (4.35). This is the claim of the lemma. □

We recall that $\beta_j^{1,n}+\beta_j^{2,n}=\bar\beta_j^n$ and that $\bar\beta_j^n$ is chosen equal to $B(\bar X_j^n)$, which implies (4.31) with $(\xi,\bar\beta,\chi,\bar\gamma)=(\bar X_j^n,\bar\beta_j^n,\zeta^*(j/n),\dot\zeta^*(j/n))$ for $j\le\kappa^n-1$. It follows from (4.37), (4.38), and (4.40) that for $j=0,1,\dots,\kappa^n-1$,

$$
\begin{aligned}
&\left\| \int_{\mathbb{R}^d} y\left[\bar\mu_j^{1,n}(dy)+\bar\mu_j^{2,n}(dy)\right] - \dot\zeta^*(j/n)\right\| \\
&= \left\|\beta_j^{1,n}+\beta_j^{2,n}-\dot\zeta^*(j/n)\right\| \\
&= \left\|\bar\beta_j^n-\dot\zeta^*(j/n)\right\| \\
&\le \bar m\left(1+L_\sigma(\zeta^*(j/n),\dot\zeta^*(j/n))\right)\|\bar X_j^n-\zeta^*(j/n)\|.
\end{aligned}
$$

From the uniform integrability property in (4.44) and the a.s. convergence of $(\bar\mu^{1,n},\bar\mu^{2,n},\kappa^n)$ to $(\bar\mu^1,\bar\mu^2,\kappa)$, we have for all $t\in[0,1]$ that

$$
\int_{\mathbb{R}^d\times[0,t\wedge\kappa^n]} y\left[\bar\mu^{1,n}(dy\times ds)+\bar\mu^{2,n}(dy\times ds)\right]
$$

converges in probability to

$$\int_{\mathbb{R}^d\times[0,t\wedge\kappa]} y\left[\bar{\mu}^1(dy\times ds)+\bar{\mu}^2(dy\times ds)\right].$$

Thus for every $t\in[0,\bar{\kappa}]$,

$$\begin{aligned}
&\left\|\int_{\mathbb{R}^d\times[0,t]} y\left[\bar{\mu}^1(dy\times ds)+\bar{\mu}^2(dy\times ds)\right]-\int_0^t \dot{\zeta}^*(s)ds\right\| \\
&\quad=\lim_{n\to\infty}\left\|\int_{\mathbb{R}^d\times[0,t]} y\left[\bar{\mu}^{1,n}(dy\times ds)+\bar{\mu}^{2,n}(dy\times ds)\right]-\int_0^t \dot{\zeta}^*(s)ds\right\| \\
&\quad\le \bar{m}\int_0^t \|\bar{X}(s)-\zeta^*(s)\|\left(1+L_\sigma(\zeta^*(s),\dot{\zeta}^*(s))\right)ds.
\end{aligned}$$

For $s\in[0,1]$, let $a(s)=1+L_\sigma(\zeta^*(s),\dot{\zeta}^*(s))$ and $b(s)=\|\bar{Z}_\sigma(s)-\zeta^*(s)\|$. Then from part (b) of Lemma 4.23 and since $\bar{Z}_\sigma=\bar{X}+\bar{U}_\sigma$, the last display implies that for $t\in[0,\bar{\kappa}]$,

$$\begin{aligned}
b(t)&=\left\|\int_{\mathbb{R}^d\times[0,t]} y\left[\bar{\mu}^1(dy\times ds)+\bar{\mu}^2(dy\times ds)\right]-\int_0^t \dot{\zeta}^*(s)ds\right\| \\
&\le \bar{m}\int_0^t \|\bar{X}(s)-\zeta^*(s)\|a(s)ds \\
&\le \bar{m}\int_0^t b(s)a(s)ds+\bar{m}\|\bar{X}-\bar{Z}_\sigma\|_\infty\int_0^1 a(s)ds.
\end{aligned}$$

Using that [again by (4.32) and (4.35)] $\int_0^1 a(s)ds\le 1+2\|F\|_\infty+\sigma+\varepsilon$, for all $\varepsilon,\sigma\in(0,1)$, Gronwall's lemma [Lemma E.2] implies

$$\begin{aligned}
\|\bar{Z}_\sigma(\cdot\wedge\bar{\kappa})-\zeta^*(\cdot\wedge\bar{\kappa})\|_\infty &\le \bar{m}\|\bar{X}-\bar{Z}_\sigma\|_\infty\int_0^1 a(s)ds\int_0^1 e^{\bar{m}\int_0^t a(s)ds}dt \\
&\le \bar{m}(2\|F\|_\infty+3)\|\bar{X}-\bar{Z}_\sigma\|_\infty e^{\bar{m}(2\|F\|_\infty+3)} \\
&= \bar{m}_1\|\bar{X}-\bar{Z}_\sigma\|_\infty, \qquad (4.46)
\end{aligned}$$

where $\bar{m}_1\doteq\bar{m}(2\|F\|_\infty+3)e^{\bar{m}(2\|F\|_\infty+3)}$. Finally, with M as in Lemma 4.18 equal to the Lipschitz constant of F, we obtain

$$\begin{aligned}
\limsup_{n\to\infty} E\left[F(\bar{Z}^n_\sigma)-F(\zeta)\right] &\le \limsup_{n\to\infty} E\left[M\|\bar{Z}^n_\sigma-\zeta\|_\infty\wedge 2\|F\|_\infty\right] \\
&\le M\left[\bar{m}_1 E\|\bar{X}-\bar{Z}_\sigma\|_\infty+\sigma\right]+2\|F\|_\infty P(\bar{\kappa}<1) \\
&\le M\left(\bar{m}_1(m(\sigma,\varepsilon))^{1/2}+\sigma\right)+2\|F\|_\infty P(\bar{\kappa}<1).
\end{aligned}$$

For the second inequality we use the fact that $\|\zeta-\zeta^*\|_\infty\le\sigma$, and we partition according to whether $\bar{\kappa}<1$, using (4.46) when this is not the case. The third inequality follows from Lemma 4.24.

The last quantity we need to control is $P(\bar{\kappa} < 1)$. Since the convergence in path space is with respect to the uniform topology, we have

$$P(\bar{\kappa} < 1) \le P\left(\left\| \bar{Z}_\sigma(\cdot \wedge \bar{\kappa}) - \zeta^*(\cdot \wedge \bar{\kappa}) \right\|_\infty \ge \bar{\eta}/2\right) \le \frac{4}{\bar{\eta}^2} \bar{m}_1^2 \, m(\sigma, \varepsilon),$$

where the first inequality follows from the definition of κ^n and the second inequality uses Chebyshev's inequality, (4.46) and Lemma 4.24. Thus (4.45) holds with

$$h(\sigma, \varepsilon) = M\left(\bar{m}_1 (m(\sigma, \varepsilon))^{1/2} + \sigma\right) + 2\|F\|_\infty \frac{4}{\bar{\eta}^2(\varepsilon)} \bar{m}_1^2 \, m(\sigma, \varepsilon),$$

where we write $\bar{\eta} = \bar{\eta}(\varepsilon)$ to emphasize its dependence on ε (recall from Lemma 4.21 that $\bar{\eta}$ does not depend on σ). From the definition of $m(\sigma, \varepsilon)$ in Lemma 4.24 it follows that $\lim_{\sigma \to 0} m(\sigma, \varepsilon) = 0$ for all $\varepsilon \in (0, 1)$. This proves that $\lim_{\sigma \to 0} h(\sigma, \varepsilon) = 0$ for every $\varepsilon \in (0, 1)$, and hence (4.43) holds. With the limit (4.43) demonstrated, we have completed the proof of the Laplace lower bound under Condition 4.8. □

4.9 Notes

Among the very first papers to treat models of this type are those of Wentzell [245–248], Freidlin and Wentzell [140], and Azencott and Ruget [8]. The conditions we use are weaker than those of the cited references. The statement of assumptions and conclusions for this chapter parallels that of Chapter 6 of [97], though the proof is different and is directly analogous to the way Cramér's theorem was obtained in Chap. 3. In particular, we first prove a process-level generalization of Sanov's theorem for the driving noises that define the recursive stochastic model. Then, assuming appropriate integrability conditions on the distribution of these noises, one obtains Laplace asymptotics for the process of interest by viewing it as a mapping on this process-level empirical measure. This leads to a somewhat simpler proof, though in all approaches the mollification aspect of the proof under Conditions 4.3 and 4.8 is technical. We also note that the proof given here corrects a gap in a proof in [97], in that the constant M defined on page 205 of [97] depends on σ, and therefore the claim in equation (6.65) of [97] is not really established.

As noted several times already, the analysis of the discrete time model of this chapter is in many ways more difficult than that of the corresponding continuous-time models, largely because for continuous time models the noise enters in an additive and affine manner.

Chapter 5
Moderate Deviations for Recursive Markov Systems

In this chapter we consider $\mathbb{R}^d$-valued discrete time processes of the same form as in Chap. 4, but instead of analyzing the large deviation behavior, we consider deviations closer to the LLN limit. Since this will require centering on the limit, we assume that the process model has the form

$$X_{i+1}^n \doteq X_i^n + \frac{1}{n} b(X_i^n) + \frac{1}{n} \nu_i(X_i^n), \quad X_0^n = x_0, \tag{5.1}$$

where $\{\nu_i(\cdot)\}_{i \in \mathbb{N}_0}$ are iid random vector fields as in Chap. 4 but with zero mean, and $b : \mathbb{R}^d \to \mathbb{R}^d$ is continuously differentiable.

As in Chap. 4, we consider the continuous time piecewise linear interpolations $\{X^n(t)\}_{0 \le t \le T}$ with $X^n(i/n) = X_i^n$. Under moment conditions that are weaker than those of Chap. 4, there is a law of large numbers limit $X^0 \in \mathscr{C}([0, T] : \mathbb{R}^d)$. Closely related to X^0 is the noiseless version of (5.1) obtained by setting $\nu_i(\cdot) = 0$, which is denoted by $\{X_i^{n,0}\}_{n \in \mathbb{N}_0}$ with piecewise linear interpolation $\{X^{n,0}(t)\}_{0 \le t \le T}$. We introduce a scaling sequence $\varkappa(n)$ that satisfies

$$\varkappa(n) \to 0 \text{ and } \varkappa(n)n \to \infty, \tag{5.2}$$

and study the rescaled difference

$$Y^n \doteq \sqrt{\varkappa(n)n}(X^n - X^{n,0}).$$

Since under Condition 5.1, b is Lipschitz continuous, we have

$$\left\| X^0 - X^{n,0} \right\|_\infty \doteq \sup_{t \in [0,T]} \left\| X^0(t) - X^{n,0}(t) \right\| = O(1/n).$$

Thus

$$\sqrt{\varkappa(n)n} \left\| X^0 - X^{n,0} \right\|_\infty = O(\sqrt{\varkappa(n)/n}),$$

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_5

and hence Y^n behaves the same asymptotically as $\sqrt{\varkappa(n)n}(X^n - X^0)$. It will be shown that under weaker conditions on the noise $\nu_i(\cdot)$ than were used in Chap. 4, Y^n satisfies the large deviation principle on $\mathscr{C}([0, T] : \mathbb{R}^d)$ with a "Gaussian"-type rate function. As is customary for this type of scaling, we refer to this as moderate deviations.

While one might expect the proof of the moderate deviations result to be similar to that of the corresponding large deviations result, there are important differences. For example, the tightness proof is significantly more complicated in the case of moderate deviations than for the case of large deviations. In Chap. 4 we were able to establish an a priori bound on certain relative entropy costs associated with any sequence of nearly minimizing controls. Under this boundedness of the relative entropy costs, empirical measures of the controlled driving noises as well as the controlled processes themselves were tight. With the scaling used for moderate deviations, and even with the information that the analogous relative entropy costs decay like $O(1/\varkappa(n)n)$, tightness of the empirical measures of the driving noise does not hold. Instead, one must consider the empirical measures of conditional means of the noises, and additional effort is required to show that the limits of these measures determine the limit of the controlled processes. This extra difficulty arises for moderate deviations (even with the vanishing relative entropy costs), because the noise is amplified by the factor $\sqrt{\varkappa(n)n}$ in the definition of Y^n.

A second way in which the proofs for large and moderate deviations differ is in their treatment of degenerate noise, i.e., problems in which the support of $\nu_i(\cdot)$ is not all of $\mathbb{R}^d$. As we saw in Chap. 4, this leads to significant difficulties in the proof of the large deviation lower bound, requiring a somewhat involved mollification argument. In contrast, the proof in the setting of moderate deviations, though more involved than the nondegenerate case, is much more straightforward.

As a potential application of these results we mention their usefulness in the design and analysis of Monte Carlo schemes for events whose probability is small but not very small. For such problems, the performance of standard Monte Carlo may not be adequate, especially if the quantity must be computed for many different parameter settings, as in, say, an optimization problem. Another instance is the situation in which the cost for even a single sample is very high, as for example in the case of stochastic partial differential equations. Then accelerated Monte Carlo may be of interest, and as is well known, such schemes (e.g., importance sampling and splitting) benefit from the use of information contained in the large deviation rate function as part of the algorithm design (e.g., [28, 76, 114, 116]). In a situation in which one considers events of small but not too small probability, one may find the moderate deviation approximation both adequate and relatively easy to apply, since moderate deviations lead to situations in which the objects needed to design an efficient scheme can be explicitly constructed in terms of solutions to the linear–quadratic regulator. These issues were first explored in [101]. Other moderate deviation analyses are presented in Chaps. 10 and 13, and an example of how moderate deviation approximations can be used to construct accelerated Monte Carlo schemes is given in Sect. 17.5.

5.1 Assumptions, Notation, and Theorem Statement

Let

$$X_{i+1}^n \doteq X_i^n + \frac{1}{n} b(X_i^n) + \frac{1}{n} v_i(X_i^n), \quad X_0^n = x_0,$$

where the $\{v_i(\cdot)\}_{i \in \mathbb{N}_0}$ are zero-mean iid vector fields whose distribution is given by a stochastic kernel $\theta(dy|x)$ on $\mathbb{R}^d$ given $\mathbb{R}^d$. For $\alpha \in \mathbb{R}^d$, define

$$H_c(x, \alpha) \doteq \log E e^{\langle \alpha, v_i(x) \rangle}.$$

The subscript c reflects the fact that this log moment generating function uses the centered distribution $\theta(\cdot|x)$, rather than $H(x, \alpha) = H_c(x, \alpha) + \langle \alpha, b(x) \rangle$ as in Chap. 4. We use the following.

Condition 5.1 (a) *There exist* $\lambda > 0$ *and* $K_{mgf} < \infty$ *such that*

$$\sup_{x \in \mathbb{R}^d} \sup_{\|\alpha\| \le \lambda} H_c(x, \alpha) \le K_{mgf}. \tag{5.3}$$

(b) *The mapping* $x \mapsto \theta(dy|x)$ *from* $\mathbb{R}^d$ *to* $\mathscr{P}(\mathbb{R}^d)$ *is continuous with respect to the topology of weak convergence.*

(c) $b(x)$ *is continuously differentiable, and the norms of both* $b(x)$ *and its derivative are uniformly bounded by a constant* $K_b < \infty$.

Throughout this chapter we let $\|\alpha\|_A^2 \doteq \langle \alpha, A\alpha \rangle$ for $\alpha \in \mathbb{R}^d$ and a symmetric nonnegative definite matrix A. Define

$$A_{ij}(x) \doteq \int_{\mathbb{R}^d} y_i y_j \theta(dy|x),$$

and note that the weak continuity of $\theta(dy|x)$ with respect to x and (5.3) ensures that $A(x)$ is continuous in x and that its norm, $\|A(x)\| \doteq \sup_{v:\|v\|=1} \|A(x)v\|$, is uniformly bounded by some constant K_A. Note that

$$\frac{\partial H_c(x, 0)}{\partial \alpha_i} = \int_{\mathbb{R}^d} y_i \theta(dy|x) = 0 \tag{5.4}$$

and

$$\frac{\partial^2 H_c(x, 0)}{\partial \alpha_i \partial \alpha_j} = \int_{\mathbb{R}^d} y_i y_j \theta(dy|x) = A_{ij}(x) \tag{5.5}$$

for all $i, j \in \{1, \ldots, d\}$ and $x \in \mathbb{R}^d$, and that $A(x)$ is nonnegative definite and symmetric. It follows that for $x \in \mathbb{R}^d$,

$$A(x) = Q(x) \Lambda(x) Q^T(x),$$

where $Q(x)$ is an orthogonal matrix whose columns are the eigenvectors of $A(x)$, and $\Lambda(x)$ is the diagonal matrix consisting of the eigenvalues of $A(x)$ in descending order. The map $x \mapsto \Lambda(x)$ is continuous. In what follows, we define $\Lambda^{-1}(x)$ to be the diagonal matrix with diagonal entries each equal to the inverse of the corresponding eigenvalue for the positive eigenvalues, and equal to ∞ for the zero eigenvalues. Then when we write

$$\|\alpha\|^2_{A^{-1}(x)} = \|\alpha\|^2_{Q(x)\Lambda^{-1}(x)Q^T(x)}, \tag{5.6}$$

we mean a value of ∞ for $\alpha \in \mathbb{R}^d$ not in the linear span of the eigenvectors corresponding to the positive eigenvalues, and the standard value for vectors $\alpha \in \mathbb{R}^d$ in that linear span. (Note that even if the definition of $A^{-1}(x)$ is ambiguous, in that for α in the range of $A(x)$ there may be more than one v such that $A(x)v = \alpha$, the value of $\|\alpha\|^2_{A^{-1}(x)}$ is not ambiguous. Indeed, since the eigenvectors can be assumed orthogonal, for all such v, $\langle v, \alpha\rangle$ coincides with $\langle \bar{v}, \alpha\rangle$, where $\bar{v}$ is the solution in the span of eigenvectors corresponding to positive eigenvalues.) Condition 5.1(a) implies there exist $K_{DA} < \infty$ and $\lambda_{DA} \in (0, \lambda]$ such that

$$\sup_{x\in\mathbb{R}^d} \sup_{\|\alpha\|\le\lambda_{DA}} \max_{i,j,k} \left| \frac{\partial^3 H_c(x,\alpha)}{\partial\alpha_i \partial\alpha_j \partial\alpha_k} \right| \le \frac{K_{DA}}{d^3}, \tag{5.7}$$

and consequently for all $\|\alpha\| \le \lambda_{DA}$ and all $x \in \mathbb{R}^d$,

$$\frac{1}{2}\|\alpha\|^2_{A(x)} - \|\alpha\|^3 K_{DA} \le H_c(x,\alpha) \le \frac{1}{2}\|\alpha\|^2_{A(x)} + \|\alpha\|^3 K_{DA}. \tag{5.8}$$

Define the continuous time piecewise linear interpolation of X^n_i by

$$X^n(t) \doteq X^n_i + [X^n_{i+1} - X^n_i](nt - i), \quad t \in [i/n, i/n + 1/n].$$

In addition, define

$$X^{n,0}_{i+1} = X^{n,0}_i + \frac{1}{n} b(X^{n,0}_i), \quad X^{n,0}_0 = x_0,$$

and let $X^{n,0}(t)$ be the analogously defined continuous time interpolation. Then $X^{n,0} \to X^0$ in $\mathscr{C}([0,T] : \mathbb{R}^d)$, where for $t \in [0,T]$,

$$X^0(t) = \int_0^t b(X^0(s))ds + x_0.$$

Since $Ev_i(x) = 0$ for all $x \in \mathbb{R}^d$, we know that $X^n \to X^0$ in $\mathscr{C}([0,T] : \mathbb{R}^d)$ in probability.

In Chap. 4 we showed, under significantly stronger assumptions, that $X^n(t)$ satisfies a large deviation principle on $\mathscr{C}([0,T] : \mathbb{R}^d)$ with scaling sequence $r(n) = 1/n$. Letting I_L denote this rate function (with L for large deviation), it takes the form

$$I_L(\phi) \doteq \inf\left[\int_0^T L_c(\phi(s), u(s))ds : \phi(t) = x_0 + \int_0^t b(\phi(s))ds + \int_0^t u(s)ds,\ t \in [0,T]\right],$$

where

$$L_c(x,\beta) \doteq \sup_{\alpha\in\mathbb{R}^d} [\langle\alpha,\beta\rangle - H_c(x,\alpha)] \tag{5.9}$$

is the Legendre transform of $H_c(x,\alpha)$. We see that I_L coincides with the rate function of Chap. 4, because

$$L(x,\beta) \doteq \sup_{\alpha\in\mathbb{R}^d} [\langle\alpha,\beta\rangle - H_c(x,\alpha) - \langle\alpha, b(x)\rangle] = L_c(x,\beta - b(x)),$$

so that $L_c(\phi(s), u(s)) = L_c(\phi(s), \dot\phi(s) - b(\phi(s))) = L(\phi(s), \dot\phi(s))$.

Assume that $\varkappa(n)$ satisfies (5.2):

$$\varkappa(n) \to 0 \text{ and } \varkappa(n)n \to \infty.$$

We define the rescaled difference

$$Y^n(t) \doteq \sqrt{\varkappa(n)n}(X^n(t) - X^{n,0}(t)).$$

Let $Db(x)$ denote the matrix of partial derivatives $(Db(x))_{ij} = \partial b_i(x)/\partial x_j$, and let $A^{1/2}(x)$ be the unique nonnegative definite square root of $A(x)$.

Theorem 5.2 *Assume Condition 5.1. Then $\{Y^n\}_{n\in\mathbb{N}}$ satisfies the Laplace principle on $\mathscr{C}([0,T]:\mathbb{R}^d)$ with scaling sequence $\varkappa(n)$ and rate function*

$$I_M(\phi) = \inf\left[\frac{1}{2}\int_0^T \|u(t)\|^2\,dt : \phi(t) = \int_0^t Db(X^0(s))\phi(s)ds + \int_0^t A^{1/2}(X^0(s))u(s)ds,\ t\in[0,T]\right].$$

The function I_M is essentially the same as what one would obtain using a linear approximation around the law of large numbers limit X^0 of the dynamics and a quadratic approximation of the costs in I_L. By our convention, proofs will be given for the case $T = 1$, with only notational modifications needed for the general case. An alternative form of the rate function that is consistent with expressions we use for continuous time models is

$$I_M(\phi) = \inf_{u\in U_\phi}\left[\frac{1}{2}\int_0^1 \|u(t)\|^2\,dt\right], \tag{5.10}$$

where for $\phi \in \mathscr{AC}_0([0,1]:\mathbb{R}^d)$ U_ϕ is the subset of $\mathscr{L}^2([0,1]:\mathbb{R}^d)$ given by

$$U_\phi \doteq \left\{ u : \phi(\cdot) = \int_0^\cdot Db(X^0(s))\phi(s)ds + \int_0^\cdot A^{1/2}(X^0(s))u(s)ds \right\}, \tag{5.11}$$

and U_ϕ is the empty set otherwise. Since $\{\phi : I_M(\phi) \le K\}$ is the image of the compact set $\{u : \int_0^1 \|u(t)\|^2\, dt \le K\}$ (with the weak topology on $\mathscr{L}^2([0,1]:\mathbb{R}^d)$) under a continuous mapping, I_M has compact level sets. To complete the proof of Theorem 5.2, we must show that for every bounded and continuous F,

$$\lim_{n\to\infty} -\varkappa(n) \log E\left[e^{-\frac{1}{\varkappa(n)}F(Y^n)}\right] = \inf_{\phi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [I_M(\phi) + F(\phi)]. \tag{5.12}$$

The argument will follow the same layout as that used in Chaps. 3 and 4. In Sect. 5.2, a representation is derived for the exponential integral in (5.12), and in Sect. 5.3, tightness of empirical measures and identification of limits for these measures and controlled processes are carried out. It is here that the moderate deviation problem differs most from the corresponding large deviation problem, in that the definition of the empirical measures is not analogous to the definition used in Chap. 4. The Laplace principle upper and lower bounds that together imply (5.12) are proved in Sects. 5.4 and 5.5, respectively.

5.2 The Representation

As usual, the first step is to identify a useful representation for the Laplace functionals. Owing to the moderate deviation scaling, the construction of the controlled processes differs slightly from that of Chap. 4.

Construction 5.3 Suppose we are given a probability measure $\mu^n \in \mathscr{P}((\mathbb{R}^d)^n)$ and decompose it in terms of conditional distributions $[\mu^n]_{i|1,\ldots,i-1}$ on the ith variable given variables 0 through $i-1$:

$$\mu^n(dv_0 \times \cdots \times dv_{n-1}) = [\mu^n]_0(dv_0)[\mu^n]_{1|0}(dv_1|v_0) \\ \times \cdots \times [\mu^n]_{n-1|0,\ldots,n-2}(dv_{n-1}|v_0,\ldots,v_{n-2}).$$

Let $\{\bar{v}_i^n\}_{i=0,\ldots,n-1}$ be random variables defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and with joint distribution μ^n. Thus conditioned on $\bar{\mathscr{F}}_i^n \doteq \sigma(\bar{v}_j^n,\, j=0,\ldots,i-1)$, $\bar{v}_i^n$ has distribution $\bar{\mu}_i^n(dv_i) \doteq [\mu^n]_{i|0,\ldots,i-1}(dv_i|\bar{v}_0^n,\ldots,\bar{v}_{i-1}^n)$. The collection $\{\bar{\mu}_i^n\}_{i=0,\ldots,n-1}$ will be called a control. Controlled processes $\bar{X}^n$, $\bar{Y}^n$ and measures $\bar{M}^n$ are recursively constructed as follows. Let $\bar{X}_0^n = x_0$, and for $i = 1,\ldots,n$, define $\bar{X}_i^n$ recursively by

$$\bar{X}^n_{i+1} = \bar{X}^n_i + \frac{1}{n} b(\bar{X}^n_i) + \frac{1}{n} \bar{v}^n_i.$$

When $\{\bar{X}^n_i\}_{i=1,\ldots,n}$ has been constructed, define

$$\bar{Y}^n_{i+1} = \bar{Y}^n_i + \sqrt{\frac{\varkappa(n)}{n}} \left(b(\bar{X}^n_i) - b(X^{n,0}_i) \right) + \sqrt{\frac{\varkappa(n)}{n}} \bar{v}^n_i, \quad \bar{Y}^n_0 = 0. \tag{5.13}$$

Note that

$$\bar{X}^n_i - X^{n,0}_i = \frac{1}{\sqrt{\varkappa(n)n}} \bar{Y}^n_i, \ i = 0, 1, \ldots, n. \tag{5.14}$$

Let $\bar{X}^n$ and $\bar{Y}^n$ be as in (4.2) the piecewise linear interpolations with $\bar{X}^n(i/n) = \bar{X}^n_i$ and $\bar{Y}^n(i/n) = \bar{Y}^n_i$. Define also the interpolated conditional mean (provided it exists)

$$\bar{w}^n(t) \doteq \int_{\mathbb{R}^d} y \bar{\mu}^n_i(dy), \quad t \in [i/n, i/n + 1/n),$$

the scaled conditional mean

$$w^n(t) \doteq \sqrt{\varkappa(n)n} \bar{w}^n(t),$$

and random measures on $\mathbb{R}^d \times [0, 1]$ by

$$\bar{M}^n(dw \times dt) \doteq \delta_{w^n(t)}(dw)dt = \delta_{\sqrt{\varkappa(n)n}\bar{w}^n(t)}(dw)dt.$$

Note that $\bar{M}^n$ is the empirical measure of the scaled conditional means and not, in contrast to Chap. 4, of the $\bar{v}^n_i$, scaled or otherwise. This additional "averaging" will be needed for tightness. We will refer to this construction when we are given $\{\bar{\mu}^n_i\}_{i=1,\ldots,n}$ to identify associated $\bar{X}^n$, $\bar{Y}^n$, w^n and $\bar{M}^n$. By Theorem 4.5, for every bounded, continuous $F : \mathscr{C}([0, 1] : \mathbb{R}^d) \to \mathbb{R}$,

$$\begin{aligned} &- \varkappa(n) \log E \left[e^{-\frac{1}{\varkappa(n)} F(Y^n)} \right] \\ &= \inf_{\{\bar{\mu}^n_i\}} E \left[\sum_{i=0}^{n-1} \varkappa(n) R(\bar{\mu}^n_i(\cdot) \| \theta(\cdot|\bar{X}^n_i)) + F(\bar{Y}^n) \right]. \end{aligned} \tag{5.15}$$

5.3 Tightness and Limits for Controlled Processes

5.3.1 *Tightness and Uniform Integrability*

When the moment-generating function is finite for all α, a variational characterization of its Legendre transform in terms of relative entropy is proved in Lemma 4.16. In this

chapter we will need only the following inequality, which holds when the moment-generating function is finite in some neighborhood of the origin. Recall that $L_c(x,\cdot)$ is the Legendre–Fenchel transform of $H_c(x,\cdot)$.

Lemma 5.4 *Assume Condition 5.1. Then for all $x, \beta \in \mathbb{R}^d$,*

$$L_c(x,\beta) \le R(\eta(\cdot)\|\theta(\cdot|x))$$

for all $\eta \in \mathscr{P}(\mathbb{R}^d)$ satisfying $\int_{\mathbb{R}^d} y\eta(dy) = \beta$.

Proof Fix $x, \beta \in \mathbb{R}^d$ and consider any $\eta \in \mathscr{P}(\mathbb{R}^d)$ that satisfies $\int_{\mathbb{R}^d} y\eta(dy) = \beta$. If $R(\eta(\cdot)\|\theta(\cdot|x)) = \infty$, the lemma is automatically true, so we assume without loss that $R(\eta(\cdot)\|\theta(\cdot|x)) < \infty$. From (5.3) we have

$$\int_{\mathbb{R}^d} e^{\lambda\|y\|/d^{1/2}}\theta(dy|x) \le 2de^{K_{\mathrm{mgf}}} < \infty.$$

Therefore, recalling the inequality $ab \le e^a + \ell(b)$ for $a, b \ge 0$, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} \frac{\lambda}{d^{1/2}}\|y\| \frac{d\eta(\cdot)}{d\theta(\cdot|x)}(y)\theta(dy|x) \\
&\le \int_{\mathbb{R}^d} e^{\lambda\|y\|/d^{1/2}}\theta(dy|x) + \int_{\mathbb{R}^d} \ell\left(\frac{d\eta(\cdot)}{d\theta(\cdot|x)}(y)\right)\theta(dy|x) \\
&\le 2de^{K_{\mathrm{mgf}}} + R(\eta(\cdot)\|\theta(\cdot|x)).
\end{aligned}$$

Consequently, for all $\alpha \in \mathbb{R}^d$,

$$\int_{\mathbb{R}^d} \|\alpha\|\,\|y\| \frac{d\eta(\cdot)}{d\theta(\cdot|x)}(y)\theta(dy|x) \le \frac{d^{1/2}\|\alpha\|}{\lambda}\left(2de^{K_{\mathrm{mgf}}} + R(\eta(\cdot)\|\theta(\cdot|x))\right) < \infty. \tag{5.16}$$

Define the bounded continuous function $F_K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by

$$F_K(y,\alpha) = \begin{cases} \langle \alpha, y\rangle & \text{if } |\langle \alpha, y\rangle| \le K, \\ \frac{K\langle \alpha, y\rangle}{|\langle \alpha, y\rangle|} & \text{otherwise.} \end{cases}$$

From (5.16) and the dominated convergence theorem, we have

$$\lim_{K\to\infty} \int_{\mathbb{R}^d} F_K(y,\alpha)\eta(dy) = \left\langle \alpha, \int_{\mathbb{R}^d} y\eta(dy)\right\rangle = \langle \alpha, \beta\rangle.$$

Another application of the monotone convergence theorem gives

$$\lim_{K\to\infty} \int_{\{y:\langle \alpha, y\rangle < 0\}} e^{F_K(y,\alpha)}\theta(dy|x) = \int_{\{y:\langle \alpha, y\rangle < 0\}} e^{\langle \alpha, y\rangle}\theta(dy|x),$$

and the monotone convergence theorem gives

$$\lim_{K\to\infty}\int_{\{y:\langle\alpha,y\rangle\geq 0\}} e^{F_K(y,\alpha)}\theta(dy|x) = \int_{\{y:\langle\alpha,y\rangle\geq 0\}} e^{\langle\alpha,y\rangle}\theta(dy|x).$$

Thus

$$\lim_{K\to\infty}\log\left(\int_{\mathbb{R}^d} e^{F_K(y,\alpha)}\theta(dy|x)\right) = H_c(x,\alpha).$$

By the Donsker–Varadhan variational formula (Proposition 2.2), for every $K \in (0,\infty)$ and $\alpha \in \mathbb{R}^d$,

$$R(\eta(\cdot)\|\,\theta(\cdot|x)) \geq \int_{\mathbb{R}^d} F_K(y,\alpha)\eta(dy) - \log\left(\int_{\mathbb{R}^d} e^{F_K(y,\alpha)}\theta(dy|x)\right).$$

Sending $K\to\infty$ and taking the supremum over $\alpha\in\mathbb{R}^d$ yields

$$R(\eta(\cdot)\|\,\theta(\cdot|x)) \geq \sup_{\alpha\in\mathbb{R}^d}[\langle\alpha,\beta\rangle - H_c(x,\alpha)] = L_c(x,\beta),$$

which completes the proof of the lemma. □

Theorem 5.5 *Assume Condition 5.1 and*

$$\sup_{n\in\mathbb{N}}\left[\varkappa(n)nE\left[\frac{1}{n}\sum_{i=0}^{n-1}R(\bar{\mu}_i^n(\cdot)\|\,\theta(\cdot|\bar{X}_i^n))\right]\right] \leq K_E < \infty. \tag{5.17}$$

Let $\{\bar{M}^n\}_{n\in\mathbb{N}}$, $\{\bar{w}^n\}_{n\in\mathbb{N}}$, *and* $\{w^n\}_{n\in\mathbb{N}}$ *be defined as in Construction 5.3. Then*

$$\sup_{n\in\mathbb{N}} E\left[\int_0^1 \sqrt{\varkappa(n)n}\,\|\bar{w}^n(t)\|\,dt\right] = \sup_{n\in\mathbb{N}} E\left[\int_0^1 \|w^n(t)\|\,dt\right] < \infty.$$

In addition, $\{\bar{M}^n\}$ *is tight (as a sequence of random probability measures) and uniformly integrable in the sense that*

$$\lim_{C\to\infty}\limsup_{n\to\infty} E\left[\int_{\mathbb{R}^d\times[0,1]} \|w\|\,1_{\{\|w\|\geq C\}}\bar{M}^n(dw\times dt)\right] = 0. \tag{5.18}$$

Proof We assume without loss of generality that $\inf_{n\in\mathbb{N}}\{\sqrt{\varkappa(n)n}\} > 0$. Let $B\in(0,\infty)$ be such that $B \leq \lambda_{DA}\inf_{n\in\mathbb{N}}\{\sqrt{\varkappa(n)n}\}$, so that $\lambda_{DA} \geq B/\sqrt{\varkappa(n)n}$ for all n. Recall L_c from (5.9), and let $\bar{K} \doteq \lambda_{DA}K_{DA} + K_A/2$, where we recall that K_A is the bound on $A(x)$ and K_{DA} was introduced in (5.8). Let e_i denote the standard unit vectors in $\mathbb{R}^d$. Then for all $i\in\{1,\ldots,d\}$ and each choice of $\pm$,

$$
\begin{aligned}
&\varkappa(n)nL_c(x,\beta)\\
&\quad=\sup_{\alpha\in\mathbb{R}^d}\left[\sqrt{\varkappa(n)n}\left\langle\alpha,\sqrt{\varkappa(n)n}\beta\right\rangle-\varkappa(n)nH_c(x,\alpha)\right]\\
&\quad\geq\pm\sqrt{\varkappa(n)n}\left\langle\frac{B}{\sqrt{\varkappa(n)n}}e_i,\sqrt{\varkappa(n)n}\beta\right\rangle-\varkappa(n)nH_c\left(x,\pm\frac{B}{\sqrt{\varkappa(n)n}}e_i\right)\\
&\quad\geq\pm B\sqrt{\varkappa(n)n}\beta_i-\frac{1}{2}B^2\left\|A(x)\right\|-B^2\lambda_{DA}K_{DA}\\
&\quad\geq\pm B\sqrt{\varkappa(n)n}\beta_i-B^2\bar{K},
\end{aligned}
$$

where the first inequality follows from making a specific choice of α and the second uses (5.8). If we multiply both sides by $|\beta_i|$, sum on i, and then divide by $\sum_{i=1}^d|\beta_i|$, we obtain

$$
d\varkappa(n)nL_c(x,\beta)+dB^2\bar{K}\geq Bd\sqrt{\varkappa(n)n}\frac{\|\beta\|^2}{\sum_{i=1}^d|\beta_i|}\geq B\sqrt{\varkappa(n)n}\left\|\beta\right\|. \tag{5.19}
$$

To slightly simplify the notation, we let $s^n(t)\doteq\lfloor nt\rfloor/n$, where $\lfloor a\rfloor$ is the integer part of a. Using the bound relating L_c and relative entropy from Lemma 5.4 together with (5.17), we obtain

$$
\begin{aligned}
d\left(\frac{K_E}{B}+B\bar{K}\right)&\geq\frac{d\varkappa(n)n}{B}E\left[\int_0^1L_c\left(\bar{X}^n\left(s^n(t)\right),\bar{w}^n(t)\right)dt\right]+dB\bar{K}\\
&\geq E\left[\int_0^1\sqrt{\varkappa(n)n}\left\|\bar{w}^n(t)\right\|dt\right],
\end{aligned}\tag{5.20}
$$

which proves the first statement in the theorem. Since by Theorem 2.10 the mapping

$$
m\mapsto\int_{\mathbb{R}^d\times[0,1]}\|w\|\,m(dw\times dt)
$$

defines a tightness function on $\mathscr{P}\left(\mathbb{R}^d\times[0,1]\right)$, it follows from Lemma 2.9 and the first claim that $\{\bar{M}^n\}_{n\in\mathbb{N}}$ is tight.

For the uniform integrability, let $C\in(1,\infty)$ be arbitrary. We note that the estimates in (5.19) and (5.20) hold for any B and n such that $B\leq\lambda_{DA}\{\sqrt{\varkappa(n)n}\}$. Consider n large enough that

$$
\min\{\lambda_{DA},1\}\geq\frac{C}{\sqrt{\varkappa(n)n}}.
$$

Then for such n, the estimates (5.19) and (5.20) hold with $B=1$ and $B=C$. Recalling $\bar{K}\doteq\lambda_{DA}K_{DA}+K_A/2$ and applying (5.20) with $B=1$, we have for such n,

$$E\left[\int_0^1 \sqrt{\varkappa(n)n}\left\|\bar{w}^n(s)\right\| ds\right] \leq K^* \doteq d\left(K_E + \frac{1}{2}K_A + \lambda_{DA}K_{DA}\right),$$

and therefore

$$E\left[\int_0^1 1_{\{\sqrt{\varkappa(n)n}\|\bar{w}^n(s)\|>C^2\}} ds\right] \leq \frac{K^*}{C^2}.$$

In the following bound, (5.19) with $C = B$ is used to get the first inequality and the last display, and (5.20) with $B = 1$ for the third inequality:

$$\begin{aligned}
&CE\left[\int_{\mathbb{R}^d\times[0,1]} \|w\| 1_{\{\|w\|\geq C^2\}} \bar{M}^n(dw \times dt)\right] \\
&\leq E\left[d\int_0^1 1_{\{\|w^n(s)\|>C^2\}}\left(\varkappa(n)nL_c\left(\bar{X}^n\left(s^n(t)\right), \bar{w}^n(t)\right) + C^2\bar{K}\right) dt\right] \\
&\leq d\varkappa(n)nE\left[\int_0^1 L_c\left(\bar{X}^n\left(s^n(t)\right), \bar{w}^n(t)\right) dt\right] + C^2 d\bar{K}E\left[\int_0^1 1_{\{\|w^n(t)\|>C^2\}} dt\right] \\
&\leq K^* d\left(1+\bar{K}\right).
\end{aligned}$$

This proves the claimed uniform integrability. □

5.3.2 *Identification of Limits*

The following theorem is a law of large numbers type result for the difference between the noises and their conditional means, and is the most complicated part of the analysis.

Theorem 5.6 *Assume Condition 5.1 and (5.17). Consider the sequence $\{\bar{v}_i^n\}$ of controlled noises and $\{\bar{w}^n(i/n)\}$ of means of the controlled noises as in Construction 5.3. For $i \in \{1, \ldots, n\}$ let*

$$W_i^n \doteq \frac{1}{n}\sum_{j=0}^{i-1}\sqrt{\varkappa(n)n}\left(\bar{v}_j^n - \bar{w}^n\left(j/n\right)\right).$$

Then for all $\delta > 0$,

$$\lim_{n\to\infty} P\left\{\max_{i\in\{1,\ldots,n\}}\left\|W_i^n\right\| \geq \delta\right\} = 0.$$

Proof According to (5.17),

$$\frac{1}{n}\sum_{i=0}^{n-1} E[R(\bar{\mu}_i^n(\cdot)\|\,\theta(\cdot|\bar{X}_i^n))] \leq \frac{K_E}{\varkappa(n)n}.$$

Because of this the (random) Radon–Nikodym derivatives,

$$f_i^n(y) = \frac{d\bar{\mu}_i^n(\cdot)}{d\theta(\cdot|\bar{X}_i^n)}(y)$$

are well defined and can be selected in a measurable way [79, Theorem V.58]. We will control the magnitude of the noise when the Radon–Nikodym derivative is large by bounding

$$\frac{1}{n}\sum_{i=0}^{n-1} E[1_{\{f_i^n(\bar{v}_i^n)\geq r\}} \left\| \bar{v}_i^n \right\|]$$

for large $r \in (0, \infty)$.

From the bound on the moment-generating function (5.3) [see (3.12)], we obtain

$$\sup_{x\in\mathbb{R}^d} \int_{\mathbb{R}^d} e^{\frac{\lambda}{\sqrt{d}}\|y\|}\theta(dy|x) \leq 2de^{K_{\text{mgf}}}.$$

Let $\sigma = \min\{\lambda/2\sqrt{d}, 1\}$ and recall $\ell(b) \doteq b\log b - b + 1$. Then

$$\frac{1}{n}\sum_{i=0}^{n-1} E\left[1_{\{f_i^n(\bar{v}_i^n)\geq r\}} \left\| \bar{v}_i^n \right\|\right] = \frac{1}{n}\sum_{i=0}^{n-1} E\left[\int_{\{y: f_i^n(y)\geq r\}} \|y\| \, f_i^n(y)\theta(dy|\bar{X}_i^n)\right],$$

and the bound $ab \leq e^a + \ell(b)$ for $a, b \geq 0$ with $a = \sigma\|y\|$ and $b = f_i^n(y)$ gives that for all i,

$$\begin{aligned} &E\left[\int_{\{y: f_i^n(y)\geq r\}} \|y\| \, f_i^n(y)\theta(dy|\bar{X}_i^n)\right] \\ &\quad \leq \frac{1}{\sigma} E\left[\int_{\{y: f_i^n(y)\geq r\}} e^{\sigma\|y\|}\theta(dy|\bar{X}_i^n)\right] + \frac{1}{\sigma} E\left[\int_{\{y: f_i^n(y)\geq r\}} \ell(f_i^n(y))\bar{\mu}_i^n(dy)\right]. \end{aligned}$$

Since $\ell(b) \geq 0$ for all $b \geq 0$, we have

$$\begin{aligned} E\left[\int_{\{y: f_i^n(y)\geq r\}} \ell\left(f_i^n(y)\right)\theta(dy|\bar{X}_i^n)\right] &\leq E\left[\int_{\mathbb{R}^d} \ell(f_i^n(y))\theta(dy|\bar{X}_i^n)\right] \\ &= E[R(\bar{\mu}_i^n(\cdot)\| \, \theta(\cdot|\bar{X}_i^n))], \end{aligned}$$

and by Hölder's inequality,

$$
\begin{aligned}
&E\left[\int_{\{y:f_i^n(y)\geq r\}} e^{\sigma\|y\|}\theta(dy|\bar{X}_i^n)\right] \\
&\quad\leq E\left[\left(\int_{\mathbb{R}^d} 1_{\{f_i^n(y)\geq r\}}\theta(dy|\bar{X}_i^n)\right)^{1/2}\left(\int_{\mathbb{R}^d} e^{2\sigma\|y\|}\theta(dy|\bar{X}_i^n)\right)^{1/2}\right] \\
&\quad= E\left[\theta(\{y : f_i^n(y)\geq r\}|\bar{X}_i^n)^{1/2}\right]\left(2de^{K_{\mathrm{mgf}}}\right)^{1/2}.
\end{aligned}
$$

In addition, for all $r > 1$, Markov's inequality gives

$$
\begin{aligned}
\theta(\{y : f_i^n(y)\geq r\}|\bar{X}_i^n) &\leq \frac{1}{r\log r}\int \log(f_i^n(y))f_i^n(y)\theta(dy|\bar{X}_i^n) \\
&= \frac{R(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n))}{r\log r}.
\end{aligned}
$$

The last four displays give the bound

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=0}^{n-1} E\left[\int_{\{f_i^n(y)\geq r\}} \|y\|\, f_i^n(y)\theta(dy|\bar{X}_i^n)\right] \\
&\quad\leq \frac{1}{\sigma}\left(2de^{K_{\mathrm{mgf}}}\right)^{1/2}\frac{1}{n}\sum_{i=0}^{n-1} E\left[\left(\frac{R(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n))}{r\log r}\right)^{1/2}\right] \\
&\qquad+\frac{1}{\sigma}\frac{1}{n}\sum_{i=0}^{n-1} E[R(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n))].
\end{aligned}
$$

Since by Jensen's inequality,

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=0}^{n-1} E\left[\left(\frac{R(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n))}{r\log r}\right)^{1/2}\right] \\
&\quad\leq \left(\frac{1}{r\log r}\right)^{1/2}\left(\frac{1}{n}\sum_{i=0}^{n-1} E[R(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n))]\right)^{1/2},
\end{aligned}
$$

we obtain the overall bound

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=0}^{n-1} E\left[1_{\{f_i^n(\bar{v}_i^n)\geq r\}}\,\|\bar{v}_i^n\|\right] \\
&\quad\leq \frac{1}{\sigma}\left(2de^{K_{\mathrm{mgf}}}\right)^{1/2}\left(\frac{1}{r\log r}\right)^{1/2}\left(\frac{1}{n}\sum_{i=0}^{n-1} E[R(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n))]\right)^{1/2} \\
&\qquad+\frac{1}{\sigma}\frac{1}{n}\sum_{i=0}^{n-1} E[R(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n))]
\end{aligned}
$$

$$\leq \frac{1}{\sigma}\frac{K_E^{1/2}}{\sqrt{\varkappa(n)n}}\left(2de^{K_{\text{mgf}}}\right)^{1/2}\left(\frac{1}{r\log r}\right)^{1/2}+\frac{1}{\sigma}\frac{K_E}{\varkappa(n)n}. \tag{5.21}$$

Using this result, we can complete the proof. Define

$$\xi_i^{n,r} \doteq \begin{cases} \bar{\nu}_i^n \text{ if } f_i^n(\bar{\nu}_i^n) < r, \\ 0 \quad \text{otherwise.} \end{cases}$$

For all $\delta > 0$,

$$\begin{aligned} &P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}\sqrt{\varkappa(n)n}\left(\bar{\nu}_i^n-\bar{w}^n\left(\frac{i}{n}\right)\right)\right\|\geq 3\delta\right\} \\ &\quad\leq P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}\sqrt{\varkappa(n)n}(\bar{\nu}_i^n-\xi_i^{n,r})\right\|\geq\delta\right\} \\ &\qquad+P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}\sqrt{\varkappa(n)n}\left(\xi_i^{n,r}-\int_{\{y:f_i^n(y)<r\}}y\bar{\mu}_i^n(dy)\right)\right\|\geq\delta\right\} \\ &\qquad+P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}\sqrt{\varkappa(n)n}\left(\bar{w}^n\left(\frac{i}{n}\right)-\int_{\{y:f_i^n(y)<r\}}y\bar{\mu}_i^n(dy)\right)\right\|\geq\delta\right\}. \end{aligned}$$

The first term satisfies

$$\begin{aligned} &P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}\sqrt{\varkappa(n)n}(\bar{\nu}_i^n-\xi_i^{n,r})\right\|\geq\delta\right\} \\ &\quad\leq\frac{1}{\delta}\sqrt{\varkappa(n)n}\frac{1}{n}\sum_{i=0}^{n-1}E\left[\left\|\bar{\nu}_i^n-\xi_i^{n,r}\right\|\right] \\ &\quad=\frac{1}{\delta}\sqrt{\varkappa(n)n}\frac{1}{n}\sum_{i=0}^{n-1}E\left[1_{\{f_i^n(\bar{\nu}_i^n)\geq r\}}\left\|\bar{\nu}_i^n\right\|\right]. \end{aligned}$$

The norm in the second term is a submartingale in k, and so by Doob's submartingale inequality [see (D.1)],

$$\begin{aligned} &P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}\sqrt{\varkappa(n)n}\left(\xi_i^{n,r}-\int_{\{y:f_i^n(y)<r\}}y\bar{\mu}_i^n(dy)\right)\right\|\geq\delta\right\} \\ &\quad\leq\frac{1}{\delta^2}E\left[\left\|\frac{1}{n}\sum_{i=0}^{n-1}\sqrt{\varkappa(n)n}\left(\xi_i^{n,r}-\int_{\{y:f_i^n(y)<r\}}y\bar{\mu}_i^n(dy)\right)\right\|^2\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\delta^2}\frac{\varkappa(n)}{n}\sum_{i=0}^{n-1} E\left[\left\|\left(\xi_i^{n,r} - \int_{\{y:f_i^n(y)<r\}} y\bar{\mu}_i^n(dy)\right)\right\|^2\right] \\
&\le \frac{1}{\delta^2}\frac{\varkappa(n)}{n}\sum_{i=0}^{n-1} E\left[\left\|\xi_i^{n,r}\right\|^2\right] \\
&= \frac{1}{\delta^2}\frac{\varkappa(n)}{n}\sum_{i=0}^{n-1} E\left[\int_{\{y:f_i^n(y)<r\}} \|y\|^2 f_i^n(y)\theta(dy|\bar{X}_i^n)\right] \\
&\le \frac{r}{\delta^2}\frac{\varkappa(n)}{n}\sum_{i=0}^{n-1} E\left[\int_{\mathbb{R}^d} \|y\|^2 \theta(dy|\bar{X}_i^n)\right] \\
&\le \frac{r}{\delta^2}\varkappa(n)K_{\mu,2},
\end{aligned}
$$

where

$$
K_{\mu,2} = \sup_{x\in\mathbb{R}^d}\int_{\mathbb{R}^d} \|y\|^2\,\theta(dy|x) < \infty,
$$

and the finiteness is due to (5.3). We can use Jensen's inequality with the third term and get the same bound that was proved for the first:

$$
\begin{aligned}
&P\left\{\max_{k=0,\dots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}\sqrt{\varkappa(n)n}\left(\bar{w}^n\left(\frac{i}{n}\right) - \int_{\{y:f_i^n(y)<r\}} y\bar{\mu}_i^n(dy)\right)\right\| \ge \delta\right\} \\
&\le \frac{1}{\delta}\sqrt{\varkappa(n)n}\frac{1}{n}\sum_{i=0}^{n-1} E\left[\left\|\left(\bar{w}^n\left(\frac{i}{n}\right) - \int_{\{y:f_i^n(y)<r\}} y\bar{\mu}_i^n(dy)\right)\right\|\right] \\
&= \frac{1}{\delta}\sqrt{\varkappa(n)n}\frac{1}{n}\sum_{i=0}^{n-1} E\left[\left\|\int_{\{y:f_i^n(y)\ge r\}} y\bar{\mu}_i^n(dy)\right\|\right] \\
&\le \frac{1}{\delta}\sqrt{\varkappa(n)n}\frac{1}{n}\sum_{i=0}^{n-1} E\left[\int_{\{y:f_i^n(y)\ge r\}} \|y\|\,\bar{\mu}_i^n(dy)\right] \\
&= \frac{1}{\delta}\sqrt{\varkappa(n)n}\frac{1}{n}\sum_{i=0}^{n-1} E\left[1_{\{f_i^n(\bar{v}_i^n)\ge r\}}\left\|\bar{v}_i^n\right\|\right].
\end{aligned}
$$

Combining the bounds for these three terms with (5.21) gives

$$
\begin{aligned}
&P\left\{\max_{k=0,\dots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}\sqrt{\varkappa(n)n}\left(\bar{v}_i^n - \bar{w}^n\left(\frac{i}{n}\right)\right)\right\| \ge 3\delta\right\} \\
&\le \frac{2}{\delta}\sqrt{\varkappa(n)n}\frac{1}{n}\sum_{i=0}^{n-1} E\left[1_{\{f_i^n(\bar{v}_i^n)\ge r\}}\left\|\bar{v}_i^n\right\|\right] + \frac{r}{\delta^2}\varkappa(n)K_{\mu,2}
\end{aligned}
$$

$$\leq \frac{2}{\sigma\delta} K_E^{1/2} \left(2de^{K_{\mathrm{mgf}}}\right)^{1/2} \left(\frac{1}{r\log r}\right)^{1/2} + \frac{2}{\sigma\delta}\frac{K_E}{\sqrt{\varkappa(n)n}} + \varkappa(n)\frac{r}{\delta^2}K_{\mu,2}.$$

Sending $n \to \infty$ and then $r \to \infty$ (and using $\varkappa(n) \to 0$, $\varkappa(n)n \to \infty$) gives

$$P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}\sqrt{\varkappa(n)n}\left(\bar{v}_i^n - \bar{w}^n\left(\frac{i}{n}\right)\right)\right\| \geq 3\delta\right\} \to 0$$

as $n \to \infty$, which completes the proof. □

The next result identifies the weak limits of controlled processes. We recall that for a probability measure γ on $\mathbb{R}^d \times [0,1]$, the marginal distribution on the second coordinate is denoted by $[\gamma]_2$, and the conditional distribution on the first coordinate given the second is given by $[\gamma]_{1|2}$. Thus for Borel sets A and B,

$$\gamma(\mathbb{R}^d \times B) = [\gamma]_2(B) \text{ and } \gamma(A \times B) = \int_B [\gamma]_{1|2}(A|s)[\gamma]_2(ds).$$

Theorem 5.7 *Let $\{\bar{\mu}_i^n\}_{i=1,\ldots,n}$ be a sequence of controls, and define the corresponding random variables as in Construction 5.3. Assume Condition 5.1 and that (5.17) holds for some $K_E < \infty$. Then $\{(\bar{M}^n, \bar{Y}^n)\}_{n\in\mathbb{N}}$ is tight in $\mathscr{P}(\mathbb{R}^d \times [0,1]) \times \mathscr{C}([0,1] : \mathbb{R}^d)$. Consider a subsequence (keeping the index n for convenience) such that $\{(\bar{M}^n, \bar{Y}^n)\}$ converges weakly to $(\bar{M}, \bar{Y})$. Then with probability 1, $[\bar{M}]_2(dt)$ is Lebesgue measure and*

$$\bar{Y}(t) = \int_0^t Db(X^0(s))\bar{Y}(s)ds + \int_0^t w(s)ds, \tag{5.22}$$

where

$$w(t) = \int_{\mathbb{R}^d} w[\bar{M}]_{1|2}(dw\,|t\,). \tag{5.23}$$

In addition,

$$\liminf_{n\to\infty} \varkappa(n)nE\left[\frac{1}{n}\sum_{i=0}^{n-1} R(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n))\right] \geq E\left[\int_0^1 \frac{1}{2}\,\|w(s)\|^2_{A^{-1}(X^0(s))}\,ds\right]. \tag{5.24}$$

The proof of this theorem is lengthy. After some preliminary discussion, several lemmas will be presented. After stating and proving the lemmas, we will return to complete the argument for Theorem 5.7.

It was shown in Theorem 5.5 that $\{\bar{M}^n\}_{n\in\mathbb{N}}$ is tight. If $\bar{M}$ is any weak limit of a subsequence of $\{\bar{M}^n\}_{n\in\mathbb{N}}$, then since for all n the second marginal of $\bar{M}^n(dw \times dt)$ is Lebesgue measure, it follows that $[\bar{M}]_2(dt)$ is Lebesgue measure with probability 1.

The ultimate goal is to show that $\bar{Y}^n \to \bar{Y}$ weakly in $\mathscr{C}([0,1]:\mathbb{R}^d)$, where $\bar{Y}(t)$ is given by (5.22) in terms of the weak limit $\bar{M}$. To achieve this, we introduce the following processes, which serve as intermediate steps. Let $\check{Y}_0^n = 0$ and

$$\check{Y}_{i+1}^n = \check{Y}_i^n + \sqrt{\frac{\varkappa(n)}{n}}\left(b\left(X_i^{n,0} + \frac{1}{\sqrt{\varkappa(n)n}}\check{Y}_i^n\right) - b\left(X_i^{n,0}\right)\right) + \sqrt{\frac{\varkappa(n)}{n}}\bar{w}^n\left(\frac{i}{n}\right),$$

together with its continuous time linear interpolation defined for $t \in [i/n, i/n + 1/n]$ by

$$\check{Y}^n(t) = (i+1-nt)\check{Y}_i^n + (nt-i)\check{Y}_{i+1}^n.$$

Also let

$$\hat{Y}^n(t) = \int_0^t Db\left(X^0(s)\right)\hat{Y}^n(s)ds + \int_0^t w^n(s)ds, \tag{5.25}$$

where

$$w^n(t) = \int_{\mathbb{R}^d} w[\bar{M}]_{1|2}^n(dw\,|\,t)$$

as in Construction 5.3. Then both $\check{Y}^n$ and $\hat{Y}^n$ are random variables taking values in $\mathscr{C}([0,1]:\mathbb{R}^d)$. Note that $\bar{Y}^n$ differs from $\check{Y}^n$, because $\bar{Y}^n$ is driven by the actual noises and $\check{Y}^n$ is driven by their conditional means. While the driving terms of $\hat{Y}^n$ and $\check{Y}^n$ are the same [recall that $\sqrt{\varkappa(n)n}\bar{w}^n(t) = w^n(t)$], they differ in that $\check{Y}^n$ is still a linear interpolation of a discrete time process, whereas $\hat{Y}^n$ satisfies an ODE. We will show that along any subsequence where $\bar{M}^n \to \bar{M}$ weakly,

$$\bar{Y}^n - \check{Y}^n \to 0,\ \check{Y}^n - \hat{Y}^n \to 0,\ \text{and } \hat{Y}^n \to \bar{Y}$$

in $\mathscr{C}([0,1]:\mathbb{R}^d)$, all in distribution, where $\bar{Y}$ is the unique solution of (5.22).

To show that $\hat{Y}^n \to \bar{Y}$, we show that $\{\hat{Y}^n\}$ is tight in $\mathscr{C}([0,1]:\mathbb{R}^d)$ and use the mapping defined by (5.25) from $\int_0^\cdot w^n$ to $\hat{Y}^n$. Recall that $\sup_{x\in\mathbb{R}^d}\|Db(x)\| \le K_b$. The following lemma uses the uniform integrability of $\{\bar{M}^n\}$ given in Theorem 5.5 to prove tightness of $\{\hat{Y}^n\}$.

Lemma 5.8 *Assume Conditions 5.1 and (5.17). The sequence $\{\hat{Y}^n\}$ defined in (5.25) in terms of the measures $\{\bar{M}^n\}$ via Construction 5.3 is tight in $\mathscr{C}([0,1]:\mathbb{R}^d)$, as is $\{\int_0^\cdot w^n ds\}$.*

Proof Tightness of $\{\int_0^\cdot w^n ds\}$ is a consequence of the fact that for $\delta, C \in (0,\infty)$,

$$\limsup_{n\to\infty} P\left(\sup_{|s-t|\le\delta}\int_s^t \left\|w^n(r)\right\| dr > \varepsilon\right) \le \delta\frac{C}{\varepsilon} + \frac{1}{\varepsilon}T(C),$$

where

$$T(C) \doteq \limsup_{n\to\infty} E\left[\int_0^1 1_{\{\|w^n(t)\|>C\}} \left\|w^n(t)\right\| dt\right]$$
$$= \limsup_{n\to\infty} E\left[\int_{\{\|w\|>C\}} \|w\| \, \bar{M}^n(dw \times dt)\right],$$

and the fact that by Theorem 5.5, $T(C) \to 0$ as $C \to \infty$. For tightness of $\{\hat{Y}^n\}$ it suffices to check that the map from $\mathscr{C}([0,1] : \mathbb{R}^d)$ to itself that takes $z \in \mathscr{C}([0,1] : \mathbb{R}^d)$ to the unique solution of

$$\phi(t) = \int_0^t Db(X^0(s))\phi(s)ds + z(t), \ t \in [0,1],$$

is continuous. However, this continuity follows directly from Gronwall's inequality. □

We still need to show that $\hat{Y}^n$ converges to $\bar{Y}$. This also relies on the uniform integrability given by Theorem 5.5.

Lemma 5.9 *Assume Conditions 5.1 and (5.17). Let the sequence $\{\hat{Y}^n(t)\}$ be defined by (5.25) and consider a weakly convergent subsequence $\{(\hat{Y}^n, \bar{M}^n)\}$ with limit $(\hat{Y}, \bar{M})$. Then w.p.1, $\hat{Y} = \bar{Y}$, where $\bar{Y}$ is defined by (5.22)–(5.23).*

Proof We can write

$$\hat{Y}^n(t) = \int_0^t Db(X^0(s))\hat{Y}^n(s)ds + \int_0^t \int_{\mathbb{R}^d} w\bar{M}^n(dw \times ds).$$

The uniform integrability proved in Theorem 5.5 and that $[\bar{M}]_2$ is Lebesgue measure w.p.1 will be used. The latter implies $E\bar{M}(\mathbb{R}^d \times \{t\}) = 0$ for $t \in [0,1]$. Sending $n \to \infty$ and using the definition of $w(s)$ in (5.23) gives

$$\hat{Y}(t) = \int_0^t Db(X^0(s))\hat{Y}(s)ds + \int_0^t \int_{\mathbb{R}^d} w\bar{M}(dw \times ds)$$
$$= \int_0^t Db(X^0(s))\hat{Y}(s)ds + \int_0^t w(s)ds.$$

By uniqueness of the solution, $\hat{Y} = \bar{Y}$ follows. □

It remains to show that $\bar{Y}^n - \check{Y}^n \to 0$ and $\check{Y}^n - \hat{Y}^n \to 0$. We begin with $\bar{Y}^n - \check{Y}^n \to 0$. Recall that the difference between $\bar{Y}^n$ and $\check{Y}^n$ is that the first is driven by the actual noises, and the second is driven by their conditional means. The following discrete version of Gronwall's inequality will be used to prove $\bar{Y}^n - \check{Y}^n \to 0$. A proof can be found in [83, p. 283].

Lemma 5.10 *If $\{z_n\}$, $\{u_n\}$, and $\{v_n\}$ are nonnegative sequences defined for $n \in \mathbb{N}_0$ that satisfy*

$$z_k \le v_k + \sum_{k=0}^{n-1} u_i z_i,$$

then

$$z_k \le v_k + \sum_{k=0}^{n-1} u_i v_i \exp\left\{\sum_{i=k+1}^{n-1} u_j\right\}.$$

Lemma 5.11 *Assume Conditions 5.1 and (5.17). Then* $\check{Y}^n - \bar{Y}^n \to 0$ *in probability.*

Proof Recall from (5.13) and (5.14) that

$$\bar{Y}_k^n = \sum_{i=0}^{k-1} \sqrt{\frac{\varkappa(n)}{n}} \left(b\left(X_i^{n,0} + \frac{1}{\sqrt{\varkappa(n)n}} \bar{Y}_i^n \right) - b\left(X_i^{n,0} \right) \right) + \sum_{i=0}^{k-1} \sqrt{\frac{\varkappa(n)}{n}} \bar{v}_i^n$$

and

$$\check{Y}_k^n = \sum_{i=0}^{k-1} \sqrt{\frac{\varkappa(n)}{n}} \left(b\left(X_i^{n,0} + \frac{1}{\sqrt{\varkappa(n)n}} \check{Y}_i^n \right) - b\left(X_i^{n,0} \right) \right) + \sum_{i=0}^{k-1} \sqrt{\frac{\varkappa(n)}{n}} \bar{w}^n \left(\frac{i}{n} \right),$$

so with W_k^n defined as in Theorem 5.6,

$$\left\| \bar{Y}_k^n - \check{Y}_k^n \right\| \le \left\| W_k^n \right\| + \sum_{i=0}^{k-1} \frac{K_b}{n} \left\| \bar{Y}_i^n - \check{Y}_i^n \right\|.$$

Using Lemma 5.10 gives, for $k \le n$,

$$\begin{aligned} \left\| \bar{Y}_k^n - \check{Y}_k^n \right\| &\le \left\| W_k^n \right\| + \sum_{i=0}^{k-1} \left\| W_i^n \right\| \frac{K_b}{n} \exp\left\{ \frac{K_b}{n} (k - i - 1) \right\} \\ &\le (1 + K_b e^{K_b}) \max_{i \in \{0,1,\ldots,k\}} \left\| W_i^n \right\|. \end{aligned}$$

From Theorem 5.6, we have $\max_{i \in \{1,\ldots,n\}} \left\| W_i^n \right\| \to 0$ in probability, and therefore

$$\max_{i \in \{1,\ldots,n\}} \left\| \bar{Y}_i^n - \check{Y}_i^n \right\| \to 0,$$

and hence $\sup_{t \in [0,1]} \left\| \bar{Y}^n(t) - \check{Y}^n(t) \right\| \to 0$ in probability. □

To complete the proof of the convergence we need to show that $\check{Y}^n - \hat{Y}^n \to 0$. Recall that these two processes have the same driving terms but different drifts, in that $\hat{Y}^n$ satisfies the ODE

$$\hat{Y}^n(t) = \int_0^t Db(X^0(s))\hat{Y}^n(s)ds + \int_0^t w^n(s)ds,$$

while $\check{Y}^n$ is the linear interpolation of the discrete time process defined by $\check{Y}_0^n = 0$ and

$$\check{Y}_{i+1}^n = \check{Y}_i^n + \sqrt{\frac{\varkappa(n)}{n}}\left(b\left(X_i^{n,0} + \frac{1}{\sqrt{\varkappa(n)n}}\check{Y}_i^n\right) - b\left(X_i^{n,0}\right)\right) + \frac{1}{n}w^n\left(\frac{i}{n}\right).$$

However, essentially the same arguments as those used in Lemma 5.8 to show tightness of $\{\hat{Y}^n\}$ can be used to prove tightness of $\{\check{Y}^n\}$, and then it easily follows as in Lemma 5.9 that any limit will satisfy the same ODE (5.22) as the limit of $\{\hat{Y}^n\}$, and therefore $\check{Y}^n - \hat{Y}^n \to 0$ follows.

Combining $\bar{Y}^n - \check{Y}^n \to 0$, $\check{Y}^n - \hat{Y}^n \to 0$, and $\hat{Y}^n \to \bar{Y}$ demonstrates that along the subsequence where $\bar{M}^n \to \bar{M}$ weakly, $\bar{Y}^n \to \bar{Y}$ in distribution, which implies that along this subsequence, $(\bar{M}^n, \bar{Y}^n) \to (\bar{M}, \bar{Y})$ weakly. We have already shown that with probability 1, $[\bar{M}]_2(dt)$ is Lebesgue measure and

$$\bar{Y}(t) = \int_0^t Db(X^0(s))\bar{Y}(s)ds + \int_0^t \int_{\mathbb{R}^d} w[\bar{M}]_{1|2}(dw\,|s)ds,$$

so the proof of convergence (i.e., the first part of Theorem 5.7) is complete.

To finish the proof of Theorem 5.7, we must prove the bound (5.24). Recall the notation $s^n(t) \doteq \lfloor nt \rfloor / n$, and note from (5.14) that the weak convergence of $\bar{Y}^n$ implies

$$\sup_{t\in[0,1]} \left\| \bar{X}^n(s^n(t)) - X^0(t) \right\| \to 0 \text{ in probability.} \tag{5.26}$$

Now define random measures on $\mathbb{R}^d \times \mathbb{R}^d \times [0,1]$ by

$$\gamma^n(dx \times dw \times dt) = \delta_{\bar{X}^n(s^n(t))}(dx)\,\bar{M}^n(dw \times dt).$$

Note that the tightness of $\{\gamma^n\}$ follows easily from (5.26) and the tightness of $\left\{\bar{M}^n\right\}$. Thus given any subsequence, we can choose a further subsequence (again we will retain n as the index for simplicity) along which $\{\gamma^n\}$ converges weakly to some $\mathscr{P}\left(\mathbb{R}^d \times \mathbb{R}^d \times [0,1]\right)$-valued random variable γ with

$$[\gamma]_{2,3}(dw \times dt) = \bar{M}(dw \times dt),$$

where $[\gamma]_{2,3}$ is the second and third marginal of γ. If we establish (5.24) for this subsequence, it holds for the original one using a standard argument by contradiction. For $\sigma > 0$, let

$$G_\sigma^{X^0} \doteq \left\{(x, w, t) : \left\|x - X^0(t)\right\| \le \sigma\right\}$$

be closed sets centered on $X^0(t)$ in the x variable, and note that by (5.26) and weak convergence, for all $\sigma > 0$,

$$1 = \limsup_{n\to\infty} E\left[\gamma^n\left(G_\sigma^{X^0}\right)\right] \le E\left[\gamma\left(G_\sigma^{X^0}\right)\right].$$

Thus

$$E\left[\gamma\left(\bigcap_{n\in\mathbb{N}} G_{1/n}^{X^0}\right)\right] = 1,$$

so with probability 1, γ puts all its mass on $\left\{(x,w,t) : x = X^0(t)\right\}$. Therefore, with probability 1, for a.e. (w,t) under $[\gamma]_{2,3}(dw\times dt)$,

$$[\gamma]_{1|2,3}(dx|\, w,t) = \delta_{X^0(t)}(dx).$$

Combined with the fact that the second marginal of $\bar{M}(dw\times dt)$ is Lebesgue measure, this gives

$$\gamma(dx\times dw\times dt) = \delta_{X^0(t)}(dx)\,\bar{M}(dw|\,t)\,dt. \tag{5.27}$$

For $\kappa\in(0,\infty)$, define

$$\bar{L}_\kappa(x,\beta) \doteq \sup_{\alpha\in\mathbb{R}^d}\left[\langle\alpha,\beta\rangle - \frac{1}{2}\|\alpha\|^2_{A(x)} - \frac{1}{2\kappa}\|\alpha\|^2\right].$$

Using (5.8),

$$\begin{aligned}
&\varkappa(n)nL_c\left(x, \frac{1}{\sqrt{\varkappa(n)n}}\beta\right)\\
&\quad= \sup_{\alpha\in\mathbb{R}^d}\left[\sqrt{\varkappa(n)n}\langle\alpha,\beta\rangle - \varkappa(n)nH_c(x,\alpha)\right]\\
&\quad\ge \sup_{\alpha\in\mathbb{R}^d}\left[\sqrt{\varkappa(n)n}\langle\alpha,\beta\rangle - \frac{\varkappa(n)n}{2}\|\alpha\|^2_{A(x)} - \varkappa(n)nK_{DA}\|\alpha\|^3\right]\\
&\quad\ge \sup_{\alpha\in\mathbb{R}^d}\left[\langle\alpha,\beta\rangle - \|\alpha\|^2_{A(x)} - \frac{1}{2\kappa}\|\alpha\|^2 - \frac{K_{DA}}{\sqrt{\varkappa(n)n}}\|\alpha\|^3\right].
\end{aligned} \tag{5.28}$$

Let K_1 be an arbitrary compact subset of $\mathbb{R}^d$. Since $\|\alpha\|^2$ is superlinear, there exists another compact set K_2 of $\mathbb{R}^d$, depending only on κ and K_1, such that whenever $\beta\in K_1$ and $x\in\mathbb{R}^d$,

$$\begin{aligned}
\sup_{\alpha\in K_2}\left[\langle\alpha,\beta\rangle - \frac{1}{2}\|\alpha\|^2_{A(x)} - \frac{1}{2\kappa}\|\alpha\|^2\right] &= \sup_{\alpha\in\mathbb{R}^d}\left[\langle\alpha,\beta\rangle - \frac{1}{2}\|\alpha\|^2_{A(x)} - \frac{1}{2\kappa}\|\alpha\|^2\right]\\
&= \bar{L}_\kappa(x,\beta).
\end{aligned}$$

Also, from (5.28),

$$\begin{aligned}\Lambda^n_{K_2}(x,\beta) &\doteq \sup_{\alpha\in K_2}\left[\langle\alpha,\beta\rangle - \|\alpha\|^2_{A(x)} - \frac{1}{2\kappa}\|\alpha\|^2 - \frac{K_{DA}}{\sqrt{\varkappa(n)n}}\|\alpha\|^3\right]\\ &\le \varkappa(n)nL_c\left(x, \frac{1}{\sqrt{\varkappa(n)n}}\beta\right)\end{aligned}$$

Note that as $n\to\infty$,

$$\sup_{(x,\beta)\in\mathbb{R}^d\times K_1}|\Lambda^n_{K_2}(x,\beta) - \bar{L}_\kappa(x,\beta)| \to 0. \tag{5.29}$$

By Lemma 5.4 and the definitions of γ^n and $\bar{M}^n$, we have

$$\begin{aligned}&\liminf_{n\to\infty}\varkappa(n)nE\left[\sum_{i=0}^{n-1}\frac{1}{n}R(\bar{\mu}^n_i(\cdot)\|\theta(\cdot|\bar{X}^n_i))\right]\\ &\quad\ge \liminf_{n\to\infty}E\left[\int_{\mathbb{R}^d\times\mathbb{R}^d\times[0,1]}\varkappa(n)nL_c\left(x,\frac{1}{\sqrt{\varkappa(n)n}}w\right)\gamma^n(dx\times dw\times dt)\right]\\ &\quad\ge \liminf_{n\to\infty}E\left[\int_{\mathbb{R}^d\times K_1\times[0,1]}\Lambda^n_{K_2}(x,w)\gamma^n(dx\times dw\times dt)\right].\end{aligned}$$

For fixed K_1, since K_2 is bounded, there is $c\in(0,\infty)$ such that $|\Lambda^n_{K_2}(x,w)|\le c(1+\|w\|)$ for all $n\in\mathbb{N}$ and also $|\bar{L}_\kappa(x,w)|\le c(1+\|w\|)$. Using these bounds and (5.18) to control contributions to the integrals from large values of $\|w\|$, it follows from (5.29) that the last quantity in the previous display is the same as

$$\liminf_{n\to\infty}E\left[\int_{\mathbb{R}^d\times K_1\times[0,1]}\bar{L}_\kappa(x,w)\gamma^n(dx\times dw\times dt)\right].$$

Using the continuity of $(x,\beta)\mapsto\bar{L}_\kappa(x,\beta)$ and Fatou's lemma thus gives

$$\begin{aligned}&\liminf_{n\to\infty}\varkappa(n)nE\left[\sum_{i=0}^{n-1}\frac{1}{n}R(\bar{\mu}^n_i(\cdot)\|\theta(\cdot|\bar{X}^n_i))\right]\\ &\quad\ge E\left[\int_{\mathbb{R}^d\times K_1\times[0,1]}\bar{L}_\kappa(x,w)\gamma(dx\times dw\times dt)\right].\end{aligned}$$

Since $\bar{L}_\kappa\ge 0$, by the monotone convergence theorem we can replace K_1 by $\mathbb{R}^d$ in the last display. Next note that as $\kappa\to\infty$,

$$\bar{L}_\kappa(x,\beta)\uparrow\frac{1}{2}\|\beta\|^2_{A^{-1}(x)}$$

for all $(x, \beta) \in \mathbb{R}^{2d}$. Finally, using the monotone convergence theorem, the decomposition (5.27), and Jensen's inequality in that order shows that

$$\begin{aligned}
&\liminf_{n\to\infty} \varkappa(n) n E\left[\sum_{i=0}^{n-1} \frac{1}{n} R(\bar{\mu}_i^n(\cdot) \| \theta(\cdot|\bar{X}_i^n))\right] \\
&\quad \geq \lim_{\kappa\to\infty} E\left[\int_{\mathbb{R}^d\times\mathbb{R}^d\times[0,1]} \bar{L}_\kappa(x, w)\, \gamma(dx \times dw \times dt)\right] \\
&\quad = E\left[\int_{\mathbb{R}^d\times\mathbb{R}^d\times[0,1]} \frac{1}{2} \|w\|^2_{A^{-1}(x)}\, \gamma(dx \times dw \times dt)\right] \\
&\quad = E\left[\int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} \|w\|^2_{A^{-1}(X^0(t))}\, \bar{M}(dw|\, t)\, dt\right] \\
&\quad \geq E\left[\frac{1}{2}\int_0^1 \|w(t)\|^2_{A^{-1}(X^0(t))}\, dt\right],
\end{aligned}$$

which is (5.24). This concludes the proof of Theorem 5.7. □

5.4 Laplace Upper Bound

In this section we prove the variational lower bound

$$\liminf_{n\to\infty} -\varkappa(n) \log E\left[e^{-\frac{1}{\varkappa(n)} F(Y^n)}\right] \geq \inf_{\phi\in\mathscr{C}([0,1]:\mathbb{R}^d)} [I_M(\phi) + F(\phi)], \tag{5.30}$$

which corresponds to the Laplace upper bound.

Suppose for each n that $\{\bar{\mu}_i^n\}$ comes within $1/n$ of achieving the infimum in (5.15), so that

$$\begin{aligned}
&\liminf_{n\to\infty} -\varkappa(n) \log E\left[e^{-\frac{1}{\varkappa(n)} F(Y^n)}\right] \\
&\quad \geq \liminf_{n\to\infty} E\left[\sum_{i=0}^{n-1} \varkappa(n) R(\bar{\mu}_i^n(\cdot) \| \theta(\cdot|\bar{X}_i^n)) + F(\bar{Y}^n)\right].
\end{aligned} \tag{5.31}$$

We also have

$$\sup_{n\in\mathbb{N}} \varkappa(n) n E\left[\sum_{i=0}^{n-1} \frac{1}{n} R(\bar{\mu}_i^n(\cdot) \| \theta(\cdot|\bar{X}_i^n))\right] \leq 2\|F\|_\infty + 1.$$

Consequently, (5.17) is satisfied with $K_E = 2\|F\|_\infty + 1$, and from Theorem 5.7 we can choose for every subsequence of $\{(\bar{M}^n, \bar{Y}^n)\}$ a further subsequence (we retain n as the index for convenience) along which $(\bar{M}^n, \bar{Y}^n)$ converges to $(\bar{M}, \bar{Y})$ in distribution,

$\bar{M}, \bar{Y}$ are related by (5.22)–(5.23), and such that (5.24) is satisfied. Combining this with (5.31) gives

$$\begin{aligned}
&\liminf_{n\to\infty} -\varkappa(n) \log E\left[e^{-\frac{1}{\varkappa(n)}F(Y^n)}\right] \\
&\quad \geq \liminf_{n\to\infty} E\left[\sum_{i=0}^{n-1} \varkappa(n) R(\bar{\mu}_i^n(\cdot) \,\|\, \theta(\cdot|\bar{X}_i^n)) + F(\bar{Y}^n)\right] \\
&\quad \geq E\left[\int_0^1 \frac{1}{2} \|w(s)\|^2_{A^{-1}(X^0(s))}\, ds + F(\bar{Y})\right].
\end{aligned}$$

Define ϕ^u for $u \in \mathscr{L}^2([0,1] : \mathbb{R}^d)$ by

$$\phi^u(t) = \int_0^t Db(X^0(s))\phi^u(s)ds + \int_0^t A^{1/2}(X^0(s))u(s)ds. \tag{5.32}$$

Recalling

$$\bar{Y}(t) = \int_0^t Db(X^0(s))\bar{Y}(s)ds + \int_0^t w(s)ds,$$

it follows using the expression for I_M in (5.10) and (5.11) that

$$\begin{aligned}
&E\left[\int_0^1 \frac{1}{2} \|w(s)\|^2_{A^{-1}(X^0(s))}\, ds + F(\bar{Y})\right] \\
&\quad \geq \inf_{u\in\mathscr{L}^2([0,1]:\mathbb{R}^d)} \left[\int_0^1 \frac{1}{2}\|u(s)\|^2\, ds + F(\phi^u)\right] \\
&\quad = \inf_{\phi\in\mathscr{C}([0,1]:\mathbb{R}^d)} \left[I_M(\phi) + F(\phi)\right],
\end{aligned}$$

which is the lower bound (5.30). □

5.5 Laplace Lower Bound

In this section we prove the variational upper bound

$$\limsup_{n\to\infty} -\varkappa(n) \log E\left[e^{-\frac{1}{\varkappa(n)}F(Y^n)}\right] \leq \inf_{\phi\in\mathscr{C}([0,1]:\mathbb{R}^d)} \left[I_M(\phi) + F(\phi)\right], \tag{5.33}$$

which is the Laplace lower bound. Note that for $u, v \in \mathscr{L}^2([0,1] : \mathbb{R}^d)$,

$$\begin{aligned}
\phi^u(t) - \phi^v(t) = &\int_0^t Db(X^0(s))\left(\phi^u(s) - \phi^v(s)\right) ds \\
&+ \int_0^t A^{1/2}(X^0(s))(u(s) - v(s))ds.
\end{aligned}$$

Thus by Gronwall's inequality,

$$\begin{aligned}
&\sup_{t\in[0,1]} \left\|\phi^u(t) - \phi^v(t)\right\| \\
&\quad \le e^{K_b} \int_0^1 \left\|A^{1/2}(X^0(s))(u(s) - v(s))\right\| ds \\
&\quad \le e^{K_b} \left(\int_0^1 \left\|A^{1/2}(X^0(s))(u(s) - v(s))\right\|^2 ds\right)^{1/2} \\
&\quad \le e^{K_b} K_A^{1/2} \left(\int_0^1 \|u(s) - v(s)\|^2 ds\right)^{1/2}. \qquad (5.34)
\end{aligned}$$

Since $\mathscr{C}([0,1]:\mathbb{R}^d)$ is dense in $\mathscr{L}^2([0,1]:\mathbb{R}^d)$, the proof of the Laplace lower bound is reduced to showing that for an arbitrary $u \in \mathscr{C}([0,1]:\mathbb{R}^d)$,

$$\limsup_{n\to\infty} -\varkappa(n) \log E\left[e^{-\frac{1}{\varkappa(n)} F(Y^n)}\right] \le \frac{1}{2}\int_0^1 \|u(s)\|^2 ds + F\left(\phi^u\right). \qquad (5.35)$$

We fix $u \in \mathscr{C}([0,1]:\mathbb{R}^d)$ for the remainder of the proof.

Remark 5.12 The proof of the lower bound for the moderate deviation problem differs substantially from the corresponding proof of the large deviation problem, especially in regard to the treatment of degenerate noise. For the case of large deviations, this was handled in Chap. 4 using a mollification. In the moderate deviations setting a simpler argument is possible. This is largely due to the form of the rate function, which is the same as that of the small noise diffusion model of Sect. 3.2, but with time-dependent drift $Db(X^0(t))$ and diffusion matrix $A^{1/2}\left(X^0(t)\right)$. As just discussed, with this form one can find a nearly optimal trajectory for the limit variational problem of the form ϕ^u, with u continuous rather than just measurable, which greatly facilitates the construction of nearly optimal controls for the prelimit in the proof of the lower bound. This is not possible for the general model of Chap. 4, since it is not useful to view X^n there as a continuous or nearly continuous mapping on an "exogenous" noise process. In this sense, the moderate deviation problem shares some of the simplifying features of the continuous time models discussed in Sect. 3.2 and at greater length in later chapters.

We now turn to the proof of (5.35) for the given $u \in \mathscr{C}([0,1]:\mathbb{R}^d)$. The main difficulty related to the possible degeneracy of the noise is the following. Since at the prelimit, the controlled processes $\bar{X}^n$ may be close to but not precisely equal to X^0, the range of $A(\bar{X}^n(i/n))$ can differ from that of $A(X^0(i/n))$ (at least in the degenerate case). Because of this, the construction of a control that approximates $A^{1/2}(X^0(i/n))u(i/n)$ with nearly optimal cost is not as straightforward as in the nondegenerate case [it is simple in that case due to the invertibility of $A^{1/2}(\bar{X}^n(i/n))$].

Recall the orthogonal decomposition of $A^{-1}(x)$ discussed above (5.6). For $\kappa \in (0,\infty)$, define

$$A_\kappa^{-1}(x) = Q(x)\Lambda_\kappa^{-1}(x)Q^T(x),$$

where $\Lambda_\kappa^{-1}(x)$ is the diagonal matrix such that $\Lambda_{ii,\kappa}^{-1}(x) = \Lambda_{ii}^{-1}(x)$ when $\Lambda_{ii}^{-1}(x) \le \kappa^2$ and $\Lambda_{ii,\kappa}^{-1}(x) = \kappa^2$ when $\Lambda_{ii}^{-1}(x) > \kappa^2$. Note that by [155, Theorem 6.2.37], $A^{1/2}(x)$, $A_\kappa^{-1}(x)$, and $A_\kappa^{1/2}(x)$ are continuous functions of $A(x)$, and consequently they are also continuous functions of $x \in \mathbb{R}^d$. In addition, define

$$u_\kappa(s) = \begin{cases} u(s) & \text{for } \|u(s)\| \le \kappa, \\ \frac{\kappa u(s)}{\|u(s)\|} & \text{for } \|u(s)\| > \kappa. \end{cases}$$

Let $\phi^{u,\kappa}(t) = \phi^{A_\kappa^{-1/2}(X^0)u_\kappa}(t)$, and note that $\phi^{u,\kappa}$ solves

$$\begin{aligned} \phi^{u,\kappa}(t) = &\int_0^t Db(X^0(s))\phi^{u,\kappa}(s)ds \\ &+ \int_0^t A(X^0(s))A_\kappa^{-1/2}(X^0(s))u_\kappa(s)ds. \end{aligned} \tag{5.36}$$

For n sufficiently large,

$$\max_{0\le i\le n-1} \frac{1}{\sqrt{\varkappa(n)n}} \left\| A_\kappa^{-1/2}\left(X^0(i/n)\right) u_\kappa(i/n) \right\| \le \frac{1}{\sqrt{\varkappa(n)n}}\kappa^2 \le \lambda_{DA},$$

and we can define the sequence $\{(\bar{X}^{n,\kappa}, \bar{Y}^{n,\kappa}, \bar{M}^{n,\kappa}, w^{n,\kappa})\}$ as in Construction 5.3 with

$$\begin{aligned} \bar{\mu}_i^{n,\kappa}(dy) = \exp\Bigg\{ &\left\langle y, \frac{1}{\sqrt{\varkappa(n)n}} A_\kappa^{-1/2}\left(X^0(i/n)\right) u_\kappa(i/n) \right\rangle \\ &- H_c\left(\bar{X}_i^{n,\kappa}, \frac{1}{\sqrt{\varkappa(n)n}} A_\kappa^{-1/2}\left(X^0(i/n)\right) u_\kappa(i/n)\right) \Bigg\} \theta(dy|\bar{X}_i^{n,\kappa}). \end{aligned}$$

We will use

$$\int_{\mathbb{R}^d} y \exp\{\langle y, \alpha\rangle - H_c(x,\alpha)\}\theta(dy|x) = D_\alpha H_c(x,\alpha)$$

and the formula

$$D_\alpha H_c(x,\alpha) = D_\alpha H_c(x,\alpha) - D_\alpha H_c(x,0) = \int_0^1 \left(\frac{d}{ds} D_\alpha H_c(x, s\alpha)\right) ds,$$

where $D_\alpha H_c(x,0) = 0$ follows from (5.4). Using (5.5) to approximate second derivatives that appear on the right side of the last display, the bounds (5.7) imply that for $\|\alpha\| \le \lambda_D A$,

$$\left\| \int_{\mathbb{R}^d} y \exp\{\langle y, \alpha\rangle - H_c(x, \alpha)\}\theta(dy|x) - A(x)\alpha \right\| \le K_{DA} \|\alpha\|^2 . \tag{5.37}$$

The next result identifies the limit in probability of the controlled processes and an asymptotic bound for the relative entropies. Recall that $u \in \mathscr{C}([0, 1] : \mathbb{R}^d)$ has been fixed.

Theorem 5.13 *Let $\kappa \in (0, \infty)$ be given. Consider the controls $\{\bar{\mu}_i^{n,\kappa}\}$ and random variables $\{(\bar{X}^{n,\kappa}, \bar{Y}^{n,\kappa}, \bar{M}^{n,\kappa}, w^{n,\kappa})\}$ as in Construction 5.3 with $\{\bar{\mu}_i^n\}$ replaced by $\{\bar{\mu}_i^{n,\kappa}\}$, and define $\phi^{u,\kappa}$ by (5.36). Then*

$$\bar{Y}^{n,\kappa} \to \phi^{u,\kappa} \tag{5.38}$$

in $\mathscr{C}([0, 1] : \mathbb{R}^d)$ in probability, and

$$\begin{aligned} &\limsup_{n\to\infty} \varkappa(n) n E\left[\frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^{n,\kappa}(\cdot) \,\middle\|\, \theta(\cdot|\bar{X}_i^{n,\kappa})\right)\right] \\ &\qquad \le \frac{1}{2}\int_0^1 \left\| A_\kappa^{-1/2}(X^0(s))u_\kappa(s)\right\|^2_{A(X^0(s))} ds. \end{aligned} \tag{5.39}$$

Proof From (5.8) and (5.37), for all n large enough that $\kappa^2/\sqrt{\varkappa(n)n} \le \lambda_{DA}$ and with $s_i^n \doteq i/n$,

$$\begin{aligned} &R\left(\bar{\mu}_i^{n,\kappa}(\cdot) \,\middle\|\, \theta(\cdot|\bar{X}_i^{n,\kappa})\right) \\ &\quad = \int_{\mathbb{R}^d} \left\langle y, \frac{1}{\sqrt{\varkappa(n)n}} A_\kappa^{-1/2}\left(X^0\left(s_i^n\right)\right) u_\kappa\left(s_i^n\right)\right\rangle \bar{\mu}_i^{n,\kappa}(dy) \\ &\qquad - H_c\left(\bar{X}_i^{n,\kappa}, \frac{1}{\sqrt{\varkappa(n)n}} A_\kappa^{-1/2}\left(X^0\left(s_i^n\right)\right) u_\kappa\left(s_i^n\right)\right) \\ &\quad \le \frac{1}{\varkappa(n)n}\left\langle A\left(\bar{X}_i^{n,\kappa}\right) A_\kappa^{-1/2}\left(X^0\left(s_i^n\right)\right) u_\kappa\left(s_i^n\right), A_\kappa^{-1/2}\left(X^0\left(s_i^n\right)\right) u_\kappa\left(s_i^n\right)\right\rangle \\ &\qquad - \frac{1}{2\varkappa(n)n}\left\langle A\left(\bar{X}_i^{n,\kappa}\right) A_\kappa^{-1/2}\left(X^0\left(s_i^n\right)\right) u_\kappa\left(s_i^n\right), \right. \\ &\qquad\qquad \left. A_\kappa^{-1/2}\left(X^0\left(s_i^n\right)\right) u_\kappa\left(s_i^n\right)\right\rangle + \frac{2}{(\varkappa(n)n)^{3/2}} K_{DA}\kappa^6 \\ &\quad = \frac{1}{2\varkappa(n)n}\left\| A_\kappa^{-1/2}\left(X^0\left(s_i^n\right)\right) u_\kappa\left(s_i^n\right)\right\|^2_{A(\bar{X}_i^{n,\kappa})} + \frac{2}{(\varkappa(n)n)^{3/2}} K_{DA}\kappa^6. \end{aligned}$$

Consequently,

$$\limsup_{n\to\infty} \varkappa(n) n E\left[\frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^{n,\kappa}(\cdot) \,\middle\|\, \theta(\cdot|\bar{X}_i^{n,\kappa})\right)\right] \tag{5.40}$$

$$\leq \limsup_{n\to\infty} \frac{1}{2} E\left[\frac{1}{n}\sum_{i=0}^{n-1} \left\| A_\kappa^{-1/2}\left(X^0(i/n)\right) u_\kappa(i/n)\right\|^2_{A(\bar{X}_i^{n,\kappa})}\right],$$

where in fact,

$$\limsup_{n\to\infty} \frac{1}{2} E\left[\frac{1}{n}\sum_{i=0}^{n-1} \left\| A_\kappa^{-1/2}\left(X^0(i/n)\right) u_\kappa(i/n)\right\|^2_{A(\bar{X}_i^{n,\kappa})}\right] \leq \frac{1}{2}\kappa^4 K_A.$$

Therefore, (5.17) is satisfied by $\{\bar{\mu}_j^{n,\kappa}\}$. Thus the conclusions of Theorem 5.7 hold with $\bar{Y}^n$, $\bar{M}^n$ replaced by $\bar{Y}^{n,\kappa}$, $\bar{M}^{n,\kappa}$. Choose a subsequence (keeping n as the index for convenience) along which $\{(\bar{M}^{n,\kappa}, \bar{Y}^{n,\kappa})\}$ converges weakly to some limit $(\bar{M}^\kappa, \bar{Y}^\kappa)$, where $[\bar{M}^\kappa]_2$ is Lebesgue measure and

$$\bar{Y}^\kappa(t) = \int_0^t Db(X^0(s))\bar{Y}^\kappa(s)ds + \int_0^t \int_{\mathbb{R}^d} w[\bar{M}^\kappa]_{1|2}(dw\,|s)ds.$$

Then $\bar{Y}^{n,\kappa} \to \bar{Y}^\kappa$ implies

$$\sup_{t\in[0,1]} \left\|\bar{X}^{n,\kappa}(t) - X^0(t)\right\| \to 0$$

in probability. Because of this and the continuity of $A^{1/2}(x)$, we have (recall $s^n(t) \doteq \lfloor nt \rfloor / n$)

$$\sup_{t\in[0,1]} \left\| A^{1/2}(\bar{X}^{n,\kappa}(s^n(t))) - A^{1/2}(X^0(s^n(t)))\right\| \to 0$$

in probability. However, the continuity of $t \mapsto A^{1/2}(X^0(t))A_\kappa^{-1/2}(X^0(t))u_\kappa(t)$ gives

$$\sup_{t\in[0,1]} \left\| A^{1/2}(X^0(s^n(t)))A_\kappa^{-1/2}(X^0(s^n(t)))u_\kappa(s^n(t)) - A^{1/2}(X^0(t))A_\kappa^{-1/2}(X^0(t))u_\kappa(t)\right\| \to 0.$$

Combining these limits, and using the fact that $A_\kappa^{-1/2}(X^0(t))u_\kappa(t)$ is uniformly bounded, shows that

$$\sup_{t\in[0,1]} \left\| A^{1/2}(\bar{X}^{n,\kappa}(s^n(t)))A_\kappa^{-1/2}(X^0(s^n(t)))u_\kappa(s^n(t)) - A^{1/2}(X^0(t))A_\kappa^{-1/2}(X^0(t))u_\kappa(t)\right\| \to 0 \tag{5.41}$$

in probability. This combined with the uniform bounds allows the use of the dominated convergence theorem to show that

$$\limsup_{n\to\infty} E\left[\frac{1}{2}\int_0^1 \left\|A_\kappa^{-1/2}(X^0(s^n(t)))u_\kappa(s^n(t))\right\|^2_{A(\bar{X}^{n,\kappa}(s^n(t)))}\,dt\right]$$
$$= \frac{1}{2}\int_0^1 \left\|A_\kappa^{-1/2}(X^0(t))u_\kappa(t)\right\|^2_{A(X^0(t))}\,dt.$$

Combining this with (5.40) demonstrates (5.39).

To prove (5.38), we will show that in fact,

$$\bar{M}^\kappa(dw\times dt) = \delta_{A(X^0(t))A_\kappa^{-1/2}(X^0(t))u_\kappa(t)}(dw)dt.$$

For all $\sigma > 0$, let

$$G_\sigma \doteq \left\{(z,t)\in\mathbb{R}^d\times[0,1] : \left\|z - A(X^0(t))A_\kappa^{-1/2}(X^0(t))u_\kappa(t)\right\| \le \sigma\right\},$$

and note that by weak convergence, $\limsup_{n\to\infty} E[\bar{M}^{n,\kappa}(G_\sigma)] \le E[\bar{M}^\kappa(G_\sigma)]$. Note also that

$$E[\bar{M}^{n,\kappa}(G_\sigma)]$$
$$\ge P\left[\sup_{t\in[0,1]}\left\|\sqrt{\varkappa(n)n}\int_{\mathbb{R}^d} y\bar{\mu}^{n,\kappa}_{\lfloor nt\rfloor}(dy) - A(X^0(t))A_\kappa^{-1/2}(X^0(t))u_\kappa(t)\right\| \le \sigma\right].$$

However, by (5.37) we can choose n large enough to make

$$\sup_{t\in[0,1]}\left\|\sqrt{\varkappa(n)n}\int_{\mathbb{R}^d} y\bar{\mu}^{n,\kappa}_{\lfloor nt\rfloor}(dy) - A\left(\bar{X}^{n,\kappa}\left(s^n(t)\right)\right)A_\kappa^{-1/2}\left(X^0\left(s^n(t)\right)\right)u_\kappa\left(s^n(t)\right)\right\|$$

arbitrarily small, and the proof that

$$\sup_{t\in[0,1]}\left\|A(\bar{X}^{n,\kappa}(s^n(t)))A_\kappa^{-1/2}(X^0(s^n(t)))u_\kappa(s^n(t))\right.$$
$$\left. - A(X^0(t))A_\kappa^{-1/2}(X^0(t))u_\kappa(t)\right\| \to 0$$

in probability is identical to the proof of (5.41). Hence $\limsup_{n\to\infty} E[\bar{M}^{n,\kappa}(G_\sigma)] = 1$ for all $\sigma > 0$, and so $E[\bar{M}^\kappa(\cap_{n\in\mathbb{N}}G_{1/n})] = 1$. This implies that with probability 1,

$$[\bar{M}]^\kappa_{1|2}(dw\,|\,t) = \delta_{A(X^0(t))A_\kappa^{-1/2}(X^0(t))u_\kappa}(dw)$$

for a.e. t. It follows that

$$\bar{Y}^\kappa(t) = \int_0^t Db(X^0(s))\bar{Y}^\kappa(s)ds + \int_0^t A(X^0(s))A_\kappa^{-1/2}(X^0(s))u_\kappa(s)ds,$$

and therefore $\bar{Y}^{n,\kappa} \to \phi^{u,\kappa}$ weakly. This implies (5.38) and completes the proof. □

The second theorem in this section allows us to approximate $F(\phi^u)$ by $F(\phi^{u,\kappa})$ and $\frac{1}{2}\int_0^1 \|u(s)\|^2\,ds$ by

$$\frac{1}{2}\int_0^1 \left\|A_\kappa^{-1/2}(X^0(s))u_\kappa(s)\right\|^2_{A(X^0(s))}\,ds.$$

Recall that $u \in \mathscr{C}([0,1]:\mathbb{R}^d)$ has been given.

Theorem 5.14 *Define ϕ^u by (5.32) and $\phi^{u,\kappa}$ by (5.36). Then $\phi^{u,\kappa} \to \phi^u$ in $\mathscr{C}([0,1]:\mathbb{R}^d)$ and*

$$\limsup_{\kappa\to\infty} \frac{1}{2}\int_0^1 \left\|A_\kappa^{-1/2}(X^0(s))u_\kappa(s)\right\|^2_{A(X^0(s))}\,ds \le \frac{1}{2}\int_0^1 \|u(s)\|^2\,ds.$$

Proof Note that

$$\left\|A^{1/2}(X^0(s))A_\kappa^{-1/2}(X^0(s))u_\kappa(s)\right\| \le \|u(s)\|$$

for all $s \in [0,1]$. Thus it is automatic that

$$\limsup_{\kappa\to\infty} \frac{1}{2}\int_0^1 \left\|A_\kappa^{-1/2}(X^0(s))u_\kappa(s)\right\|^2_{A(X^0(s))}\,ds \le \frac{1}{2}\int_0^1 \|u(s)\|^2\,ds.$$

Also, for $x \in \mathbb{R}^d$,

$$A(x)A_\kappa^{-1/2}(x) = Q(x)\Lambda(x)\Lambda_\kappa^{-1/2}(x)Q^T(x) \to Q(x)\Lambda^{1/2}(x)Q^T(x) = A^{1/2}(x).$$

Since $u_\kappa(s) \to u(s)$ for all $s \in [0,1]$,

$$A(X^0(s))A_\kappa^{-1/2}(X^0(s))u_\kappa(s) \to A^{1/2}(X^0(s))u(s) \tag{5.42}$$

pointwise. Since $u \in \mathscr{L}^2([0,1]:\mathbb{R}^d)$, by the dominated convergence theorem, (5.42) also holds in $\mathscr{L}^2([0,1]:\mathbb{R}^d)$. Combining this with the second inequality in (5.34) shows that $\phi^{u,\kappa} \to \phi^u$ in $\mathscr{C}([0,1]:\mathbb{R}^d)$. □

Using (5.15) and the fact that any given control is suboptimal yields

$$-\varkappa(n)\log E\left[e^{-\frac{1}{\varkappa(n)}F(Y^n)}\right] \le E\left[\sum_{i=0}^{n-1}\varkappa(n)R\left(\bar{\mu}_i^{n,\kappa}(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^{n,\kappa})\right) + F(\bar{Y}^{n,\kappa})\right].$$

Using Theorem 5.13, this implies

$$\limsup_{n\to\infty} -\varkappa(n)\log E\left[e^{-\frac{1}{\varkappa(n)}F(Y^n)}\right] \le \frac{1}{2}\int_0^1 \left\|A_\kappa^{-1/2}(X^0(s))u_\kappa(s)\right\|^2_{A(X^0(s))}\,ds + F(\phi^{u,\kappa}).$$

Sending $\kappa \to \infty$ and using Theorem 5.14 gives (5.35), which completes the proof of (5.33). □

5.6 Notes

Among the earliest papers to study moderate deviations are those by Rubin and Sethuraman [222], Ghosh [146], and Michel [200]. See the introduction of [41] for a more complete discussion of work in this area. While a number of settings have been considered, to the authors' knowledge the first papers to consider moderate deviations for small noise processes around solutions to general nonlinear ODEs (rather than constant-velocity trajectories) are [100] for discrete time models, upon which this chapter is based, and [41] for continuous time processes. As noted in the introduction to the chapter, the proof of the moderate deviation principle presented here is neither uniformly harder nor easier than its large deviation counterpart, at least when one is using weak convergence methods. In particular, the large deviation upper bound is made more difficult due to difficulties in using tightness in the convergence analysis. (The case of solutions to SDEs driven by Brownian motion, which is given as an example in Chap. 3, is in fact much easier, owing to the fact that the driving noise is already Gaussian.) Also, the assumed conditions are not strictly weaker, mainly in that additional smoothness is needed for the proper centering and rescaling. Moderate deviation principles will also appear in Chaps. 10 and 13.

Chapter 6
Empirical Measure of a Markov Chain

In this chapter we develop the large deviation theory for the empirical measure of a Markov chain, thus generalizing Sanov's theorem from Chap. 3. The ideas developed here are useful in other contexts, such as proving sample path large deviation properties of processes with multiple time scales as described in Sect. 7.3.

To focus on the main issues, we first consider Markov chains with a compact Polish state space S. Proving the large deviation upper bound in the case of a general Polish space requires the existence of a suitable Lyapunov function, which is discussed in Sect. 6.10. The Lyapunov function is used to prove tightness of a controlled empirical measure when relative entropy costs are bounded, a tightness that is automatic in the compact case. For examples in which such Lyapunov functions exist for noncompact state space models see [87, 88, 97].

Thus let $\{X_i, i \in \mathbb{N}_0\}$ denote a Markov chain with transition kernel $p(x, dy)$ and compact state space S. The object of interest is the empirical or occupation measure defined by

$$L^n(dx) \doteq \frac{1}{n} \sum_{i=0}^{n-1} \delta_{X_i}(dx). \tag{6.1}$$

Under ergodicity there is a unique invariant measure $\pi \in \mathscr{P}(S)$, and by the ergodic theorem, $L^n \to \pi$ in the weak topology, w.p.1. Large deviation theory gives approximations to the probability of $\{L^n \in A\}$ when A does not contain π, along with related expected values.

6.1 Applications

While there are many applications of large deviation estimates for the empirical measure to problems in the physical sciences and engineering, we mention here uses in other areas.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_6

6.1.1 Markov Chain Monte Carlo

One of the most important uses of the empirical measure is for the numerical approximation of integrals of the form $\int_S f(x)\pi(dx)$, and in particular in the special case in which π is a Gibbs measure, e.g., $\pi(dx) = e^{-V(x)/\tau}dx/Z$, where $S \subset \mathbb{R}^d$, V is a potential function, τ is a parameter, and Z is a normalization that makes the indicated measure a probability measure. There are well-known methods to construct ergodic Markov processes $\{X_i, i \in \mathbb{N}_0\}$ for which π is the unique invariant distribution, and thus $\int_S f(x)L^n(dx)$ gives a convergent approximation to $\int_S f(x)\pi(dx)$. This technique has a tremendous number of practical applications in the physical and biological sciences, engineering, statistics, and elsewhere [6, 190].

However, for many problems, S is in some sense very large, and moreover the methods that generate the chain from V have the property that when V has many deep local minima, parts of the state space communicate poorly under the dynamics $p(x, dy)$. When this happens, and it happens frequently, the issue of good algorithm design becomes crucial.

In order to compare algorithms, one needs a criterion for good performance. Since it focuses on the object of interest, i.e., the empirical measure, it would seem that the large deviation rate is a natural measure. The rate function I depends of course on the dynamics, though for any chain leading to π as an invariant distribution, $I(\mu) = 0$ if and only $\mu = \pi$. Different algorithms lead to different rate functions, and the rate functions give one a great deal of information that can be used to compare algorithms.

The rate function can be compared with other measures that have been traditionally used to compare chains, such as the subdominant eigenvalue. Let $p(x, dy)$ be an ergodic transition kernel with invariant distribution π. Under suitable conditions, $p(x, \cdot)$ has a single eigenvalue equal to 1 corresponding to the eigenvector π, and the magnitude $|\lambda_2|$ of the next-largest eigenvalue is often used to characterize the performance of the associated empirical measure. However, the second eigenvalue provides information only on convergence of the n-step transition kernel $p^{(n)}(x, dy) = P\{X_n \in dy | X_0 = x\}$, and in particular does not give any direct information regarding the empirical measure.

A work that effectively applies the large deviation rate as a measure of rate of convergence is [108], and further development of its use in algorithm design is ongoing [86].

6.1.2 Markov Modulated Dynamics

The process models considered in Chap. 4 can be made more general and appropriate for a broader range of applications by allowing the distribution of the driving noise to depend on an exogenous Markov chain. For example, the homogeneous random walk model [i.e., $\theta(dy|x) = \theta(dy)$] occurs in problems of insurance risk, with the noises v_i representing the difference between income and payouts at time i. A more

realistic model would allow the distribution $\theta(dy)$ to depend on a finite state Markov chain ξ_i representing, e.g., the state of the economy and other factors. Such a process would be called *Markov modulated*.

Similarly, in the more general case with state dependence one could replace $\theta(dy|x)$ by $\theta(dy|x,\xi)$ with $\xi \in S$. Suppose $\{\xi_i\}$ is a Markov chain on S that is independent of $\{v_i(x,\xi), i \in \mathbb{N}_0, x \in \mathbb{R}^d, \xi \in S\}$, and replace $v_i(X_i^n)$ as in Chap. 4 by $v_i(X_i^n, \xi_i)$. In such a situation, owing to time scale separation (i.e., X_i^n varies more slowly than ξ_i when n is large), the large deviation properties of the empirical measures of the sequences $\{v_i(x,\xi_i), i \in \mathbb{N}_0\}$ for various $x \in \mathbb{R}^d$ are needed to define the local rate function for the continuous time interpolation $\{X_i^n\}$. These empirical measures can be analyzed using the same methods we use in this chapter to study the empirical measure of just the $\{\xi_i\}$. A statement of the form of the resulting rate function on path space is given in Sect. 7.3.

6.2 The Representation

Throughout this chapter, S is a Polish space. Although in the beginning of the chapter we focus on the case in which S is compact, several results needed later on such as the representation are stated and proved for the general case. Whenever compactness of S is assumed, this will be explicitly noted.

We first introduce some needed notation. Just as in the proof of the representation used for Sanov's theorem (Proposition 3.1), the representation follows by applying the chain rule to the "high-level" representation stated in Proposition 2.2. In contrast with the setting of Sanov's theorem, where the base measure was product measure, here the base measure is the Markov measure

$$\theta(dx_1 \times \cdots \times dx_n) = p(x_0, dx_1)p(x_1, dx_2) \times \cdots \times p(x_{n-1}, dx_n)$$

on S^n, where x_0 is some fixed initial condition. For the high-level representation we consider more or less arbitrary alternative measures μ on S^n. Just as in the proof of Proposition 3.1, we factor μ appearing in $R(\mu \| \theta)$ by conditioning, and then apply the chain rule to decompose the relative entropy on product space as a sum of relative entropies. This gives the following proposition, in which $\bar{X}_0 = x_0$ by definition. Except for the form of the base measure θ, the proof is the same as the proof of Proposition 3.1 and hence omitted.

Proposition 6.1 *Let $n \in \mathbb{N}$ and let $\{X_i\}_{i\in\{1,\ldots,n\}}$ be S-valued random variables with joint distribution θ. If $G \in \mathscr{M}_b(S^n)$, then*

$$-\frac{1}{n}\log E e^{-nG(X_1,\ldots,X_n)} = \inf E\left[G(\bar{X}_1^n,\ldots,\bar{X}_n^n) + \frac{1}{n}\sum_{i=1}^{n} R\left(\bar{\mu}_i^n(\cdot)\,\|\,p(\bar{X}_{i-1}^n,\cdot)\right)\right],$$

with the infimum over all collections of random probability measures $\{\bar{\mu}_i^n\}_{i\in\{1,\dots,n\}}$ *that satisfy the following two conditions:*

1. $\bar{\mu}_i^n$ *is measurable with respect to the* σ*-algebra generated by* $\mathscr{F}_{i-1}^n$*, where* $\mathscr{F}_0^n = \{\emptyset, \Omega\}$ *and for* $i \in \{1,\dots,n\}$, $\mathscr{F}_i^n = \sigma\{\bar{X}_1^n,\dots,\bar{X}_i^n\}$.
2. *The conditional distribution of* $\bar{X}_i^n$*, given* $\mathscr{F}_{i-1}^n$*, is* $\bar{\mu}_i^n$.

Later we will consider uniformity of the large deviation estimates with respect to x_0 in some compact subset of S. When dealing with this issue, we denote the initial condition explicitly by writing E_{x_0}. As before, we consider $\{\bar{X}_j\}_{i\in\{1,\dots,n\}}$ to be a *controlled* version of the original sequence $\{X_i\}_{i\in\{1,\dots,n\}}$, with the control $\bar{\mu}_j^n$ selecting the (conditional) distribution of $\bar{X}_j$. Let $\bar{L}^n$ be the controlled empirical measure (with $\bar{X}_0^n = x_0$):

$$\bar{L}^n(dx) = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\bar{X}_i^n}(dx). \tag{6.2}$$

Then by Proposition 6.1,

$$-\frac{1}{n}\log E e^{-nF(L^n)} = \inf_{\{\bar{\mu}_i^n\}} E\left[F(\bar{L}^n) + \frac{1}{n}\sum_{i=1}^{n} R\left(\bar{\mu}_i^n(\cdot)\,\|\,p(\bar{X}_{i-1}^n,\cdot)\right)\right] \tag{6.3}$$

for every bounded and measurable $F : \mathscr{P}(S) \to \mathbb{R}$. To prove an LDP for $\{L^n\}_{n\in\mathbb{N}}$, it will be enough to consider bounded and continuous F.

6.3 Form of the Rate Function

In the setting of Sanov's theorem, the minimizing controls were found, a posteriori, to be asymptotically product measure (see Remark 3.7), reflecting the form of the base measure on the collection $\{X_i\}_{i\in\mathbb{N}}$. One might suspect something analogous here, which is that nearly optimizing controls for large n might be of the Markov form $\bar{\mu}_i^n(dx_i) = q(\bar{X}_{i-1}^n, dx_i)$ for some transition kernel q. With this in mind, we rewrite the relative entropy using the chain rule [Theorem 2.6]:

$$\begin{aligned} R\left(\bar{\mu}_i^n(\cdot)\,\|\,p(\bar{X}_{i-1}^n,\cdot)\right) &= R\left(\bar{\mu}_i^n(\cdot)\,\|\,p(\bar{X}_{i-1}^n,\cdot)\right) + R\left(\delta_{\bar{X}_{i-1}^n}(\cdot)\,\middle\|\,\delta_{\bar{X}_{i-1}^n}(\cdot)\right) \\ &= R\left(\delta_{\bar{X}_{i-1}^n}(dx)\bar{\mu}_i^n(dy)\,\middle\|\,\delta_{\bar{X}_{i-1}^n}(dx)p(x,dy)\right). \end{aligned}$$

The measure $\delta_{\bar{X}_{i-1}^n}(dx)\bar{\mu}_i^n(dy)$ records the control that is used to pick the distribution of $\bar{X}_i^n$ given the location of $\bar{X}_{i-1}^n$, and it will be used to identify the form of $q(x,dy)$.

We will try to guess the form of the rate function, and at the same time sketch the proof of the large deviation upper bound without giving details; precise statements

and proofs will be given in later sections. Suppose that the controls $\{\bar{\mu}_i^n\}$ come within $1/n$ of the infimum. By Jensen's inequality and the joint convexity of relative entropy, we have

$$
\begin{aligned}
&E\left[F(\bar{L}^n)+\frac{1}{n}\sum_{i=1}^{n} R\left(\delta_{\bar{X}_{i-1}^n}(dx)\bar{\mu}_i^n(dy)\,\middle\|\,\delta_{\bar{X}_{i-1}^n}(dx)p(x,dy)\right)\right] \qquad (6.4)\\
&\geq E\left[F(\bar{L}^n)+R\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\bar{X}_{i-1}^n}(dx)\bar{\mu}_i^n(dy)\,\middle\|\,\frac{1}{n}\sum_{i=1}^{n}\delta_{\bar{X}_{i-1}^n}(dx)p(x,dy)\right)\right]\\
&= E\left[F(\bar{L}^n)+R\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\bar{X}_{i-1}^n}(dx)\bar{\mu}_i^n(dy)\,\middle\|\,\bar{L}^n(dx)p(x,dy)\right)\right].
\end{aligned}
$$

Let

$$
\lambda^n(dx\times dy)\doteq\frac{1}{n}\sum_{i=1}^{n}\delta_{\bar{X}_{i-1}^n}(dx)\bar{\mu}_i^n(dy). \qquad (6.5)
$$

Since S and hence S^2 are compact, so are $\mathscr{P}(S)$ and $\mathscr{P}(S^2)$, and so automatically $\left\{\left(\lambda^n,\bar{L}^n\right),n\in\mathbb{N}\right\}$ is tight. Note that $\bar{L}^n$ is the first marginal of λ^n, and that $\bar{\mu}_i^n(dy)$ picks the distribution of $\bar{X}_i^n$. Hence the martingale generalization of the Glivenko–Cantelli lemma as formulated in Lemma 3.5 can be used to show that asymptotically, the first and second marginals of λ^n are the same. The precise result will be given in Lemma 6.12.

Thus if $\left(\lambda^n,\bar{L}^n\right)\to\left(\lambda,\bar{L}\right)$ in distribution along a subsequence, then $[\lambda]_1(dx)=[\lambda]_2(dx)=\bar{L}(dx)$, where $[\lambda]_1$ and $[\lambda]_2$ denote the first and second marginals of λ. We will assume that $p(x,dy)$ satisfies the **Feller property**, i.e., that the mapping $x\to p(x,\cdot)$ is continuous in the topology of weak convergence. This will imply $\bar{L}^n(dx)p(x,dy)\to\bar{L}(dx)p(x,dy)$ in distribution. We can then compute a lower bound along the weakly converging subsequence using Fatou's lemma, the continuity of F, and lower semicontinuity of $R(\cdot\|\cdot)$:

$$
\begin{aligned}
&\liminf_{n\to\infty}-\frac{1}{n}\log Ee^{-nF(L^n)}\\
&\quad\geq\liminf_{n\to\infty}E\left[F(\bar{L}^n)+R\left(\lambda^n(dx\times dy)\,\middle\|\,\bar{L}^n(dx)p(x,dy)\right)\right]\\
&\quad\geq E\left[F(\bar{L})+R\left(\lambda(dx\times dy)\|\,\bar{L}(dx)p(x,dy)\right)\right].
\end{aligned}
$$

Given $\mu\in\mathscr{P}(S)$ and a probability transition kernel p on S, let $(\mu\otimes p)(dx\,dy)$ denote the probability measure on S^2 given by $\mu(dx)p(x,dy)$, and let

$$
A(\mu)\doteq\left\{\gamma\in\mathscr{P}(S^2):[\gamma]_1=[\gamma]_2=\mu\right\}. \qquad (6.6)
$$

Suppose we define

$$
I(\mu)=\inf_{\gamma\in A(\mu)}R(\gamma\|\,\mu\otimes p).
$$

Then since $[\lambda]_1 = [\lambda]_2 = \bar{L}$, we have shown that

$$\liminf_{n\to\infty} -\frac{1}{n} \log E e^{-nF(L^n)} \geq \inf_{\mu\in\mathscr{P}(S)} [F(\mu) + I(\mu)] ,$$

suggesting that I may in fact be the rate function.

To complete the proof we will have to prove the reverse inequality with the same function I. Note that if $\gamma \in A(\mu)$, then we can factor $\gamma(dx \times dy)$ in the form $\gamma(dx \times dy) = \mu(dx)q(x, dy)$ for some transition kernel q, and that $[\gamma]_2 = \mu$ is exactly the statement that μ is an invariant distribution for q. This will suggest how to construct a control for the reverse inequality, and also verify the claim that asymptotically, there are nearly optimal measures that are Markovian and stationary.

6.4 Assumptions and Statement of the LDP

For the purposes of proving a Laplace principle upper bound and identifying the rate function, we will assume that $p(x, dy)$ satisfies the Feller property: if $\{x_n, n \in \mathbb{N}\}$ is any sequence in S such that $x_n \to x \in S$, then $p(x_n, \cdot) \Rightarrow p(x, \cdot)$.

Condition 6.2 *The transition probability kernel p satisfies the Feller property.*

Under the Feller property, $\mu_n \Rightarrow \mu$ implies $\mu_n \otimes p \Rightarrow \mu \otimes p$ [Lemma 6.7], which greatly simplifies the analysis and the form of the rate function. If the Feller property does not hold, then sets that are negligible under a stationary distribution of p may be significant from the large deviation perspective [118].

Given a transition probability function $q(x, dy)$ on S and $k \in \mathbb{N}$, let $q^{(1)}(x, dy) \equiv q(x, dy)$ and let $q^{(k)}(x, dy)$ denote the k-step transition probability function defined recursively by

$$q^{(k+1)}(x, A) = \int_S q(y, A)\, q^{(k)}(x, dy)$$

for Borel sets A. The following transitivity assumption, weaker than Hypothesis H in [88], is a slight variation of Hypothesis (gH) in [59].

Condition 6.3 *The transition kernel p satisfies the following transitivity condition. There exist positive integers l_0 and n_0 such that for all x and ζ in S,*

$$\sum_{i=l_0}^{\infty} \frac{1}{2^i} p^{(i)}(x, dy) \ll \sum_{j=n_0}^{\infty} \frac{1}{2^j} p^{(j)}(\zeta, dy) , \tag{6.7}$$

where $p^{(k)}$ denotes the k-step transition probability.

Thus if one allows a sufficient delay, the weighted distributions of the chain at future times are in some sense mutually absolutely continuous for all pairs of starting points.

The following condition and the Feller property (Condition 6.2) will imply the existence of an invariant probability distribution for $p(x, dy)$.

Condition 6.4 *The collection* $\{EL^n\}_{n\in\mathbb{N}}$ *is tight in* $\mathscr{P}(S)$.

For the proof of the large deviation upper bound we will need a stronger stability property than Condition 6.4. Such a property can be formulated in terms of a suitable Lyapunov function (see Sect. 6.10). However, to keep the presentation simple, we first consider the case in which Condition 6.4 holds automatically, which occurs when S is compact.

Remark 6.5 If p corresponds to a finite state ergodic Markov chain, then Conditions 6.2, 6.3, and 6.4 all hold.

Theorem 6.6 *Assume Conditions 6.2 and 6.3 on the Markov chain* $\{X_i\}_{i\in\mathbb{N}_0}$. *Furthermore, suppose that S is compact. Let* $\{L^n\}_{n\in\mathbb{N}}$ *be the empirical measures defined in* (6.1). *Then* $\{L^n\}_{n\in\mathbb{N}}$ *satisfies an LDP with rate function*

$$I(\mu) = \inf_{\gamma\in A(\mu)} R(\gamma \| \mu \otimes p). \tag{6.8}$$

Let $d(\cdot,\cdot)$ *denote the metric on S. Suppose in addition that for each* $x \in S$, *there are* $\beta \in \mathscr{P}(S)$, $k \in \mathbb{N}$, *and* $c > 0$ *such that for all* $\zeta \in S$ *satisfying* $d(\zeta, x) < c$ *and all* $A \in \mathscr{B}(S)$,

$$\sum_{j=1}^{k} p^{(j)}(\zeta, A) \geq c\beta(A). \tag{6.9}$$

Then the large deviation estimates are uniform in the initial condition x_0, *and in particular,* $\{L^n\}_{n\in\mathbb{N}}$ *satisfies a uniform Laplace principle in the sense of Definition 1.11 with the rate function* $I_{x_0} = I$.

The form of the rate function given in (6.8) [see also Remark 6.9] differs from the standard Donsker–Varadhan rate function given in [87, 88]. The rate function in (6.8) can be interpreted in terms of the minimal cost for ergodic control (or average cost per unit time) problems, which is not at all surprising, given the representation (6.3). The proof of Theorem 6.6 is split over the four sections that follow. In particular, the Laplace upper bound is proved in Proposition 6.13, the Laplace lower bound in Proposition 6.15, and the uniform Laplace principle in Proposition 6.18. Finally, in Sect. 6.10 we discuss the case in which S is not compact.

6.5 Properties of the Rate Function

We begin with a result used in the arguments of Sect. 6.3.

Lemma 6.7 *Assume that Condition 6.2 holds and that $\mu_n \in \mathscr{P}(S)$ converges weakly to μ. Then $\mu_n \otimes p$ converges weakly to $\mu \otimes p$ in $\mathscr{P}(S^2)$.*

Proof Let $f \in \mathscr{C}_b(S^2)$ be given. Then the mapping $G : (x, \gamma) \to \int_S f(x, y)\gamma(dy)$ from $S \times \mathscr{P}(S)$ to $\mathbb{R}$ is bounded and continuous [this is easy to see, for example, using the Skorokhod epresentation theorem (Theorem A.8)]. Since $x \mapsto p(x, \cdot)$ is continuous in the weak topology, the mapping $x \mapsto G(x, p(x, \cdot)) = \int_S f(x, y)p(x, dy)$ is bounded and continuous. Since $\mu_n \Rightarrow \mu$,

$$\begin{aligned}\int_{S^2} f(x, y)(\mu_n \otimes p)(dx, dy) &= \int_S \int_S f(x, y)p(x, dy)\mu_n(dx) \\ &\to \int_S \int_S f(x, y)p(x, dy)\mu(dx) \\ &= \int_{S^2} f(x, y)(\mu \otimes p)(dx, dy),\end{aligned}$$

proving the claim. □

Recall the sets $A(\mu)$, $\mu \in \mathscr{P}(S)$, introduced in (6.6). A distribution $\mu \in \mathscr{P}(S)$ is said to be invariant (or an invariant measure or stationary distribution) for a transition kernel $q(x, dy)$ if $\mu(A) = \int_S q(x, A)\mu(dx)$ for all $A \in \mathscr{B}(S)$.

Lemma 6.8 (a) *If $\gamma \in A(\mu)$, then there exists a transition kernel $q(x, dy)$ such that $\gamma(dx \times dy) = \mu(dx)q(x, dy)$, and μ is an invariant distribution for q.*

(b) *For every $\mu \in \mathscr{P}(S)$ with $I(\mu) < \infty$ there is $\gamma \in A(\mu)$ that achieves the infimum in the definition of $I(\mu)$, i.e.,*

$$I(\mu) = R(\gamma \| \mu \otimes p).$$

Proof If $\gamma \in A(\mu)$, then the existence of regular conditional probabilities (Theorem B.2) guarantees the existence of $q(x, dy)$ with the properties of a probability transition function and such that $\gamma(dx\, dy) = \mu(dx)q(x, dy)$. Since $\gamma \in A(\mu)$ implies $[\gamma]_2 = \mu$, it follows that $\int_S \mu(dx)q(x, A) = \mu(A)$ for Borel sets A, and hence μ is invariant under q. This proves part (a).

Since $I(\mu) < \infty$, given $n \in \mathbb{N}$ there is $\gamma_n \in A(\mu)$ such that $R(\gamma_n \| \mu \otimes p) \le I(\mu) + 1/n$. Since for $i = 1, 2$, $[\gamma_n]_i = \mu$, it follows that $\{[\gamma_n]_i\}_{n\in\mathbb{N}}$ is tight, which implies in turn that $\{\gamma_n\}_{n\in\mathbb{N}}$ is tight. So γ_n will converge (at least along a subsequence) to a limit $\gamma \in A(\mu)$, and the lower semicontinuity of $R(\cdot \| \mu \otimes p)$ implies $R(\gamma \| \mu \otimes p) \le I(\mu)$. However, $\gamma \in A(\mu)$ implies $I(\mu) \le R(\gamma \| \mu \otimes p)$, which completes the proof for part (b). □

Remark 6.9 If γ achieves the infimum in the definition of $I(\mu)$ and $q(x, dy)$ is the factorization as in the first part of the lemma, then the chain rule implies

$$I(\mu) = \int_S R\left(q(x,\cdot) \| \, p(x,\cdot)\right) \mu(dx).$$

Thus one may interpret the rate function as follows. There is a relative entropy cost to perturb the dynamics from the underlying Markov chain p to q. We consider all such perturbations that admit μ as a stationary distribution, pay the relative entropy cost based on the frequency with which states are visited, and then minimize over q.

Lemma 6.10 *Define $I : \mathscr{P}(S) \to [0, \infty]$ by (6.8). Then the following hold.*
(a) *I is convex.*
(b) *$I(\mu) = 0$ if and only if μ is an invariant probability distribution for p.*
(c) *If Condition 6.2 holds, then I is lower semicontinuous.*
(d) *If in addition S is compact, then the level set $\{\mu : I(\mu) \le M\}$ is compact for each $M < \infty$.*

Proof (a) For $i = 1, 2$, given $\mu_i \in \mathscr{P}(S)$ with $I(\mu_i) < \infty$, there is $\gamma_i \in A(\mu_i)$ such that $I(\mu_i) = R\left(\gamma_i \| \, \mu_i \otimes p\right)$ [part (b) of Lemma 6.8]. Given $\lambda \in (0, 1)$,

$$[\lambda\gamma_1 + (1-\lambda)\gamma_2]_1 = [\lambda\gamma_1 + (1-\lambda)\gamma_2]_2 = \lambda\mu_1 + (1-\lambda)\mu_2.$$

Using the definition of I and that $R(\cdot \| \cdot)$ is convex on $\mathscr{P}(S^2)^2$, for such λ, we have

$$\begin{aligned} I(\lambda\mu_1 + (1-\lambda)\mu_2) &\le R\left(\lambda\gamma_1 + (1-\lambda)\gamma_2 \| \, \lambda\mu_1 \otimes p + (1-\lambda)\mu_2 \otimes p\right) \\ &\le \lambda R\left(\gamma_1 \| \, \mu_1 \otimes p\right) + (1-\lambda) R\left(\gamma_2 \| \, \mu_2 \otimes p\right) \\ &= \lambda I(\mu_1) + (1-\lambda) I(\mu_2). \end{aligned}$$

Note that the inequality is also true if $I(\mu_1)$ or $I(\mu_2)$ is ∞. Thus I is convex.

(b) We next show that $I(\mu) = 0$ if and only if μ is an invariant under p. If μ is invariant under p, then $\gamma = \mu \otimes p$ satisfies $\gamma \in A(\mu)$, and thus $I(\mu) = 0$. Conversely, if $I(\mu) = 0$, then by part (b) of Lemma 6.8, there is $\gamma \in A(\mu)$ such that $R\left(\gamma \| \, \mu \otimes p\right) = 0$, which implies $\gamma = \mu \otimes p$. Since $[\mu \otimes p]_2 = [\gamma]_2 = \mu$, μ is invariant for p.

(c) We next show that I is lower semicontinuous under Condition 6.2. Let $\{\mu_n\}_{n\in\mathbb{N}}$ converge weakly to μ. Without loss of generality we can assume that $I(\mu_n) < \infty$ for each n. Thus for each n there is $\gamma_n \in A(\mu_n)$ such that $I(\mu_n) = R\left(\gamma_n \| \, \mu_n \otimes p\right)$. It suffices to prove that for every subsequence of $\{\mu_n\}$ there is a subsubsequence (relabeled as $\{n\}$) such that

$$\liminf_{n\to\infty} I(\mu_n) \ge I(\mu).$$

Since each marginal of γ_n is μ_n and $\mu_n \Rightarrow \mu$, $\{\gamma_n\}$ is tight. Hence there is a subsequence such that $\gamma_n \Rightarrow \gamma \in A(\mu)$. Since $\mu_n \otimes p \Rightarrow \mu \otimes p$ along this subsequence

by Lemma 6.7, the lower semicontinuity of relative entropy implies

$$\liminf_{n\to\infty} I(\mu_n) = \liminf_{n\to\infty} R(\gamma_n \| \mu_n \otimes p) \geq R(\gamma \| \mu \otimes p) \geq I(\mu),$$

completing the proof that I is lower semicontinuous.

(d) Using the lower semicontinuity established in part (c), $\{\mu : I(\mu) \leq M\}$ is closed for each $M < \infty$. Compactness is now immediate on observing that when S is compact, so is $\mathscr{P}(S)$. □

The following lemma gives the existence of an invariant measure for $p(x, dy)$.

Lemma 6.11 *Suppose that Conditions 6.2 and 6.4 hold. Then there is at least one invariant probability for p.*

Proof Since $\{EL^n\}_{n\in\mathbb{N}}$ is tight, along some subsequence, EL^n converges to a probability measure π. Since $p(X_i, \cdot)$ gives the conditional distribution of X_{i+1} given $\sigma\{X_0, \ldots X_i\}$, it follows that for all $f \in \mathscr{C}_b(S)$,

$$\begin{aligned}
&\left| \int_S f(x) EL^n(dx) - \int_{S^2} f(y) p(x, dy) EL^n(dx) \right| \\
&= \frac{1}{n} \left| E \sum_{i=0}^{n-1} f(X_i) - E \sum_{i=0}^{n-1} \int_S f(y) p(X_i, dy) \right| \\
&= \frac{1}{n} |Ef(X_0) - Ef(X_n)| \\
&\leq \frac{2}{n} \|f\|_\infty \\
&\to 0
\end{aligned}$$

as $n \to \infty$. The Feller property implies that $x \to \int_S f(y) p(x, dy)$ is bounded and continuous. Hence taking the limit of EL^n along the subsequence, we have $\int_S f(x)\pi(dx) = \int_{S^2} f(y) p(x, dy)\pi(dx)$ for every $f \in \mathscr{C}_b(S)$, and thus π is invariant. □

6.6 Tightness and Weak Convergence

As in the proof of Sanov's theorem, we will need to relate the weak limits of the controlled empirical measure $\bar{L}^n$ defined in (6.2) to the random measures λ^n defined in (6.5), which links $\bar{\mu}_i^n$, the measure used to pick $\bar{X}_i^n$, to the value of $\bar{X}_{i-1}^n$. The proof of the following result is almost identical to that of Lemma 3.5, and is included here only for completeness. Since it will be used for proving the uniform versions of the large deviation bounds, we use the notation E_{x_0} to denote that the uncontrolled and controlled chains start at x_0.

Lemma 6.12 *Let $\{x_0^n\}_{n\in\mathbb{N}}$ be a sequence in S. Let $\{\bar{X}_i^n\}$ be as in Sect. 6.2 with $\bar{X}_0^n = x_0^n$ for $n \in \mathbb{N}$. Define $\bar{L}^n$ and λ^n through (6.2) and (6.5). Suppose that $\{(\bar{L}^n, \lambda^n)\}_{n\in\mathbb{N}}$ converges weakly along a subsequence to $(\bar{L}, \lambda)$. Then $\bar{L} = [\lambda]_1 = [\lambda]_2$ a.s.*

Proof Since $\bar{L}^n = [\lambda^n]_1$, we have $\bar{L} = [\lambda]_1$. Thus we need only show that $\bar{L} = [\lambda]_2$. We use the fact that for every Polish space S there is a countable separating class $\{f_m\}_{m\in\mathbb{N}}$ of bounded continuous functions (see Appendix A). Define $\Delta_{m,i}^n \doteq f_m\left(\bar{X}_i^n\right) - \int_S f_m(y)\,\bar{\mu}_i^n(dy)$ and $K_m \doteq \|f_m\|_\infty$. For all $\varepsilon > 0$,

$$P_{x_0^n}\left\{\left|\frac{1}{n}\sum_{i=1}^n f_m\left(\bar{X}_i^n\right) - \frac{1}{n}\sum_{i=1}^n \int_S f_m(y)\,\bar{\mu}_i^n(dy)\right| > \varepsilon\right\} \le \frac{1}{\varepsilon^2} E_{x_0^n}\left[\frac{1}{n^2}\sum_{i,j=1}^n \Delta_{m,i}^n \Delta_{m,j}^n\right].$$

Recall that $\mathscr{F}_j^n = \sigma(\bar{X}_i^n, i = 1, \ldots, j)$. The off-diagonal terms in the expected value vanish: for $i > j$,

$$\begin{aligned} E_{x_0^n}\left[\Delta_{m,i}^n \Delta_{m,j}^n\right] &= E_{x_0^n}\left[E_{x_0^n}\left[\Delta_{m,i}^n \Delta_{m,j}^n \middle| \mathscr{F}_{i-1}^n\right]\right] \\ &= E_{x_0^n}\left[E_{x_0^n}\left[\Delta_{m,i}^n \middle| \mathscr{F}_{i-1}^n\right] \Delta_{m,j}^n\right] \\ &= 0. \end{aligned}$$

Suppose n is large enough that $2K_m/n \le \varepsilon/2$. Since $|\Delta_{m,i}^n| \le 2K_m$ and $[\lambda^n]_2 = \frac{1}{n}\sum_{i=1}^n \bar{\mu}_i^n(dx)$, we have

$$\begin{aligned} &P_{x_0^n}\left\{\left|\int_S f_m(x)\,\bar{L}^n(dx) - \int_{S^2} f_m(y)\lambda^n(dx \times dy)\right| > \varepsilon\right\} \\ &= P_{x_0^n}\left\{\left|\frac{1}{n}\sum_{i=0}^{n-1} f_m(\bar{X}_i^n) - \frac{1}{n}\sum_{i=1}^n \int_S f_m(y)\bar{\mu}_i^n(dy)\right| > \varepsilon\right\} \\ &\le P_{x_0^n}\left\{\left|\frac{1}{n}\sum_{i=1}^n f_m\left(\bar{X}_i^n\right) - \frac{1}{n}\sum_{i=1}^n \int_S f_m(y)\,\bar{\mu}_i^n(dy)\right| > \varepsilon/2\right\} \\ &\le \frac{16K_m^2}{n\varepsilon^2}. \end{aligned}$$

Since $(\bar{L}^n, \lambda^n) \Rightarrow (\bar{L}, \lambda)$ and $\varepsilon > 0$ is arbitrary, by Fatou's lemma we have

$$P\left\{\int_S f_m(x)\,\bar{L}(dx) = \int_S f_m(y)\,\lambda(dx \times dy)\right\} = 1.$$

Since $\{f_m\}$ is countable and separating, we conclude that $\bar{L} = [\lambda]_2$ a.s. □

6.7 Laplace Upper Bound

In this section we prove the Laplace upper bound, which is the variational lower bound. In fact, the result will give a uniform Laplace upper bound in the sense of Definition 1.11. Recall that P_{x_n} denotes the probability measure under which $X_0 = x_n$ a.s.

Proposition 6.13 *Assume that S is compact and that Condition 6.2 holds. Let $\{L^n\}_{n\in\mathbb{N}}$ be the empirical measures defined by (6.1). Let $\{x_n\} \subset S$ be any sequence. Define $I : \mathscr{P}(S) \to [0, \infty]$ by (6.8). Then*

$$\liminf_{n\to\infty} -\frac{1}{n} \log E_{x_n} e^{-nF(L^n)} \geq \inf_{\mu\in\mathscr{P}(S)} [F(\mu) + I(\mu)]. \tag{6.10}$$

Proof An application of the representation (6.3) and (6.4) gives

$$-\frac{1}{n} \log E_{x_n} e^{-nF(L^n)} + \frac{1}{n} \geq E_{x_n} \left[F(\bar{L}^n) + R\left(\lambda^n \,\middle\|\, \bar{L}^n \otimes p\right)\right], \tag{6.11}$$

where $\bar{L}^n$ is the controlled empirical measure associated with a control $\left\{\bar{\mu}_i^n\right\}$ that comes within $1/n$ of the infimum, and λ^n is defined as in (6.5). Since $\{(\bar{L}^n, \lambda^n)\}$ take values in a compact set, there exists a subsequence along which we have convergence in distribution to a limit $(\bar{L}, \lambda)$. It suffices by a standard argument by contradiction to prove (6.10) for this subsequence. Letting $n \to \infty$ in (6.11) gives the following:

$$\begin{aligned}
\liminf_{n\to\infty} -\frac{1}{n} \log E_{x_n} e^{-nF(L^n)} &\geq E\left[F(\bar{L}) + R\left(\lambda \| \, \bar{L} \otimes p\right)\right] \\
&\geq \inf_{\mu\in\mathscr{P}(S)} \left[F(\mu) + \inf_{\gamma\in A(\mu)} R(\gamma \| \, \mu \otimes p) \right] \\
&= \inf_{\mu\in\mathscr{P}(S)} [F(\mu) + I(\mu)].
\end{aligned}$$

The first inequality uses the weak convergence, boundedness, and continuity of F, Fatou's lemma, lower semicontinuity of $R(\cdot\|\cdot)$, and Lemma 6.7. The second inequality follows from Lemma 6.12, and the last equality uses the definition of I in (6.8). □

Remark 6.14 The proof of the upper bound just given uses the compactness of the state space to argue the tightness of $\{(\bar{L}^n, \lambda^n)\}$. In Sect. 6.10 we outline an argument for the case in which S is not compact but the Markov chain $\{X_i\}$ satisfies a suitable stability condition.

6.8 Laplace Lower Bound

In this section we prove the Laplace lower bound. A uniform Laplace lower bound under the stronger condition (6.9) will be proved in Sect. 6.9. Recall that P_x denotes the probability measure under which $X_0 = x$ a.s.

Proposition 6.15 *Assume Conditions 6.2, 6.3 and 6.4 and let $\{L^n\}_{n\in\mathbb{N}}$ be the empirical measures defined by (6.1). Let $x \in S$ be given and define $I : \mathscr{P}(S) \to [0,\infty]$ by (6.8). Then*

$$\limsup_{n\to\infty} -\frac{1}{n} \log E_x e^{-nF(L^n)} \le \inf_{\mu\in\mathscr{P}(S)} [F(\mu) + I(\mu)]\,. \tag{6.12}$$

In Lemma 6.16 below we will present two key facts that hold under Conditions 6.2, 6.3, and 6.4. The first is that $p(x,dy)$ has a unique stationary distribution π and the Markov chain associated with π and $p(x,dy)$ is ergodic (see [34, Sect. 6.3]), and the second is that $I(\mu) < \infty$ implies $\mu \ll \pi$.

The bound (6.12) almost follows from just these facts. Indeed, let $\varepsilon > 0$ and let ν satisfy $F(\nu) + I(\nu) \le \inf_\mu [F(\mu) + I(\mu)] + \varepsilon$. From the definition of I there is $\gamma \in \mathscr{P}(S^2)$ such that $[\gamma]_1 = [\gamma]_2 = \nu$ and

$$R(\gamma \| \nu \otimes p) \le I(\nu) + \varepsilon.$$

Since $[\gamma]_1 = [\gamma]_2 = \nu$, there is $q(x,dy)$ such that $\gamma(dx \times dy) = \nu(dx)q(x,dy)$ and ν is invariant under q. Moreover, by the chain rule,

$$\infty > R(\gamma \| \nu \otimes p) = \int_S R(q(x,\cdot) \| p(x,\cdot))\, \nu(dx).$$

If q were ergodic, we could use it to define controls for the representation via $\bar{\mu}_i^n(\cdot) = q(\bar{X}_{i-1}^n, \cdot)$. Using these controls, the representation and the L^1-ergodic theorem would show that (at least for certain initial conditions x)

$$\limsup_{n\to\infty} -\frac{1}{n} \log E_x e^{-nF(L^n)} \le \inf_{\mu\in\mathscr{P}(S)} [F(\mu) + I(\mu)] + 2\varepsilon,$$

and since $\varepsilon > 0$ is arbitrary we would be done. However, the problem is that $q(x,dy)$ need not be ergodic. To deal with this, we add a little bit of p to q, which, as we will see in Lemma 6.17, makes the combination ergodic.

Lemma 6.16 *Suppose Condition 6.3 holds and that the transition probability function $p(x,dy)$ has an invariant probability measure π. Let I be as in (6.8). Then the following conclusions hold.*

(a) *π is the unique invariant probability measure for $p(x,dy)$, and the Markov chain having π as its initial distribution and $p(x,dy)$ as its transition probability function is ergodic.*

(b) *Let A be a Borel set such that $p^{(\ell_0)}(x_0, A) > 0$ for some $x_0 \in S$, where ℓ_0 is as in Condition 6.3. Then $\pi(A) > 0$.*

(c) *For $\nu \in \mathscr{P}(S)$, $I(\nu) < \infty$ implies $\nu \ll \pi$.*

Proof (a) We prove that $p(x, dy)$ is indecomposable, i.e., that there do not exist disjoint Borel sets A_1 and A_2 such that

$$p(x, A_1) = 1 \text{ for all } x \in A_1 \text{ and } p(y, A_2) = 1 \text{ for all } y \in A_2. \tag{6.13}$$

Then from [34, Theorem 7.16], π is unique, and the Markov chain associated with π and $p(x, dy)$ is ergodic, completing the proof of part (a).

Suppose that there exist disjoint Borel sets A_1 and A_2 such that formula (6.13) holds. Then for all ℓ and n in $\mathbb{N}$ we have $p^{(\ell)}(x, A_1) = 1$ for all $x \in A_1$ and $p^{(n)}(y, A_2) = 1$ for all $y \in A_2$. According to Condition 6.3, the first of these equalities with $\ell = \ell_0$ guarantees that for each $y \in A_2$, there exists $j \geq n_0$ such that $p^{(j)}(y, A_1) > 0$. Since A_1 and A_2 are disjoint, for $y \in A_2$ this inequality is incompatible with the equality $p^{(n)}(y, A_2) = 1$ for all $n \in \mathbb{N}$. This contradiction shows that $p(x, dy)$ is indecomposable.

(b) Let A be a set as in the statement of part (b). Define a function ι mapping S into $\mathbb{N}$ by

$$\iota(\zeta) \doteq \min\{j \in \mathbb{N} : j \geq n_0 \text{ and } p^{(j)}(\zeta, A) > 0\}.$$

Condition 6.3 guarantees that $\iota(\zeta)$ is nonempty, and so ι is a well-defined measurable map. The set S can be written as the disjoint union of Borel sets $\bigcup_{j=n_0}^{\infty} \Sigma^{(j)}$, where

$$\Sigma^{(j)} \doteq \{\zeta \in S : \iota(\zeta) = j\}.$$

Clearly, $\pi(\Sigma^{(k)}) > 0$ for some $k \in \mathbb{N}$ satisfying $k \geq n_0$. In addition, $p^{(k)}(\zeta, A) > 0$ for all $\zeta \in \Sigma^{(k)}$. Hence

$$\pi(A) = \int_S p^{(k)}(\zeta, A)\, \pi(d\zeta) \geq \int_{\Sigma^{(k)}} p^{(k)}(\zeta, A)\, \pi(d\zeta) > 0.$$

(c) Let $\nu \in \mathscr{P}(S)$ satisfy $I(\nu) < \infty$. By part (b) of Lemma 6.8 there exists a transition probability function $q(x, dy)$ that has ν as an invariant measure and satisfies

$$\int_S R(q(x, \cdot) \| p(x, \cdot))\, \nu(dx) = I(\nu) < \infty.$$

The set $\Delta \doteq \{x \in S : q(x, \cdot) \ll p(x, \cdot)\}$ satisfies $\nu(\Delta) = 1$. Since $q(x, dy)$ has ν as an invariant measure, the transition probability function

$$\tilde{q}(x, \cdot) \doteq \begin{cases} q(x, \cdot) & \text{if } x \in \Delta, \\ p(x, \cdot) & \text{if } x \in \Delta^c, \end{cases}$$

also has ν as an invariant measure. We will prove that

$$\tilde{q}^{(\ell_0)}(x,\cdot) \ll p^{(\ell_0)}(x,\cdot) \text{ for all } x \in \Delta, \tag{6.14}$$

where ℓ_0 is the number appearing in part (a) of Condition 6.3.

Part (c) of the present lemma is an immediate consequence of this assertion. Indeed, suppose that $\nu(A) > 0$ for some Borel set A. Iterating the equation $\nu\tilde{q} = \nu$ yields

$$\int_S \tilde{q}^{(\ell_0)}(x, A)\, \nu(dx) = \nu(A) > 0.$$

Hence there exists a Borel set B such that $\nu(B) > 0$ and $\tilde{q}^{(\ell_0)}(x, A) > 0$ for all $x \in B$. By formula (6.14), $p^{(\ell_0)}(x_0, A) > 0$ for all $x_0 \in B \cap \Delta$. (Note that $B \cap \Delta$ is nonempty, since $\nu(B \cap \Delta) = \nu(B) > 0$.) Part (b) implies that $\pi(A) > 0$. This shows that $\nu \ll \pi$ and completes the proof of part (c).

We now prove (6.14). In fact, we will prove by induction that for each $k \in \mathbb{N}$,

$$\tilde{q}^{(k)}(x,\cdot) \ll p^{(k)}(x,\cdot) \text{ for all } x \in \Delta. \tag{6.15}$$

For $k = 1$ this assertion is immediate by the definition of Δ. For any $j \in \mathbb{N}$, we now assume (6.15) for all $k \in \{1, 2, \ldots, j\}$ and prove it for $k = j + 1$. For $x \in \Delta$, let C be a Borel set such that

$$p^{(j+1)}(x, C) = \int_S p(y, C)\, p^{(j)}(x, dy) = 0.$$

This implies that there exists a Borel set Γ such that $p^{(j)}(x, \Gamma) = 1$ and $p(y, C) = 0$ for all $y \in \Gamma$. Since $x \in \Delta$, the inductive hypothesis implies that $\tilde{q}^{(j)}(x, \Gamma) = 1$. It also implies that for $y \in \Gamma \cap \Delta$, $\tilde{q}(y, C) = 0$. Hence for all $x \in \Delta$,

$$\begin{aligned}
\tilde{q}^{(j+1)}(x, C) &= \int_S \tilde{q}(y, C)\, \tilde{q}^{(j)}(x, dy) \\
&= \int_\Gamma \tilde{q}(y, C)\, \tilde{q}^{(j)}(x, dy) \\
&= \int_{\Gamma\cap\Delta} \tilde{q}(y, C)\, \tilde{q}^{(j)}(x, dy) \; + \int_{\Gamma\cap\Delta^c} \tilde{q}(y, C)\, \tilde{q}^{(j)}(x, dy) \\
&= \int_{\Gamma\cap\Delta} \tilde{q}(y, C)\, \tilde{q}^{(j)}(x, dy) \; + \int_{\Gamma\cap\Delta^c} p(y, C)\, \tilde{q}^{(j)}(x, dy) \\
&= 0,
\end{aligned}$$

where the next to last equality follows on noting that $\tilde{q} = p$ on Δ^c. This proves formula (6.15) for $k = j + 1$. The proof of the lemma is complete. □

We denote the total variation norm of any signed measure γ on S by

$$\|\gamma\|_{\text{TV}} \doteq \sup_f \left| \int_S f(x)\gamma(dx) \right|,$$

where the supremum is taken over all measurable functions bounded by 1. Note that under Conditions 6.2, 6.3, and 6.4, from Lemmas 6.16 and 6.11 there is a unique invariant probability measure for $p(x, dy)$, which we denote by π.

Lemma 6.17 *Assume Conditions 6.2, 6.3, and 6.4. Let $\nu \in \mathscr{P}(S)$ satisfy $I(\nu) < \infty$. Then given $\delta > 0$, there exists $\nu^* \in \mathscr{P}(S)$ with the following properties:*

(a) $\|\nu^* - \nu\|_{TV} \le \delta$;

(b) $\pi \ll \nu^*$ *and* $\nu^* \ll \pi$, *where π is the unique invariant measure of $p(x, dy)$;*

(c) *there exists a transition probability function $q^*(x, dy)$ on S such that ν^* is an invariant measure of $q^*(x, dy)$ (in fact, the unique invariant measure) and the associated Markov chain is ergodic. In addition,*

$$I(\nu^*) \le \int_S R(q^*(x, \cdot)\|p(x, \cdot))\, \nu^*(dx) \le I(\nu). \tag{6.16}$$

Proof (a) For $\kappa \in (0, 1)$, the probability measure $\nu^\kappa \doteq (1-\kappa)\nu + \kappa\pi$ satisfies $\|\nu^\kappa - \nu\|_{\text{TV}} = \kappa\|\pi - \nu\|_{\text{TV}} \le 2\kappa$. Hence $\|\nu^\kappa - \nu\|_{\text{TV}} \le \delta$ for all $\kappa \in (0, \delta/2]$. In the statement of the lemma, we take $\nu^* \doteq \nu^\kappa$ for any $\kappa \in (0, \delta/2]$.

(b) Since $\kappa > 0$ and $\nu^*(A) \ge \kappa\pi(A)$ for all Borel sets A, it follows that $\pi \ll \nu^*$. Since $I(\nu) < \infty$, part (c) of Lemma 6.16 implies that $\nu \ll \pi$. Thus $\nu^* \ll \pi$.

(c) Using part (b) of Lemma 6.8, we choose a transition probability function $q(x, dy)$ on S that has ν as an invariant measure and satisfies

$$\int_S R(q(x, \cdot)\|p(x, \cdot))\, \nu(dx) = I(\nu).$$

We then define probability measures γ, θ, and γ^* on $S \times S$ by

$$\gamma \doteq \nu \otimes q,\ \theta \doteq \pi \otimes p, \text{ and } \gamma^* \doteq (1-\kappa)\gamma + \kappa\theta.$$

Since both of the marginals of γ (resp. θ) equal ν (resp. π), both of the marginals of γ^* equal ν^*. Hence there exists a transition probability function $q^*(x, dy)$ on S such that $\gamma^* = \nu^* \otimes q^*$ and ν^* is an invariant measure of $q^*(x, dy)$ [part (a) of Lemma 6.8].

We next verify formula (6.16), using the fact that $R(\cdot\|\cdot)$ is convex and satisfies $R(\alpha\|\alpha) = 0$. We have

$$
\begin{aligned}
I(\nu^*) &\le \int_S R(q^*(x,\cdot)\|p(x,\cdot))\,\nu^*(dx) \qquad (6.17)\\
&= R(\gamma^*\|\nu^*\otimes p)\\
&= R((1-\kappa)\nu\otimes q+\kappa\pi\otimes p\|(1-\kappa)\nu\otimes p+\kappa\pi\otimes p)\\
&\le (1-\kappa)\,R(\nu\otimes q\|\nu\otimes p)+\kappa\,R(\pi\otimes p\|\pi\otimes p)\\
&\le I(\nu).
\end{aligned}
$$

This proves formula (6.16).

We will now show that ν^* is the unique invariant measure for a modification of the transition probability function $x\mapsto q^*(x,\cdot)$ over a set of ν^*-measure zero, and that the associated Markov chain is ergodic. For this, we would like to apply part (a) of Lemma 6.16 with π replaced with ν^* and p replaced by the modification of q^*.

According to part (b) of the present lemma, we have $\nu^*\ll\pi$. Since $\nu^*(A)\ge\kappa\,\pi(A)$ for all Borel sets A, the Radon–Nikodym derivative $f(x)\doteq\frac{d\nu^*}{d\pi}(x)$ satisfies $f(x)\in[\kappa,\infty)$ π-a.s. for $x\in S$. For all Borel sets A and B,

$$
\int_A q^*(x,B)\,f(x)\,\pi(dx)=\gamma^*(A\times B)\ge\kappa\,\theta(A\times B)=\kappa\int_A p(x,B)\,\pi(dx).
$$

This together with separability of S implies that π-a.s. for $x\in S$,

$$
q^*(x,B)\ge\frac{\kappa}{f(x)}\,p(x,B)
$$

for all Borel B, and consequently, for such x,

$$
p(x,\cdot)\ll q^*(x,\cdot). \qquad (6.18)
$$

The formula

$$
\int_S R(q^*(x,\cdot)\|p(x,\cdot))\,\nu^*(dx)\le I(\nu)<\infty,
$$

implied by (6.17), shows that ν^*-a.s. for $x\in S$,

$$
q^*(x,\cdot)\ll p(x,\cdot). \qquad (6.19)
$$

Since π and ν^* are mutually absolutely continuous (part (b)), there exists a Borel set C such that $\pi(C)=0=\nu^*(C)$ and both (6.18) and (6.19) hold on the complement of C. We now redefine $q^*(x,dy)$ to equal $p(x,dy)$ for $x\in C$. Since $\nu^*(C)=0$, ν^* remains an invariant measure of the modified $q^*(x,dy)$. In addition, (6.18) and (6.19) are then valid for all $x\in S$, implying in turn that for each $j\in\mathbb{N}$ and all $x\in S$ the measures $p^{(j)}(x,\cdot)$ and $q^{*(j)}(x,\cdot)$ are mutually absolutely continuous. It follows that the absolute continuity condition (6.7) on $p(x,dy)$ implies the following absolute continuity property for $q^*(x,dy)$

$$\sum_{i=\ell_0}^{\infty} \frac{1}{2^i} (q^*)^{(i)}(x, dy) \ll \sum_{j=n_0}^{\infty} \frac{1}{2^j} (q^*)^{(j)}(\zeta, dy)$$

Applying part (a) of Lemma 6.16 to the pair (ν^*, q^*) completes the proof of part (c) of the present lemma. □

Proof (*of Proposition* 6.15) In order to prove the Laplace principle lower bound stated in Proposition 6.15, it suffices to consider only bounded Lipschitz continuous functions F [Corollary 1.10]. More precisely, the space $\mathscr{P}(S)$ that is considered with the usual weak convergence topology can be metrized using the Dudley metric (see [92]) d_{BL} defined by

$$d_{\mathrm{BL}}(\nu_1, \nu_2) \doteq \sup_f \left| \int_S f(x)\nu_1(dx) - \int_S f(x)\nu_2(dx) \right|,$$

where the supremum is taken over all real bounded and Lipschitz continuous functions f on S that are bounded by 1 and whose Lipschitz constant is also bounded by 1. A Lipschitz function F on $\mathscr{P}(S)$ is one for which

$$\sup_{\nu_1 \neq \nu_2} \frac{|F(\nu_1) - F(\nu_2)|}{d_{\mathrm{BL}}(\nu_1, \nu_2)} < \infty.$$

For a subset Φ of S to be determined below, we will first prove for $x \in \Phi$ and every bounded Lipschitz continuous function F the lower bound

$$\liminf_{n\to\infty} \frac{1}{n} \log E_x e^{-nF(L^n)} \geq - \inf_{\mu \in \mathscr{P}(S)} [F(\mu) + I(\mu)]. \tag{6.20}$$

Afterward, we will see how to convert this into a proof of the Laplace principle lower bound for all $x \in S$.

As usual, the proof will be carried out by working with the representation

$$-\frac{1}{n} \log E_x e^{-nF(L^n)} = \inf_{\{\bar{\mu}_i^n\}} E_x \left[F(\bar{L}^n) + \frac{1}{n} \sum_{i=1}^{n} R\left(\bar{\mu}_i^n(\cdot) \,\|\, p(\bar{X}_{i-1}^n, \cdot)\right) \right].$$

Using this representation formula, we will prove (6.20) by proving the upper limit

$$\limsup_{n\to\infty} \inf_{\{\bar{\mu}_i^n\}} E_x \left[F(\bar{L}^n) + \frac{1}{n} \sum_{i=1}^{n} R\left(\bar{\mu}_i^n(\cdot) \,\|\, p(\bar{X}_{i-1}^n, \cdot)\right) \right] \leq \inf_{\mu \in \mathscr{P}(S)} [F(\mu) + I(\mu)]. \tag{6.21}$$

Let $\varepsilon > 0$ be given and choose $\nu \in \mathscr{P}(S)$ such that

$$F(\nu) + I(\nu) \leq \inf_{\mu \in \mathscr{P}(S)} [F(\mu) + I(\mu)] + \varepsilon < \infty. \tag{6.22}$$

Since F is continuous and since convergence in total variation norm implies weak convergence, Lemma 6.17 yields the existence of $\nu^* \in \mathscr{P}(S)$ and a transition probability function $q^*(x, dy)$ with the following properties: ν^* is an invariant measure of $q^*(x, dy)$, the Markov chain with initial distribution ν^* and transition probability function $q^*(x, dy)$ is ergodic, and

$$F(\nu^*) \le F(\nu) + \varepsilon, \quad \int_S R(q^*(x,\cdot)\|p(x,\cdot))\,\nu^*(dx) \le I(\nu) < \infty. \tag{6.23}$$

For each $j \in \{0, 1, \dots, n-1\}$ we define

$$\bar{\mu}_j^n(dy) = q^*(\bar{X}_{j-1}^n, dy)$$

and consider two different choices for the distribution of $\bar{X}_0^n$: δ_x and ν^*. In the first case, the corresponding measure on $(\Omega, \mathscr{F})$ is denoted by P_x. In the second case, the controlled process $\{\bar{X}_j^n, j = 0, 1, \dots, n\}$ equals the first $(n+1)$ steps of the ergodic Markov chain with initial distribution ν^* and transition probability function $q^*(x, dy)$. The corresponding measure on $(\Omega, \mathscr{F})$ is denoted by P^*.

We first study the convergence of the running costs

$$\frac{1}{n}\sum_{j=0}^{n-1} R\big(\bar{\mu}_j^n(\cdot)\|p\left(\bar{X}_j^n,\cdot\right)\big) = \frac{1}{n}\sum_{j=0}^{n-1} R\big(q^*\left(\bar{X}_j^n,\cdot\right)\|p\left(\bar{X}_j^n,\cdot\right)\big).$$

Define

$$D^n \doteq E^*\left[\left|\frac{1}{n}\sum_{j=0}^{n-1} R\big(q^*\left(\bar{X}_j^n,\cdot\right)\|p\left(\bar{X}_j^n,\cdot\right)\big) - \int_S R\left(q^*(\xi,\cdot)\|p(\xi,\cdot)\right)\nu^*(d\xi)\right|\right],$$

where E^* denotes expectation with respect to P^*, and

$$D_x^n \doteq E_x\left[\left|\frac{1}{n}\sum_{j=0}^{n-1} R\big(q^*(\bar{X}_j^n,\cdot)\|p(\bar{X}_j^n,\cdot)\big) - \int_S R(q^*(\xi,\cdot)\|p(\xi,\cdot))\,\nu^*(d\xi)\right|\right],$$

where E_x denotes expectation with respect to P_x. Since the relative entropy is nonnegative and from (6.17), we have

$$E^*\big[R\big(q^*(\bar{X}_j^n,\cdot)\|p(\bar{X}_j^n,\cdot)\big)\big] = \int_S R\big(q^*(\xi,\cdot)\|p(\xi,\cdot)\big)\,\nu^*(d\xi) \le I(\nu) < \infty,$$

we can apply the L^1-ergodic theorem [34, Sect. 6.5], which implies that

$$\lim_{n\to\infty} D^n = \lim_{n\to\infty}\int_S D_x^n\,\nu^*(dx) = 0.$$

Hence by Chebyshev's inequality, for all $c > 0$,

$$\lim_{n\to\infty} \nu^*\{x \in S : D_x^n \geq c\} = 0.$$

This convergence in probability guarantees that for every subsequence of $\{n\}$ there exist a further subsequence (relabeled as $\{n\}$) and a Borel set Φ_1 such that $\nu^*(\Phi_1) = 1$ and such that whenever $x \in \Phi_1$,

$$\lim_{n\to\infty} D_x^n = \lim_{n\to\infty} E_x \left| \frac{1}{n} \sum_{j=0}^{n-1} R\big(q^*(\bar{X}_j^n, \cdot) \| p(\bar{X}_j^n, \cdot)\big) - \int_S R(q^*(\xi, \cdot) \| p(\xi, \cdot))\, \nu^*(d\xi) \right| = 0. \tag{6.24}$$

We now fix the subsubsequence and consider the convergence of the corresponding sequence of controlled empirical measures $\{\bar{L}^n\}$. Since S is a Polish space, as noted in Appendix A there exists a countable convergence determining class Ξ of bounded and continuous functions on S. For $g \in \Xi$ define

$$A(g) \doteq \left\{ \omega \in \Omega : \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} g(\bar{X}_j^n(\omega)) = \int_S g\, d\nu^* \right\}.$$

The pointwise ergodic theorem [34, Sect. 6.5] implies that

$$P^*\{A(g)\} = \int_S P_x\{A(g)\}\, \nu^*(dx) = 1,$$

which in turn implies that there exists a Borel set $\Phi_2(g)$ such that $\nu^*(\Phi_2(g)) = 1$ and $P_x(A(g)) = 1$ whenever $x \in \Phi_2(g)$. Define $\Phi_2 \doteq \cap_{g\in\Xi} \Phi_2(g)$. Then whenever $x \in \Phi_2$, we have P_x-a.s.

$$\lim_{n\to\infty} \int_S g\, d\bar{L}^n = \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} g(\bar{X}_j^n) = \int_S g\, d\nu^*$$

for all $g \in \Xi$. Since Ξ is a convergence determining class, it follows that for all $x \in \Phi_2$, $\bar{L}^n \Rightarrow \nu^*$, P_x-a.s. The continuity of F on $\mathscr{P}(S)$ then implies that

$$\lim_{n\to\infty} F(\bar{L}^n) = F(\nu^*).$$

We now put these facts together. Define $\Phi \doteq \Phi_1 \cap \Phi_2$. Then $\nu^*(\Phi) = 1$, and whenever $x \in \Phi$,

$$\begin{aligned}
&\limsup_{n\to\infty} -\frac{1}{n}\log E_x e^{-nF(L^n)} \\
&\quad\le \lim_{n\to\infty} E_x\left[F(\bar{L}^n) + \frac{1}{n}\sum_{j=0}^{n-1} R\big(q^*(\bar{X}_j^n,\cdot)\|p(\bar{X}_j^n,\cdot)\big)\right] \\
&\quad= F(\nu^*) + \int_S R(q^*(\xi,\cdot)\|p(\xi,\cdot))\,\nu^*(d\xi) \\
&\quad\le F(\nu) + I(\nu) + \varepsilon \\
&\quad\le \inf_{\mu\in\mathscr{P}(S)}[F(\mu) + I(\mu)] + 2\varepsilon,
\end{aligned}$$

where the equality on the third line is from (6.24), the inequality on the fourth line uses (6.23), and the inequality on the last line is from (6.22). Since the chosen subsequence of $\{n\}$ was arbitrary, sending $\varepsilon \to 0$ yields the upper limit (6.21) whenever $x \in \Phi$.

A short argument will extend this upper limit to all $x \in S$, thus proving the Laplace principle lower bound. This will be carried out using the Lipschitz continuity of F. Let ℓ_0 be the number occurring in Condition 6.3. Since $\mu \ll \nu^*$ [Lemma 6.17 (b)] and $\nu^*(\Phi) = 1$, we have that $\mu(\Phi) = 1$. We claim that this implies

$$p^{(\ell_0)}(x, \Phi) = 1 \quad \text{for all } x \in S. \tag{6.25}$$

If this claim were not true, then for some $x_0 \in S$ we would have $p^{(\ell_0)}(x_0, \Phi^c) > 0$ and thus, by part (b) of Lemma 6.16, $\mu(\Phi^c) > 0$. Since this contradicts $\mu(\Phi) = 1$, the claim (6.25) is proved.

For $n \in \mathbb{N}$ we define $\mathscr{P}(S)$-valued random variables $\tilde{L}^n$ by

$$\tilde{L}^n \doteq \frac{1}{n}\sum_{j=\ell_0}^{n+\ell_0-1} \delta_{X_j}.$$

Then for all $\omega \in \Omega$,

$$\left\|\tilde{L}^n - L^n\right\|_{\text{TV}} \le \frac{2\ell_0}{n}.$$

Hence, denoting by $M < \infty$ the Lipschitz constant of F with respect to the Dudley metric, and noting that by the definition of the two distances, $d_{\text{BL}}(L^n, \tilde{L}^n) \le \|L^n - \tilde{L}^n\|_{\text{TV}}$, we have for all $\omega \in \Omega$,

$$F(L^n) \le F(\tilde{L}^n) + M\, d_{\text{BL}}(L^n, \tilde{L}^n) \le F(\tilde{L}^n) + 2\ell_0 M/n.$$

For each $x \in S$ we now have for all $n \in \mathbb{N}$,

$$
\begin{aligned}
&E_x \exp\{-n\,F(L^n)\} \\
&\quad \geq \exp\{-2\ell_0 M\} E_x \exp\{-n\,F(\tilde{L}^n)\} \\
&\quad = \exp\{-2\ell_0 M\} \int_S E\Big[\exp\{-n\,F(\tilde{L}^n)\} | X_{\ell_0} = y\Big]\, p^{(\ell_0)}(x, dy) \\
&\quad = \exp\{-2\ell_0 M\} \int_{\Phi} E_y\big[\exp\{-n\,F(L^n)\}\big]\, p^{(\ell_0)}(x, dy), \qquad (6.26)
\end{aligned}
$$

where the last equality uses the Markov property and (6.25).

Let $\varepsilon > 0$ be given. Since we have already established the estimate (6.20) for $y \in \Phi$, we have for each such y an $N(y, \varepsilon) \in \mathbb{N}$ such that for all $n \geq N(y, \varepsilon)$,

$$
-\frac{1}{n} \log E_y e^{-nF(L^n)} \leq \inf_{\mu \in \mathscr{P}(S)} [F(\mu) + I(\mu)] + \varepsilon.
$$

We assume that $N(y, \varepsilon)$ is the minimal positive integer with this property. Then the function $y \mapsto N(y, \varepsilon)$ is a measurable map from S to $\mathbb{N}$, and Φ can be written as the disjoint union of Borel sets $\cup_{i=1}^{\infty} \Phi^{(i)}$, where

$$
\Phi^{(i)} \doteq \{y \in \Phi : N(y, \varepsilon) = i\}.
$$

Since $p^{(\ell_0)}(x, \Phi) = 1$, there exists $i_0 \in \mathbb{N}$ such that $p^{(\ell_0)}(x, \Phi^{(i_0)}) > 0$. Hence using (6.26), we have for all $n \geq i_0$,

$$
\begin{aligned}
&E_x \exp\{-n\,F(L^n)\} \\
&\quad \geq \exp\{-2\ell_0 M\} \int_{\Phi} E_y[\exp\{-n\,F(L^n)\}]\, p^{(\ell_0)}(x, dy) \\
&\quad \geq \exp\{-2\ell_0 M\} \int_{\Phi^{(i_0)}} E_y\big[\exp\{-n\,F(L^n)\}\big]\, p^{(\ell_0)}(x, dy) \\
&\quad \geq \exp\{-2\ell_0 M\} \exp\left[-n \left(\inf_{\mu \in \mathscr{P}(S)} [F(\mu) + I(\mu)] + \varepsilon\right)\right] p^{(\ell_0)}(x, \Phi^{(i_0)}).
\end{aligned}
$$

This gives

$$
\limsup_{n \to \infty} -\frac{1}{n} \log E_x e^{-nF(L^n)} \leq \inf_{\mu \in \mathscr{P}(S)} [F(\mu) + I(\mu)] + \varepsilon.
$$

Sending $\varepsilon \to 0$, we have that for each $x \in S$,

$$
\limsup_{n \to \infty} -\frac{1}{n} \log E_x e^{-nF(L^n)} \leq \inf_{\mu \in \mathscr{P}(S)} [F(\mu) + I(\mu)]. \qquad (6.27)
$$

The proof of the Laplace principle lower bound is complete. □

6.9 Uniform Laplace Principle

We now prove the second part of Theorem 6.6, which gives a uniform Laplace principle.

Proposition 6.18 *Assume Conditions 6.2 and 6.3 on the Markov chain* $\{X_i\}_{i\in\mathbb{N}_0}$ *and suppose that* S *is compact. Assume also that for each* $x \in S$ *there are* $\beta \in \mathscr{P}(S)$, $k \in \mathbb{N}$, *and* $c > 0$ *such that for all* $y \in S$ *satisfying* $d(y, x) < c$ *and all* $A \in \mathscr{B}(S)$, (6.9) *is satisfied. Let* $\{L^n\}_{n\in\mathbb{N}}$ *denote the sequence of empirical measures defined by* (6.1). *Then for every bounded and continuous* F,

$$\lim_{n\to\infty} \sup_{x\in S} \left| \frac{1}{n} \log E_x e^{-nF(L^n)} + \inf_{\mu\in\mathscr{P}(S)} [F(\mu) + I(\mu)] \right| = 0.$$

Proof The proof is split into upper and lower bounds. If

$$\liminf_{n\to\infty} \inf_{x\in S} -\frac{1}{n} \log E_x e^{-nF(L^n)} \geq \inf_{\mu\in\mathscr{P}(S)} [F(\mu) + I(\mu)] \tag{6.28}$$

is not valid, then there exist $\varepsilon > 0$, a subsequence of n (labeled once more as n), and a sequence $\{x_n\} \subset S$ for which

$$\liminf_{n\to\infty} -\frac{1}{n} \log E_{x_n} e^{-nF(L^n)} \leq \inf_{\mu\in\mathscr{P}(S)} [F(\mu) + I(\mu)] - \varepsilon.$$

However, this contradicts Proposition 6.13, and hence (6.28) is valid.

Next we argue that

$$\limsup_{n\to\infty} \sup_{x\in S} -\frac{1}{n} \log E_x e^{-nF(L^n)} \leq \inf_{\mu\in\mathscr{P}(S)} [F(\mu) + I(\mu)]. \tag{6.29}$$

Here we make use of the additional assumption in (6.9). For $x \in S$ and $c > 0$, let $B(x, c)$ be the open ball $\{y \in S : d(y, x) < c\}$. Let us assume that we have proved

$$\limsup_{n\to\infty} \sup_{y\in B(x,c)} -\frac{1}{n} \log E_y e^{-nF(L^n)} \leq \inf_{\mu\in\mathscr{P}(S)} [F(\mu) + I(\mu)] \tag{6.30}$$

for every bounded Lipschitz continuous function F, $x \in S$, and c associated with x as in the statement of the proposition. This will be taken care of in the next paragraph. Since S is compact, there exist finitely many points $x_1, x_2, \ldots, x_r$ such that S is covered by $\{B(x_i, c_i), i = 1, 2, \ldots, r\}$, where c_i is the radius associated with x_i as in the statement of the proposition. Hence (6.30) implies (6.29).

We now prove the upper limit (6.30). By a similar conditioning argument as in (6.26), for each $y \in B(x, c)$, each $m \in \{1, 2, \ldots, k\}$, and all $n \in \mathbb{N}$,

$$E_y\big[\exp\{-n\,F(L^n)\}\big] \geq \int_S \exp\{-2mM\}\,E_z[\exp\{-n\,F(L^n)\}]\,p^{(m)}(y,dz)$$
$$\geq \int_S \exp\{-2kM\}\,E_z[\exp\{-n\,F(L^n)\}]\,p^{(m)}(y,dz),$$

where M denotes the Lipschitz constant of F. Hence for all $n \in \mathbb{N}$ and $y \in B(x,c)$,

$$E_y[\exp\{-n\,F(L^n)\}]$$
$$\geq \exp\{-2kM\}\int_S E_z[\exp\{-n\,F(L^n)\}]\left(\frac{1}{k}\sum_{m=1}^{k} p^{(m)}(y,dz)\right)$$
$$\geq \frac{c}{k}\exp\{-2kM\}\int_S E_z[\exp\{-n\,F(L^n)\}]\,\beta(dz),$$

where the last inequality uses (6.9). The rest of the proof is similar to the argument used for proving the bound in (6.27) from (6.26), and so we give only a sketch. We will apply Proposition 6.15. Given $\varepsilon > 0$ and $y \in S$, let $N(y,\varepsilon) \in \mathbb{N}$ be the minimal positive integer such that for all $n \geq N(y,\varepsilon)$,

$$-\frac{1}{n}\log E_y e^{-nF(L^n)} \leq \inf_{\mu\in\mathscr{P}(S)}[F(\mu)+I(\mu)]+\varepsilon.$$

Define $\Phi^{(i)} \doteq \{y \in S : N(y,\varepsilon) = i\}$. Then there exists $i_0 \in \mathbb{N}$ such that $\beta(\Phi^{(i_0)}) > 0$. Thus for $y \in B(x,c)$ and $n \geq i_0$,

$$E_y[\exp\{-n\,F(L^n)\}]$$
$$\geq \frac{c}{k}\exp\{-2kM\}\int_{\Phi^{(i_0)}} E_z[\exp\{-n\,F(L^n)\}]\,\beta(dz)$$
$$\geq \frac{c}{k}\exp\{-2kM\}\beta(\Phi^{(i_0)})\exp\left\{-n\left(\inf_{\mu\in\mathscr{P}(S)}[F(\mu)+I(\mu)]+\varepsilon\right)\right\},$$

and therefore

$$\limsup_{n\to\infty}\sup_{y\in B(x,c)} -\frac{1}{n}\log E_y[\exp\{-n\,F(L^n)\}] \leq \inf_{\mu\in\mathscr{P}(S)}[F(\mu)+I(\mu)]+\varepsilon.$$

The inequality in (6.30) now follows by sending ε to 0. □

6.10 Noncompact State Space

The LDP already proved in this chapter (Theorem 6.6) is established under the assumption that the state space S of the Markov chain is compact. As we discuss in this section, this assumption can be relaxed to a suitable stability condition on the Markov chain given in terms of an appropriate Lyapunov function.

Recall that Proposition 6.15 did not require S to be compact. The proof of the lower bound used only the Feller property (Condition 6.2), the transitivity condition (Condition 6.3), and the tightness of $\{EL^n\}$ (Condition 6.4). However, the proof of the upper bound (Proposition 6.13) and the proof of compactness of level sets (part (d) of Lemma 6.10) make use of the compactness assumption in an important way. For both these results, compactness of the state space automatically implied the tightness of certain collections of (random) measures; for the upper bound, the collection $\{\lambda^n\}$ defined by (6.5), and for level set compactness, the collection $\{\mu \in \mathscr{P}(S) : I(\mu) \leq M\}$. We now show how a condition formulated in terms of a Lyapunov function implies these key tightness properties. Once the tightness is established, the rest of the proof for the upper bound and the compactness of level sets follows from the same arguments as in Sects. 6.7 and 6.5. A uniform Laplace principle over initial conditions in any compact set can also then be established as in Sect. 6.9.

Condition 6.19 *There exists a measurable function $U : S \to [0, \infty]$ such that the following properties hold.*

(a) $\inf_{x \in S} \left[U(x) - \log \int_S e^{U(y)} p(x, dy) \right] > -\infty$.
(b) *For each $M < \infty$,*

$$Z(M) \doteq \left\{ x \in S : U(x) - \log \int_S e^{U(y)} p(x, dy) \leq M \right\}$$

is a relatively compact subset of S.
(c) *For every compact $K \subset S$ there exists $C_K < \infty$ such that*

$$\sup_{x \in K} U(x) \leq C_K.$$

We refer the reader to [97, Example 8.2.3] for a natural class of models in which Condition 6.19 is satisfied. We note that parts (a) and (b) of the condition say that if

$$c(x) \doteq U(x) - \log \int_S e^{U(y)} p(x, dy), \ \ x \in S$$

and if κ denotes the infimum in part (a), then $\bar{c}(x) \doteq c(x) - \kappa$ is nonnegative and a tightness function. This property will be used in the proof of Proposition 6.20 below.

An inequality that allows the use of Condition 6.19 in proving tightness comes from the Donsker–Varadhan variational formula [part (a) of Lemma 2.4], according to which for every $k \in \mathbb{N}$ and $\mu_1, \mu_2 \in \mathscr{P}(S)$,

$$\int_S U_k(y) \mu_2(dy) - \log \int_S e^{U_k(y)} \mu_1(dy) \leq R(\mu_2 \| \mu_1), \tag{6.31}$$

where $U_k \doteq U \wedge k$. This inequality will be key in the tightness arguments to follow.

We next argue the tightness of the sequence $\{\lambda^n\}$ that appears on the right side of (6.11). We will assume that the sequence $\{x_n\}$ is contained in some compact $K \subset S$. Since F is bounded, we can restrict to sequences $\{\lambda^n\}$ that satisfy

$$\sup_n E_{x_n}\left[R\left(\lambda^n \middle\| \bar{L}^n \otimes p\right)\right] \le M_F \doteq 2\|F\|_\infty + 1 < \infty. \tag{6.32}$$

This is true, in particular, for the sequence in (6.11). The following is the key tightness result needed for the proof of the upper bound.

Proposition 6.20 *Suppose that Condition 6.19 and (6.32) hold. Then $\{\lambda^n\}$ is a tight sequence of $\mathscr{P}(S \times S)$-valued random variables.*

Proof It suffices to argue that $\{[\lambda^n]_i\}$ is tight in $\mathscr{P}(S)$ for $i = 1, 2$. Also note that $[\lambda^n]_1 = \bar{L}^n$, and since $\bar{\mu}_j^n$ gives the conditional distribution of $\bar{X}_j^n$, we have $\|E[\lambda^n]_1 - E[\lambda^n]_2\|_{\mathrm{TV}} \le 2/n$. Thus in view of Theorem 2.11, it suffices to argue the tightness of $\{\bar{L}^n\}$.

Next note that for $k < \infty$ and $j = 0, 1, \ldots, n-1$,

$$E_{x_n}[U_k(\bar{X}_{j+1}^n) \mid \mathscr{F}_j^n] = \int_S U_k(y)\bar{\mu}_{j+1}^n(dy).$$

It follows that

$$\begin{aligned}
E_{x_n}\left[U_k(\bar{X}_{j+1}^n) - U_k(\bar{X}_j^n)\right] &= E_{x_n}\left[\int_S U_k(y)\mu_{j+1}^n(dy) - U_k(\bar{X}_j^n)\right] \\
&= E_{x_n}\left[\int_S U_k(y)\mu_{j+1}^n(dy) - \log\int_S e^{U_k(y)}p(\bar{X}_j^n, dy)\right] \\
&\quad + E_{x_n}\left[\log\int_S e^{U_k(y)}p(\bar{X}_j^n, dy) - U_k(\bar{X}_j^n)\right] \\
&\le E_{x_n}\left[R\left(\mu_{j+1}^n(\cdot)\|p(\bar{X}_j^n, \cdot)\right)\right] - E_{x_n}\left[c_k(\bar{X}_j^n)\right],
\end{aligned}$$

where

$$c_k(x) \doteq U_k(x) - \log\int_S e^{U_k(y)}p(x, dy), \quad x \in S,$$

and the last inequality follows from (6.31). Summing the inequality over $j \in \{0, 1, \ldots, n-1\}$ gives

$$E_{x_n}\left[U_k(\bar{X}_n^n) - U_k(x_n)\right] \le E_{x_n}\left[\sum_{j=0}^{n-1} R\left(\mu_{j+1}^n\|p(\bar{X}_j^n, \cdot)\right)\right] - E_{x_n}\left[\sum_{j=0}^{n-1} c_k(\bar{X}_j^n)\right].$$

Noting that $c_k \geq \kappa \wedge 0$, applying Fatou's lemma, and recalling that $U \geq U_k \geq 0$, we have

$$\begin{aligned} E_{x_n}\left[\int_S c(x)\bar{L}^n(dx)\right] &= E_{x_n}\left[\frac{1}{n}\sum_{j=0}^{n-1} c(\bar{X}_j^n)\right] \\ &\leq E_{x_n}\left[\frac{1}{n}\sum_{j=0}^{n-1} R\left(\mu_{j+1}^n \| p(\bar{X}_j^n, \cdot)\right)\right] + \frac{1}{n}U(x_n) \\ &\leq M_F + \frac{1}{n}U(x_n), \end{aligned} \tag{6.33}$$

where the last line is from (6.32). Recall that $\bar{c} \doteq c - \kappa$ is a tightness function. From part (c) of Condition 6.19, $\sup_n U(x_n) < \infty$, and therefore

$$\sup_{n\in\mathbb{N}} E_{x_n}\left[\int_S \bar{c}(x)\bar{L}^n(dx)\right] \leq M_F + \sup_{n\in\mathbb{N}} \frac{1}{n}U(x_n) - \kappa < \infty.$$

Recalling that $\bar{c}$ is a tightness function on S, we have that $\{E_{x_n}\bar{L}^n\}$ is tight. By Theorem 2.11, the tightness of $\{\bar{L}^n\}$ follows, completing the proof of the proposition. □

Finally, we show that Condition 6.19 also gives the relative compactness of the level sets of I.

Proposition 6.21 *Suppose that Condition 6.19 holds. Then for every $M < \infty$, the level set $\{\mu \in \mathscr{P}(S) : I(\mu) \leq M\}$ is relatively compact.*

Proof Let $\mu \in \mathscr{P}(S)$ satisfy $I(\mu) \leq M$. From Lemma 6.8 there exist a $\gamma \in A(\mu)$ and a transition probability kernel $q(x, dy)$ such that μ is invariant for $q(x, dy)$ and

$$I(\mu) = R(\gamma \| \mu \otimes p) = \int_S R(q(x, \cdot) \| p(x, \cdot))\, \mu(dx).$$

Applying (6.31), we have that for every $k \in \mathbb{N}$ and $x \in S$,

$$\int_S U_k(y)q(x, dy) - \log \int_S e^{U_k(y)} p(x, dy) \leq R(q(x, \cdot) \| p(x, \cdot)). \tag{6.34}$$

Since μ is invariant for $q(x, dy)$, it follows that

$$\begin{aligned} \int_S c_k(x)\mu(dx) &= \int_S U_k(x)\mu(dx) - \int_S \left(\log \int_S e^{U_k(y)} p(x, dy)\right)\mu(dx) \\ &= \int_{S\times S} U_k(y)q(x, dy)\mu(dx) - \int_S \left(\log \int_S e^{U_k(y)} p(x, dy)\right)\mu(dx) \\ &= \int_S \left(\int_S U_k(y)q(x, dy) - \log \int_S e^{U_k(y)} p(x, dy)\right)\mu(dx) \end{aligned}$$

$$\leq \int_S R\left(q(x,\cdot)\|\, p(x,\cdot)\right)\mu(dx)$$
$$= I(\mu),$$

where the last inequality is from (6.34). Applying Fatou's lemma, we see that

$$\int_S \bar{c}(x)\mu(dx) \leq I(\mu) - \kappa \leq M - \kappa < \infty.$$

The result now follows on recalling that $\bar{c}$ is a tightness function on S. □

As discussed previously, Propositions 6.20 and 6.21 allow the weakening of compactness of S in Theorem 6.6 to Condition 6.19. More precisely, the following result holds, the proof of which follows, using the same arguments as in Sects. 6.5–6.9. We note that under Condition 6.19, the collection $\{E_{x_n} L^n\}$ is tight. Indeed, one can choose $\mu_j^n(\cdot) = p(\bar{X}_{j-1}^n, \cdot)$ in (6.33), in which case $\bar{L}^n = L^n$ and (6.32) applies. Then the tightness of $\{E_{x_n} L^n\}$ follows from Proposition 6.20.

Theorem 6.22 *Assume Conditions 6.2, 6.3, and 6.19 on the Markov chain $\{X_i\}_{i\in\mathbb{N}_0}$. Then the empirical measure sequence $\{L^n\}_{n\in\mathbb{N}}$ defined by (6.1) satisfies an LDP with rate function I defined by (6.8). Suppose in addition that for each $x \in S$, there are $\beta \in \mathscr{P}(S)$, $k \in \mathbb{N}$, and $c > 0$ such that for all $\zeta \in S$ satisfying $d(\zeta, x) < c$ and all $A \in \mathscr{B}(S)$, (6.9) is satisfied. Then the Laplace principle holds uniformly over the initial condition x_0 in compact sets in the sense of Definition 1.11.*

6.11 Notes

The pioneering work on large deviations for the empirical measures of a Markov chain is that of Donsker and Varadhan [87, 88] and Gärtner [144]. Our approach to this problem uses the same weak convergence analysis as [97, Chap. 8], which in turn adapted many ideas from [88] (a distinction between the present work and [97] is the derivation of the representation). Related works include [59, 72, 73, 80, 82, 84, 118, 162, 204, 205]. All of these authors who prove the large deviation upper bound save [118] assume that the transition probability function $p(x, dy)$ of the Markov chain satisfies the Feller property, as we do in our Theorem 6.6. The Feller property stipulates that the function mapping $x \in S \mapsto p(x,\cdot)$ is continuous in the weak topology on $\mathscr{P}(S)$. If one can guarantee that the discontinuity points of this mapping are in an appropriate sense very unlikely to be visited often, then one can weaken the Feller condition and still prove the large deviation principle with the same rate function as in Theorem 6.6 (see [97, Sect. 9.2]). However, in general, many unusual behaviors are possible when the Feller property is dropped, and examples and a discussion on this issue can be found in [118].

Representations for continuous time processes can be used to give a direct proof of the analogous results for the empirical measure [106], and interesting new behaviors

can occur for the case of pure jump processes in continuous time owing to the relatively heavy tails of the exponentially distributed holding times [107]. Large deviation theorems for multivariate empirical measures and for infinite dimensional extensions known as empirical processes have been proved by a number of authors including [29, 73, 80, 82, 89, 122, 123]. Since Brownian motion and reflected Brownian motion do not satisfy a strong recurrence property, the rate function for its empirical measure should reflect the possibility that mass may tend to infinity. The paper [40] elucidates these issues.

Chapter 7
Models with Special Features

7.1 Introduction

Chapters 4 through 6 considered small noise large deviations of stochastic recursive equations, small noise moderate deviations for processes of the same type, and large deviations for the empirical measure of a Markov chain. These chapters thus consider models that are both standard and fairly general for each setting. In this chapter we consider discrete time models that are somewhat less standard, with the aim being to show how the weak convergence methodology can be adapted. The examples presented are just for illustrative purposes, and processes featuring other challenges are referenced at the end of the chapter.

We first consider occupancy models, which were originally introduced as simplified models for problems from physics. There are other interesting problems, such as the "coupon collector's problem" [165], that can be formulated in terms of occupancy models. In principle these problems can be treated using combinatorics. However, when the number of objects (e.g., distinct coupons) is large, combinatorial methods become numerically difficult, and large deviation approximations and related numerical methods can be more tractable. One can reformulate many occupancy problems in terms of Markov models of the type considered in Chap. 4, but owing to the fact that certain transition probabilities can be arbitrarily small the processes do not satisfy the conditions of that chapter. As will be discussed, the large deviation upper bound can be proved using essentially the same argument as in Chap. 4, but the lower bound requires a more careful analysis near points where the transition probabilities vanish [see Sect. 7.2.4]. A positive feature of these models is that for many occupancy-type problems one can solve to a fairly explicit degree for the optimal trajectories in variational problems that result from a large deviations analysis, and one can also construct explicit solutions to the related partial differential equations [see Sect. 7.2.5]. These, in turn, can be used to construct subsolutions for the accelerated Monte Carlo schemes discussed in Chaps. 14–17.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_7

The second class of models, discussed in Sect. 7.3, are discrete time recursive Markov models with two time scales. Such models and their continuous time counterparts occur in many applications, such as stochastic approximations [182] and chemical reaction networks [4]. Owing to a time scale separation, the large deviation properties of empirical measures are relevant, and these models can be analyzed using a combination of the arguments used for the small noise model of Chap. 4 and those applied in Chap. 6.

7.2 Occupancy Models

Occupancy problems center on the distribution of r balls that have been thrown into n urns. In the simplest scenario each ball is equally likely to fall in any of the urns, i.e., each ball is independently assigned to a given urn with probability $1/n$. In this case, we say that the urn model uses *Maxwell-Boltzmann* (MB) statistics. This model has been studied for decades and applied in diverse fields such as computer science, biology, and statistics. See [53, 154, 165] and the references therein. However, balls may also enter the urns in a nonuniform way. An important generalization is to allow the likelihood that the ball lands in a given urn to depend on its contents prior to the throw, as in *Bose-Einstein* (BE) and *Fermi-Dirac* (FD) statistics [129, 165, 219].

For MB statistics, many results have been obtained using "exact" methods. For example, combinatorial methods are used in [130] and methods that use generating functions are discussed in [165]. Although they do not directly involve approximations, the implementation of these methods can be difficult. For example, in combinatorial methods one has to deal with the difference of events using the inclusion-exclusion formula and the resulting computations can involve large errors. In the moment generating function approach in [165] similar difficulties occur. Large deviations approximations can give a useful alternative to both of these approaches. As we have discussed previously for other models, using large deviation theory one can often obtain useful qualitative insights. This is particularly true for problems of occupancy type, since in many cases variational problems involving the rate function can be solved explicitly.

In this section we consider a parametric family of models, of which the previously mentioned MB, BE and FD statistics are all special cases. We assume there are n urns and that $\lfloor Tn \rfloor$ balls are thrown into them (where $\lfloor s \rfloor$ denotes the integer part of s), and analyze the asymptotic properties as n goes to ∞. (In contrast with previous chapters we do not simplify notation by considering just the case $T = 1$. The reason for this is because there can be a link between the parameter that characterizes the particular statistics of the model and a limit on corresponding number of balls that may be thrown, which can constrain the value of T away from 1.) A typical problem of interest is to characterize the large deviation asymptotics of the empirical distribution after all the balls are thrown. For example, one may wish to estimate the probability that at most half of the urns are empty after all the balls are thrown. A direct analysis of this problem is hard, and instead we lift the problem to the process level and analyze the large deviation asymptotics at this process level.

Although we formulate occupancy models in terms of a stochastic recursive equation of the same general type as considered in Chap. 4, there are several interesting features, both qualitative and technical, that distinguish occupancy models from the processes studied in Chap. 4. The most significant of these as far as the proof is concerned are certain vanishing transition probabilities. A second very interesting feature which was commented on previously is that one can explicitly solve the variational problems that arise in the process level approximations. Such explicit formulas have many uses and add significantly to the practical value of the large deviation approximations.

In Sect. 7.2.1 the parametric family of occupancy problem is described in detail. A dynamical characterization of the occupancy model is given, and the representation for exponential integrals is stated. In Sect. 7.2.2 we prove the lower bound for the variational problem, which corresponds to the large deviation upper bound. Section 7.2.3 analyzes the rate function I, and proves properties that will allow us to deal with the technical difficulties of the vanishing transition probabilities. In Sect. 7.2.4, we prove the variational upper bound which corresponds to the large deviation lower bound. Finally, in Sect. 7.2.5 we give an explicit formula for the minimum of the rate function subject to a terminal constraint.

7.2.1 Preliminaries and Main Result

In this section, we formulate the problem of interest and state the LDP. The proof is given in sections that follow.

The general occupancy problem has the same structure as the MB occupancy problem, except that in the general problem urns are distinguished according to the number of balls contained therein. The full collection of models will be indexed by a parameter a. This parameter takes values in the set $(0, \infty] \cup \{-1, -2, \ldots\}$, and its interpretation is as follows. An urn is said to be of *category* i if it contains i balls. A ball is thrown in any given urn with probability proportional to $a + i$, where i denotes the category of the urn. In particular, suppose that a ball is about to be thrown, and that any two urns (labeled say A and B) are selected. Suppose that urn A is of category i, while B is of category j. Then the probability that the ball is thrown into urn A, conditioned on the state of all the urns and that the ball is thrown into either urn A or B, is

$$\frac{a+i}{(a+i)+(a+j)}.$$

When $a = \infty$ we interpret this to mean that the two urns are equally likely. Also, when $a < 0$ we use this ratio to define the probabilities only when $0 \le i \vee j \le -a$ and $i < -a$ or $j < -a$, so the formula gives a well defined probability. The probability that a ball is placed in an urn of category $-a$ is 0. Thus under this model, urns can only be of category $0, 1, \ldots, -a$, and we only throw balls into categories $0, 1, \ldots, -a-1$. Note that the case $a = 0$ is in some sense not interesting, in that as soon as there is an urn of category $j > 0$ all balls will be placed in that urn. Likewise the cases $a < 0$ but not an integer are hard to interpret.

In this setup, certain special cases are distinguished. The cases $a = 1$, $a = \infty$, $a \in -\mathbb{N}$ correspond to Bose-Einstein statistics, Maxwell-Boltzmann statistics, and Fermi-Dirac statistics, respectively.

Suppose that before we throw a ball there are already tn balls in all the urns, and further suppose that the occupancy state is $(x_0, x_1, \ldots, x_{J+})$. Here $x_i, i = 0, 1, \ldots, J$ denotes the fraction of urns that contain i balls, and x_{J+} denotes the fraction containing more than J balls. When $a < 0$, we will take $J = -a - 1$. The "un-normalized" or "relative" probability of throwing into a category i urn with $i \leq J$ is simply $(a + i)x_i$. Let us temporarily abuse notation, and let $x_{J+1}, x_{J+2}, \ldots$ denote the exact fraction in each category i with $i > J$. Since there are tn balls in the urns before we throw, $\sum_{i=0}^{\infty} i x_i = t$. Thus the (normalized and true) probability that the ball is placed in an urn that contains exactly i balls, $i = 0, 1, \ldots, J$, is $(a + i)x_i/(a + t)$, and the probability that the ball is placed in an urn that has more than J balls is $1 - \sum_{j=0}^{J}[(a + j)/(a + t)]x_j$.

An explicit construction of this process is as follows. To simplify, we assume the empty initial condition, i.e., all urns are empty. One can consider other initial conditions, with only simple notational changes in the results to be stated below. We introduce a time variable t that ranges from 0 to T. At a time t that is of the form l/n, with $0 \leq l \leq \lfloor nT \rfloor$ an integer, l balls have been thrown. Let $X^n(t) = (X_0^n(t), X_1^n(t), \ldots X_J^n(t), X_{J+}^n(t))'$ be the **occupancy state** at that time. As noted previously, $X_i^n(t)$ denotes the fraction of urns that contain i balls at time t, $i = 0, 1, \ldots, J$, and $X_{J+}^n(t)$ the fraction of urns that contain more than J balls. As usual, the definition of X^n is extended to all $t \in [0, T]$ not of the form l/n by piecewise linear interpolation. Note that for each t $X^n(t)$ is a probability vector in $\mathbb{R}^{J+2}$. Denoting $\Lambda \doteq \{0, 1, \ldots, J + 1\}$ and with an abuse of notation

$$\mathscr{P}(\Lambda) \doteq \left\{ x \in \mathbb{R}^{J+2} : x_i \geq 0, 0 \leq i \leq J + 1 \text{ and } \sum_{i=0}^{J+1} x_i = 1 \right\},$$

then for any $t \in [0, T]$, $X^n(t) \in \mathscr{P}(\Lambda)$. Thus X^n takes values in $\mathscr{C}([0, T] : \mathscr{P}(\Lambda))$. We equip $\mathscr{C}([0, T] : \mathscr{P}(\Lambda))$ with the usual supremum norm and on $\mathscr{P}(\Lambda)$ we take the $\mathscr{L}^1$-norm, which will be denoted by $\|\cdot\|_1$.

It will be convenient to work with the following dynamical representation. For $x \in \mathbb{R}^{J+2}$ and $t \in [0, -a1_{\{a<0\}} + \infty 1_{\{a>0\}})$ define the vector $\rho(t, x) \in \mathbb{R}^{J+2}$ by

$$\rho_k(t, x) = \frac{a + k}{a + t} x_k, \qquad \text{for } k = 0, 1, \ldots, J, \tag{7.1}$$

$$\rho_{J+1}(t, x) = 1 - \sum_{k=0}^{J} \frac{a + k}{a + t} x_k,$$

where, as before, when $a = \infty$ the fraction $(a + k)/(a + t)$ is taken to be 1. Then $\rho_\cdot(x, t)$ will play a role analogous to that of $\theta(\cdot|x)$ in Chap. 4 in identifying the conditional distribution of the increment of the process. Differences are that here

there is time dependence, and also that the increment is identified by (but not equal to) the k index in $\rho_k(t,x)$ [see (7.3)]. A straightforward calculation shows that if

$$x \in \mathscr{P}(\Lambda) \quad \text{and} \quad \sum_{k=0}^{J} k x_k \le t, \tag{7.2}$$

then $\rho_{J+1}(t,x) \ge 0$ and $\rho(t,x)$ is therefore a probability vector in $\mathbb{R}^{J+2}$, i.e., $\rho(t,x) \in \mathscr{P}(\Lambda)$. Also $\rho(t,x)$ is Lipschitz continuous in $(t,x) \in [0,T] \times \mathscr{P}(\Lambda)$, as long as $T < -a$ when $a \in -\mathbb{N}$. We then construct a family of independent random functions

$$\left\{ \nu_i^n(\cdot) : i = 0, 1, \ldots, \lfloor nT \rfloor - 1, \lfloor nT \rfloor \right\}$$

that take values in

$$\Lambda \doteq \{0, 1, \ldots, J+1\}$$

and with distributions

$$P\left\{ \nu_i^n(x) = k \right\} = \rho_k\left(i/n, x\right), \quad k \in \Lambda. \tag{7.3}$$

The mapping that takes an index $k \in \Lambda$ into a change in the occupancy numbers is

$$\gamma[k] = e_{k+1} - e_k, \quad 0 \le k \le K, \quad \gamma[J+1] = 0, \tag{7.4}$$

where for $j = 0, 1, \ldots, J+1$, e_j denotes the unit vector in $\mathbb{R}^{J+2}$ with 1 in the $j+1$th coordinate. Finally, we define X_l^n recursively by $X_0^n = (1, 0, \ldots, 0)' = e_0$ and

$$X_{i+1}^n = X_i^n + \frac{1}{n} \gamma[\nu_i^n\left(X_i^n\right)].$$

For the continuous time interpolation let $X^n(i/n) = X_i^n$, and for t not of the form i/n define $X^n(t)$ by piecewise linear interpolation. Observe that the conditional distribution of the increment $\left\{ \nu_i^n\left(X_i^n\right) \right\}$ is determined by $\rho\left(i/n, X_i^n\right)$. Thus the process X^n at the discrete times i/n is Markovian and will have the same distribution as the occupancy process described previously.

Define the $J+2$ by $J+2$ matrix

$$M \doteq \begin{pmatrix} -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{pmatrix}.$$

Let $\varphi \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$ be given with $\varphi_0(0) = 1$. Suppose there is a Borel measurable function $\theta : [0,T] \to \mathscr{P}(\Lambda)$ such that for all $t \in [0,T]$

$$\varphi(t) = \varphi(0) + \int_0^t M\theta(s)ds. \tag{7.5}$$

Note that $M\theta = \sum_{k=0}^{K} \gamma[k]\theta_k$ if $\theta \in \mathscr{P}(\Lambda)$. For $i = 0, 1, \ldots J$ (resp. $i = J+1$) we interpret $\theta_i(s)$ as the rate at which balls are thrown into urns that contain i balls (resp. greater than J balls) at time s. The rates $\theta(s)$ are unique in the sense that if another $\tilde{\theta} : [0,T] \to \mathscr{P}(\Lambda)$ satisfies (7.5) then $\tilde{\theta} = \theta$ a.e. on $[0,T]$. We call φ a **valid occupancy state process** if there exists $\theta : [0,T] \to \mathscr{P}(\Lambda)$ satisfying (7.5). In this case θ is called the **occupancy rate process** associated with φ. Using the observation that $\sum_{k=1}^{J+2}(k-1)M_{kj} = 1$ for all $j = 1, \ldots, J+1$, it is easy to check that if φ is valid then $\varphi(s)$ satisfies (7.2) with x replaced by $\varphi(s)$ and t by s, for all $s \in [0,T]$. This shows that $\rho(s, \varphi(s)) \in \mathscr{P}(\Lambda)$.

When two probability vectors θ and $\nu \in \mathscr{P}(\Lambda)$ appear in the relative entropy function, we interpret them as probability measures on $\{0, 1, \ldots, J, J+1\}$, and thus

$$R(\theta \,\|\, \nu) \doteq \sum_{k=0}^{J+1} \theta_k \log \frac{\theta_k}{\nu_k}.$$

As observed before, when φ is valid, $\rho(s, \varphi(s)) \in \mathscr{P}(\Lambda)$, so $R(\theta(s) \,\|\, \rho(s, \varphi(s)))$ is well defined for all $s \in [0,T]$. For such φ define

$$I(\varphi) \doteq \int_0^T R(\theta(s) \,\|\, \rho(s, \varphi(s)))ds. \tag{7.6}$$

If φ is not valid then define $I(\varphi) = \infty$.

As usual, representation formulas for exponential integrals will be used to prove the Laplace principle. The representation needed here is a special case of the one proved in Chap. 4, and we therefore just state the form of the representation. The controlled process $\bar{X}^n(t)$ is constructed as follows. The conditional distributions of controlled random integers $\{\bar{\nu}_i^n\}$ will be specified by a sequence $\{\bar{\mu}_i^n\}$ of controls. Each $\bar{\mu}_i^n$ is measurable with respect to the σ-algebra generated by $\{\bar{\nu}_j^n\}_{0,1,\ldots,i-1}$, and identifies the conditional distribution of $\bar{\nu}_i^n$. The controlled process is determined for t of the form j/n by $\bar{X}_0^n = e_0$ and

$$\bar{X}_{i+1}^n = \bar{X}_i^n + \frac{1}{n}\gamma[\bar{\nu}_i^n] \qquad \text{for } i = 0, 1, \ldots, \lfloor nT \rfloor,$$

with γ as in (7.4). The random quantities $\bar{X}_i^n$ and $\bar{\nu}_i^n$ are defined recursively in the order

$$\bar{X}_0^n, \bar{\nu}_0^n, \bar{X}_1^n, \bar{\nu}_1^n, \bar{X}_2^n, \ldots, \bar{X}_{\lfloor nT \rfloor + 1}^n,$$

We set $\bar{X}^n(i/n) = \bar{X}_i^n$ and use piecewise linear interpolation elsewhere.

Define

$$r_i^n(\{k\}) \doteq \rho_k\left(i/n, \bar{X}_i^n\right),$$

where $\rho\,(t, x)$ is given in (7.1).

Remark 7.1 As noted previously, there is an abuse of notation, in that we sometimes think of r_i^n as the probability vector with components $r_i^n(k)$ but at other times as the probability measure with values $r_i^n(\{k\})$. To reinforce the fact that certain probability measures are on the discrete set Λ, we write such measures with the differential dk. Also, note that k will appear both as a subscript, as in $\rho_k\left(i/n, \bar{X}_i^n\right)$, and as an argument, as in $r_i^n(\{k\})$.

Let L^n, $\bar{L}^n$, $\bar{\mu}^n$ and λ^n be measures on $\{0, 1, \ldots, J+1\} \times [0, T]$ defined as in Construction 4.4, except that λ^n uses $\rho.\left(i/n, \bar{X}_i^n\right)$ in place of $\theta(\cdot|\bar{X}_i^n)$, and the measures are of mass T and defined on subsets of $[0, T]$ rather than $[0, 1]$ in the second marginal. Specifically, for $A \subset \{0, 1, \ldots, J+1\}$ and $B \in \mathscr{B}([0, T])$,

$$L^n(A \times B) \doteq \int_B L^n(A|t)dt, \quad \bar{L}^n(A \times B) \doteq \int_B \bar{L}^n(A|t)dt, \tag{7.7}$$

$$\bar{\mu}^n(A \times B) \doteq \int_B \bar{\mu}^n(A|t)dt, \quad \lambda^n(A \times B) \doteq \int_B \lambda^n(A|t)dt, \tag{7.8}$$

where for $t \in [i/n, i/n + 1/n)$, $i = 0, \ldots, \lfloor nT \rfloor$

$$\begin{array}{ll} L^n(A|t) \doteq \delta_{\nu_i^n(X_i^n)}(A), & \bar{L}^n(A|t) \doteq \delta_{\bar{\nu}_i^n}(A), \\ \bar{\mu}^n(A|t) \doteq \bar{\mu}_i^n(A), & \lambda^n(A|t) \doteq r_i^n(A) \end{array} \tag{7.9}$$

The random measures L^n, $\bar{L}^n$, $\bar{\mu}^n$ and λ^n take values in the collection of nonnegative measures on $\mathscr{P}(\Lambda) \times [0, T]$ of total mass T. The topology used is the weak topology, where these measures are renormalized to have mass one, i.e., probability measures, and since $\mathscr{P}(\Lambda) \times [0, T]$ is compact this space is compact as well. If G is any bounded measurable function the space to $\mathbb{R}$, then

$$-\frac{1}{n}\log E\exp\left[-nG(L^n)\right] = \inf_{\{\bar{\mu}_i^n\}} E\left[G\left(\bar{L}^n\right) + \frac{1}{n}\sum_{i=0}^{\lfloor nT \rfloor} R\left(\bar{\mu}_i^n \,\|\, r_i^n\right)\right], \tag{7.10}$$

where the infimum is over all the admissible control sequences $\left\{\bar{\mu}_i^n\right\}$. Since

$$X^n(t) = e_0 + \int_0^t \gamma[k]L^n(dk \times ds), \tag{7.11}$$

this also gives a representation for functions of X^n: for any bounded and continuous $F : \mathscr{C}\left([0, T] : \mathscr{P}(\Lambda)\right) \to \mathbb{R}$,

$$-\frac{1}{n}\log E\exp\left[-nF(X^n)\right]=\inf_{\{\bar{\mu}_i^n\}}E\left[F(\bar{X}^n)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar{\mu}_i^n\,\|r_i^n\right)\right]. \tag{7.12}$$

Here we have used the fact that (7.11) defines a measurable map that takes L^n to X^n, and let $\bar{X}^n$ be the image of $\bar{L}^n$ under that map.

A convention for the case $a \in -\mathbb{N}$. When $a \in -\mathbb{N}$ it is only possible to throw balls into the categories $0, 1, \ldots, -a-1$, and the only possible categories are $0, 1, \ldots, -a$. Thus if there are n urns there can at most be $-an$ balls thrown, and therefore $T \le -a$. When $T = -a$ all the urns have exactly $-a$ balls, which is not an interesting case to study. As a consequence, throughout this chapter we assume $T < -a$. Also, as was noted previously, because of the restriction on the possible categories we (without loss) assume that $J = -a-1$. Thus throughout this section for $a < 0$ we assume

$$T < -a, \quad J = -a-1. \tag{7.13}$$

7.2.2 Laplace Upper Bound

In this section, we prove the variational lower bound

$$\liminf_{n\to\infty}-\frac{1}{n}\log E\exp\left[-nF(X^n)\right]\ge\inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))}\left[I(\varphi)+F(\varphi)\right],$$

which corresponds to the Laplace upper bound. By (7.12) it is enough to show that

$$\liminf_{n\to\infty}\inf_{\{\bar{\mu}_i^n\}}E\left[F\left(\bar{X}^n\right)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar{\mu}_i^n\,\|r_i^n\right)\right]\ge\inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))}\left[I(\varphi)+F(\varphi)\right].$$

The upper bound is actually covered by the analysis of Chap. 4, since the occupancy model satisfies Condition 4.3 if one appends time as a state variable. However, for completeness we include the (short) proof here.

Recall the definitions in (7.8) and (7.9). Note that because relative entropy is nonnegative and $(\lfloor nT\rfloor+1)/n \ge T$,

$$\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar{\mu}_i^n\,\|r_i^n\right)\ge\int_0^T R\left(\bar{\mu}^n(\cdot|t)\,\|\lambda^n(\cdot|t)\right)dt. \tag{7.14}$$

As usual, we will need to understand conditions for tightness, and how the weak limits of $\bar{L}^n$, $\bar{\mu}^n$, λ^n and $\bar{X}^n$ are all related. As noted previously, tightness of the first three is automatic since they take values in a compact space. In addition, the process $\bar{X}^n$ takes values in a space of continuous trajectories that start at e_0 and which are Lipschitz continuous with the Lipschitz constant bounded by

$$\left\| \sup_{0\le t\le T} \int \gamma[k]\bar{L}^n(dk|t) \right\|_1 \le \sup_{k\in\Lambda} \|\gamma[k]\|_1 \le 2,$$

where $\|\cdot\|_1$ is the $\mathscr{L}^1$-norm on $\mathbb{R}^{J+2}$. Since the space of all such trajectories is also compact, $\{\bar{X}^n\}$ is also automatically tight. The relations between the limits can be determined using the same argument as in Lemma 4.12, save that $\rho_\cdot(t,x)$, which plays the role of $\theta(\cdot|x)$ in Chap. 4, is here time dependent, and whereas the dynamics of Chap. 4 take the form (4.1), if we consider $\rho_\cdot(t,x)$ as determining the noises that drive the system then these noises enter the system only after passing through $\gamma[\cdot]$ as in (7.11).

Rewritten for these differences, the analogue of Lemma 4.12 is as follows.

Lemma 7.2 *Consider the sequence $\{(\bar{X}^n, \bar{L}^n, \bar{\mu}^n, \lambda^n)\}_{n\in\mathbb{N}}$ as defined in (7.7) and (7.8), and with $\bar{X}^n$ as in (7.11) with L^n replaced by $\bar{L}^n$. Let $\{(\bar{X}^n, \bar{L}^n, \bar{\mu}^n, \lambda^n)\}$ denote a weakly converging subsequence, which for notational convenience we again label by n, with limit $(\bar{X}, \bar{L}, \bar{\mu}, \lambda)$. Then w.p.1 $\bar{L} = \bar{\mu}$, and $\bar{\mu}(dk \times dt)$ can be decomposed as $\bar{\mu}(dk|t)dt$, where $\bar{\mu}(dk|t)$ is a stochastic kernel on $\{0, 1, \ldots, J, J+1\}$ given $[0, T]$, and w.p.1 for all $t \in [0, T]$,*

$$\begin{aligned} \bar{X}(t) &= e_0 + \int_{\mathbb{R}^d\times[0,t]} \gamma[k]\bar{\mu}(dk \times ds) \\ &= e_0 + \int_{\mathbb{R}^d\times[0,t]} \gamma[k]\bar{\mu}(dk|s)ds. \end{aligned} \tag{7.15}$$

In addition, λ and $\bar{X}$ are related through

$$\lambda(\{k\} \times B) = \int_B \rho_k(t, \bar{X}(t))dt, \ k \in \{0, 1, \ldots, J+1\}, \ B \in \mathscr{B}([0, T]). \tag{7.16}$$

Theorem 7.3 *Define I by (7.6) for any of the occupancy models described in Sect. 7.2.1. If $F : \mathscr{C}([0, T] : \mathscr{P}(\Lambda)) \to \mathbb{R}$ is bounded and continuous, then*

$$\liminf_{n\to\infty} -\frac{1}{n} \log E \exp\left[-nF(X^n)\right] \ge \inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))} [I(\varphi) + F(\varphi)].$$

Proof Owing to the representation formula (7.10) it is enough to show that

$$\liminf_{n\to\infty} \inf_{\{\bar{\mu}_i^n\}} E\left[F\left(\bar{X}^n\right) + \frac{1}{n} \sum_{i=0}^{\lfloor nT\rfloor} R\left(\bar{\mu}_i^n \,\middle\|\, r_i^n\right) \right] \ge \inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))} [I(\varphi) + F(\varphi)]. \tag{7.17}$$

Consider any admissible sequence $\left\{\bar{\mu}_i^n\right\}$. Then (7.14) implies

$$E\left[F\left(\bar{X}^n\right)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar{\mu}_i^n\,\|r_i^n\right)\right]\geq E\left[F\left(\bar{X}^n\right)+\int_0^T R\left(\bar{\mu}^n(\cdot\,|t)\,\|\lambda^n(\cdot|t)\right)dt\right]$$
$$=E\left[F\left(\bar{X}^n\right)+TR\left(\bar{\mu}^n/T\,\|\lambda^n/T\right)\right].$$

Applying Fatou's lemma and using the lower semicontinuity of relative entropy,

$$\begin{aligned}&\liminf_{n\to\infty}E\left[F\left(\bar{X}^n\right)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar{\mu}_i^n\,\|r_i^n\right)\right]\\&\quad\geq\liminf_{n\to\infty}E\left[F\left(\bar{X}^n\right)+TR\left(\bar{\mu}^n/T\,\|\lambda^n/T\right)\right]\\&\quad\geq E\left[F\left(\bar{X}\right)+TR\left(\bar{\mu}/T\,\|\lambda/T\right)\right]\qquad(7.18)\\&\quad=E\left[F\left(\bar{X}\right)+\int_0^T R\left(\bar{\mu}(\cdot\,|t)\,\|\rho_\cdot(t,\bar{X}(t))\right)dt\right].\end{aligned}$$

If $\theta(s)=\sum_{k=0}^{J+1}e_k\bar{\mu}(\{k\}\,|s)$, then using $M\theta(s)=\sum_{k=0}^{J}\gamma[k]\theta_k(s)$ we see from Lemma 7.2 that $\bar{X}(t)=e_0+\int_0^t M\theta(s)ds$. Therefore by the definition (7.6) of the rate function $I(\varphi)$,

$$\int_0^T R\left(\bar{\mu}(\cdot\,|t)\,\|\rho_\cdot(t,\bar{X}(t))\right)dt=I(\bar{X}).$$

Thus (7.18) yields (7.17), and completes the proof of the Laplace upper bound. □

7.2.3 *Properties of the Rate Function*

In this section we prove important properties of the rate function, some of which will be used later on to prove the Laplace lower bound.

Theorem 7.4 *Let I be defined as in (7.6). Then for any $K\in[0,\infty)$ the level set $\{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda)):I(\varphi)\leq K\}$ is compact.*

Proof By adding time as a state variable we see that the occupancy model satisfies Condition 4.3 of Chap. 4. Thus the conclusion follows from Theorem 4.13.

Theorem 7.5 (ZERO COST TRAJECTORY) *For $t\in[0,T]$ let $f(t)\doteq\left(1+\frac{t}{a}\right)^{-a}$ when $a<\infty$ and $f(t)\doteq e^{-t}$ in the case $a=\infty$. Define*

$$\bar{\phi}_i(t)\doteq\frac{(-t)^i}{i!}f^{(i)}(t)\qquad\text{for } 0\leq i\leq J,$$

where $f^{(i)}(t)$ is the ith derivative of $f(t)$, and let $\bar{\phi}_{J+1}(t)\doteq 1-\sum_{i=0}^{J}\bar{\phi}_i(t)$. Then $I(\bar{\phi})=0$.

Proof We first assume $a \neq \infty$. It is easy to check that for any $0 \leq i < \infty$,

$$\frac{(-t)^i}{i!} f^{(i)}(t) \geq 0 \quad \text{and} \quad \sum_{i=0}^{\infty} \frac{(-t)^i}{i!} f^{(i)}(t) = 1. \tag{7.19}$$

Thus $\bar{\phi}$ as defined in the statement of the theorem is indeed a probability vector. It is also a continuously differentiable function and satisfies $\sum_{k=0}^{J} \bar{\phi}_k(t) \leq t$ for all $t \in [0, T]$. We will show that

$$\frac{d}{dt}\bar{\phi}(t) = M\rho(t, \bar{\phi}(t)). \tag{7.20}$$

If so, then the occupancy rate process $\bar{\theta}$ associated to $\bar{\phi}$ is $\rho(t, \bar{\phi}(t))$, and thus by the definition of rate function

$$I(\bar{\phi}) = \int_0^T R\left(\bar{\theta}(t) \,\|\, \rho(t, \bar{\phi}(t))\right) dt = 0.$$

To show (7.20) we calculate $\bar{\phi}_i(t) = \frac{(-t)^i}{i!} f^{(i)}(t)$ for $0 \leq i \leq J$ explicitly:

$$\bar{\phi}_i(t) = \frac{t^i}{i!} \frac{\prod_{j=0}^{i-1}(a+j)}{a^i} \left(1 + \frac{t}{a}\right)^{-a-i}.$$

Hence the derivative satisfies

$$\begin{aligned} \frac{d}{dt}\bar{\phi}_i(t) &= \frac{a+i-1}{a+t}\bar{\phi}_{i-1}(t) - \frac{a+i}{a+t}\bar{\phi}_i(t) \\ &= \rho_{i-1}(t, \bar{\phi}(t)) - \rho_i(t, \bar{\phi}(t)) = (M\rho(t, \bar{\phi}(t))_i, \end{aligned}$$

where ρ_{-1} is taken to be 0 and the second equality is due to the definition of $\rho(t, \bar{\phi}(t))$ in (7.1). The case of $\phi_{J+1}(t)$ follows on observing that

$$\frac{d}{dt}\bar{\phi}_{J+1}(t) = -\sum_{i=0}^{J} \frac{d}{dt}\bar{\phi}_i(t) = -\sum_{i=0}^{J} (M\rho(t, \bar{\phi}(t))_i = (M\rho(t, \bar{\phi}(t))_{J+1},$$

where the last identity is a consequence of the fact that $1^T M = 0$.

Next we consider the case when $a = \infty$. In this case $f(t) = e^{-t}$, and (7.20) is immediate on observing that $\bar{\phi}_i(t) = t^i e^{-t}/i!$ and so $\frac{d}{dt}\bar{\phi}_i(t) = \bar{\phi}_{i-1}(t) - \bar{\phi}_i(t)$. □

Lemma 7.6 *Let $\bar{\phi}$ be the zero-cost trajectory from Theorem 7.5. For every choice of the parameter a there exist $\delta > 0$ and $0 < K < \infty$ so that*

$$\bar{\phi}_i(t) \geq \delta t^K \tag{7.21}$$

for any $0 \leq i \leq J + 1$ and $t \in [0, T]$.

Proof Note that when $a > 0$, $0 \le i \le J$ and $0 \le t \le T$,

$$\bar{\phi}_i(t) = \frac{t^i}{i!} \frac{\prod_{j=0}^{i-1}(a+j)}{a^i} \left(1 + \frac{t}{a}\right)^{-a-i} \ge \frac{t^i}{J!} \left(1 + \frac{T}{a}\right)^{-a-J},$$

and because of (7.19) we have

$$\bar{\phi}_{J+1}(t) = 1 - \sum_{i=0}^{J} \bar{\phi}_i(t) \ge \frac{(-t)^{J+1}}{(J+1)!} f^{(J+1)}(t) \ge \frac{t^{J+1}}{(J+1)!} \left(1 + \frac{T}{a}\right)^{-a-J-1}.$$

Thus in this case, with $\bar{\delta} = \frac{1}{(J+1)!} \left(1 + \frac{T}{a}\right)^{-a-J-1}$,

$$\bar{\phi}_i(t) \ge \bar{\delta} t^i, \text{ for all } i = 0, 1, \ldots, J+1, \ t \in [0, T]. \tag{7.22}$$

For the case $a < 0$, by (7.13) $T < -a$ and $a = -J - 1$. If $0 \le i \le J$, then since $a + j \le -1$ for each $0 \le j \le J$,

$$\bar{\phi}_i(t) = \frac{t^i}{i!} \frac{\prod_{j=0}^{i-1}(a+j)}{a^i} \left(1 + \frac{t}{a}\right)^{-a-i} \ge \frac{t^i}{J!} \frac{1}{(-a)^i} \left(1 + \frac{t}{a}\right)^{-a-i}.$$

Moreover since $a < 0$, $t/a \in (-1, 0)$ for $t \in [0, T]$, and $-(a+i) \ge 1$, for each $i \le J$, $\left(1 + \frac{t}{a}\right)^{-a-i}$ is monotone decreasing in $t \in [0, T]$. Therefore

$$\bar{\phi}_i(t) \ge \frac{t^i}{J!} \left(-\frac{1}{a}\right)^i \left(1 + \frac{T}{a}\right)^{-a-i}.$$

For $\bar{\phi}_{J+1}(t)$ we have

$$\begin{aligned}
\bar{\phi}_{J+1}(t) &= 1 - \sum_{i=0}^{J} \bar{\phi}_i(t) \\
&\ge \frac{(-t)^{J+1}}{(J+1)!} f^{(J+1)}(t) \\
&= \frac{t^{J+1}}{(J+1)!} \frac{\prod_{j=0}^{J}(a+j)}{a^{J+1}} \\
&\ge \frac{t^{J+1}}{(J+1)!} \left(-\frac{1}{a}\right)^{J+1}.
\end{aligned}$$

Thus in this case (7.22) holds with $\bar{\delta} = (-a)^{-J-1}/(J+1)!$.

Finally, for the case $a = \infty$, using the fact that $\bar{\phi}_i(t) = t^i e^{-t}/i!$ for $i \le J$ and $\bar{\phi}_{J+1}(t) \ge t^{J+1}e^{-t}/(J+1)!$, we have that (7.22) holds with $\bar{\delta} = e^{-T}/(J+1)!$. The result now follows on taking $K = J+1$ and $\delta = \bar{\delta}(T^{-J-1} \wedge 1)$. □

For $f : [0,T] \to \mathbb{R}^{J+2}$, let $\|f\|_{\infty,T} \doteq \sup_{0\le t\le T} \|f(t)\|_1$, where $\|\cdot\|_1$ as before is the $\mathscr{L}^1$ norm on $\mathbb{R}^{J+2}$.

Lemma 7.7 *For a given value of a let the parameters δ and K be as in (7.21). Let $\varphi \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$ satisfy $I(\varphi) < \infty$. Then for any $\varepsilon > 0$ there exists $\varphi^\varepsilon \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$ such that*

(a) $I(\varphi^\varepsilon) \le I(\varphi)$,
(b) $\|\varphi - \varphi^\varepsilon\|_{\infty,T} \le \varepsilon$,
(c) $\varphi_i^\varepsilon(t) \ge \varepsilon\delta t^K$ *for all* $t \in [0,T]$ *and* $i = 0, 1, \ldots J, J+1$.

Proof For any $\varepsilon > 0$ and $\varphi \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$, let

$$\varphi^\varepsilon = (1-\varepsilon)\,\varphi + \varepsilon\bar{\varphi},$$

where $\bar{\varphi}$ is the zero cost trajectory from Theorem 7.5. Then $\varphi^\varepsilon \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$. From the definition of $\rho(t,x)$ in (7.1) it follows that $\rho(t,x)$ has the following linearity property in x. Suppose we are given $t \in [0,T]$ and $x, \tilde{x} \in \mathscr{P}(\Lambda)$ that satisfy (7.2). Then for any $\alpha \in [0,1]$, $\alpha x + (1-\alpha)\tilde{x}$ satisfies (7.2) and

$$\alpha\rho(t,x) + (1-\alpha)\rho(t,\tilde{x}) = \rho(t, \alpha x + (1-\alpha)\tilde{x}).$$

Hence recalling the definition of $I(\varphi)$ in (7.6) and the joint convexity of relative entropy, we find that $I(\varphi)$ is convex in φ. Therefore

$$I(\varphi^\varepsilon) \le (1-\varepsilon)\,I(\varphi) + \varepsilon I(\bar{\varphi}) = (1-\varepsilon)\,I(\varphi) \le I(\varphi).$$

Since $\|\varphi - \bar{\varphi}\|_{\infty,T} \le 2$

$$\|\varphi - \varphi^\varepsilon\|_{\infty,T} \le \varepsilon\,\|\varphi - \bar{\varphi}\|_{\infty,T} \le 2\varepsilon,$$

and also from Lemma 7.6, $\varphi_i^\varepsilon(t) \ge \varepsilon\bar{\varphi}_i(t) \ge \varepsilon\delta t^K$. □

The final theorem of this section is useful in proving the Laplace lower bound.

Definition 7.8 We call an occupancy path $\varphi \in \mathscr{C}([0,T] : \mathscr{P}(\Lambda))$ a **good path** if $\varphi(0) = e_0$ and there exist constants $0 < \delta', K' < \infty$ so that $\varphi_i(t) \ge \delta' t^{K'}$ for $t \in [0,T]$ and $0 \le i \le J+1$.

Definition 7.9 We call an occupancy rate control $\theta : [0,T] \to \mathscr{P}(\Lambda)$ a **good control** if (i) there exist a finite number of intervals $[r_i, s_i]$, $1 \le i \le m$ so that $[0,T] = \cup_{i=1}^m [r_i, s_i]$, and $\theta(t)$ is a constant vector on each (r_i, s_i), (ii) there exists $0 < \sigma < T$ so that θ is "pure" on $[0,\sigma)$, in the sense that for any interval of constancy $(r,s) \subset [0,\sigma)$, there exists i, $0 \le i \le J+1$ such that $\theta_i(t) = 1$ for $t \in (r,s)$.

Theorem 7.10 *For a good path $\varphi \in \mathscr{C}([0,T]:\mathscr{P}(\Lambda))$ assume $I(\varphi) < \infty$. Let δ', K' be the associated constants in the definition of a good path. For any $\varepsilon > 0$ there exists a good control θ^* and associated $\sigma > 0$ so that if φ^* is the occupancy path associated to θ^*, namely (7.5) holds with (φ,θ) replaced by (φ^*,θ^*), then there is $\delta'' \in (0,\infty)$ such that*

(a) $I(\varphi^*) \le I(\varphi) + \varepsilon$,
(b) $\|\varphi^* - \varphi\|_{\infty,T} \le \varepsilon$,
(c) *if $t < \sigma$ and $\theta_i^*(t) = 1$ then $\varphi_i^*(t) \ge \delta'' \sigma^{K'}$.*

Proof For a $\sigma \in (0,T)$ that will be specified later on, we construct a pure control $\theta^*(t)$, $t \in [0,\sigma)$ as follows. For $0 \le i \le J$ let $\theta_i^*(t) = 1$ if

$$\sum_{j=0}^{i} j\varphi_j(\sigma) + i\sum_{k=i+1}^{J+1}\varphi_k(\sigma) \le t < \sum_{j=0}^{i} j\varphi_j(\sigma) + (i+1)\sum_{k=i+1}^{J+1}\varphi_k(\sigma),$$

and let $\theta_{J+1}^*(t) = 1$ if

$$\sum_{j=0}^{J} j\varphi_j(\sigma) + (J+1)\varphi_{J+1}(\sigma) \le t < \sigma. \tag{7.23}$$

Observe that the component φ_i^* for $i > 0$ will increase only during the interval when $\theta_{i-1}^*(t) = 1$, and that it decreases to its final value while $\theta_i^*(t) = 1$. Observe also that $\varphi^*(\sigma) = \varphi(\sigma)$. Hence for $t < \sigma$, if $\theta_i^*(t) = 1$ then $\varphi_i^*(t) \ge \varphi_i^*(\sigma) \ge \delta'\sigma^{K'}$.

Now assume that $0 < a < \infty$. For i and t such that $t < \sigma$ and $\theta_i^*(t) = 1$,

$$\rho_i(t,\varphi^*(t)) = \frac{a+i}{a+t}\varphi_i^*(t) \ge \frac{a}{a+T}\delta'\sigma^{K'} = \delta''\sigma^{K'}, \tag{7.24}$$

where $\delta'' \doteq \frac{a}{a+T}\delta'$.

Recall that when $a < 0$ we assume without loss that $J = -a-1$, and that no balls are placed in urns that currently contain more than J balls. Thus $\rho_{J+1}(t,\phi(t)) = 0$ and consequently $\theta_{J+1}(t) = 0$ for all t. From (7.5) and recalling that $\sum_{j=0}^{J+1} jM_{(j+1),i} = 1$ for all $i = 1,\dots,J+1$, it follows that

$$\sum_{j=0}^{J+1} j\varphi_j(\sigma) = \sigma.$$

It then follows from (7.23) that $\theta_{J+1}^*(t) = 0$ for all $t \in [0,\sigma]$. For $0 \le i \le J$, we have, when $t < \sigma$ and $\theta_i^*(t) = 1$, that

$$\rho_i(t,\varphi^*(t)) \ge \frac{a+i}{a+t}\delta'\sigma^{K'} \ge \frac{a+J}{a+t}\delta'\sigma^{K'} \ge -\frac{1}{a}\delta'\sigma^{K'}.$$

Thus for $0 \le i \le J$, with $\delta'' = -a^{-1}\delta'$, $\rho_i(t, \varphi^*(t)) \ge \delta'' \sigma^{K'}$ when $\theta_i^*(t) = 1$ and $t \in [0, \sigma]$.

Finally, when $a = \infty$ we can choose $\delta'' = \delta'$ and (7.24) will hold. Thus in all cases (7.24) holds with some $\delta'' > 0$, that is independent of the choice of σ.

This completes the construction of θ^* and φ^* on $[0, \sigma)$. The lower bounds on the ρ_i and the fact that θ^* is pure on $[0, \sigma]$ imply

$$\int_0^\sigma R\left(\theta^*(t) \,\|\, \rho(t, \varphi^*(t))\right) dt \le -\sigma \log\left(\delta'' \sigma^{K'}\right).$$

Now choose $\sigma > 0$ small enough so that

$$-\sigma \log\left(\delta'' \sigma^{K'}\right) \le \varepsilon/2 \text{ and } \sup_{t \in [0,\sigma)} \left\|\varphi^*(t) - \varphi(t)\right\|_1 \le \varepsilon.$$

Note that the latter property can be satisfied by choosing σ sufficiently small since $\varphi(0) = \varphi^*(0) = e_0$ implies $\sup_{t\in[0,\sigma]} \|\varphi(t) - \varphi^*(t)\|_1 \le (J+1)|\varphi_0(0) - \varphi_0(\sigma)|$. Also, recall that under the construction $\varphi^*(\sigma) = \varphi(\sigma)$.

The construction of controls on $[\sigma, T]$ is easier. Let $\theta(t)$ be the rate process associated with $\varphi(t)$ by (7.5). For $N \in \mathbb{N}$ we partition $[\sigma, T]$ into N subintervals of length $c_N = (T - \sigma)/N$. For each s that $\sigma + l c_N \le s \le \sigma + (l+1) c_N$ where $0 \le l \le (N-1)$, let

$$\theta^{(N)}(s) = \frac{\int_{\sigma + l c_N}^{\sigma + (l+1) c_N} \theta(t) dt}{c_N}.$$

Let $\varphi^{(N)}$ be the occupancy path associated with $\theta^{(N)}$ over the interval $[\sigma, T]$, i.e.

$$\varphi^{(N)}(t) = \varphi^{(N)}(\sigma) + \int_\sigma^t M\theta^{(N)}(s) ds, \; t \in [\sigma, T], \; \varphi^{(N)}(\sigma) = \varphi(\sigma).$$

Then it is easy to check that $\varphi^{(N)}(t)$ coincides with $\varphi(t)$ on the "partition points" in $[\sigma, T]$, i.e., those points of the form $\{\sigma + l c_N : 0 \le l \le (N-1)\}$. Thus, since $\|\theta(t)\|_1 = 1$, for N large enough [e.g., $N > (T - \sigma)/\varepsilon$], $\sup_{t\in[\sigma,T]} \|\varphi^{(N)}(t) - \varphi(t)\|_1 \le \varepsilon$.

Because $\varphi(t)$ is good, when $t > \sigma$, we have $\varphi_i(t) \ge \delta' t^{K'} \ge \delta' \sigma^{K'} > 0$ for all $0 \le i \le J+1$. Therefore $\varphi(t)$ is uniformly bounded away from the boundary after time σ, and thus for sufficiently large N, so is $\varphi^{(N)}(t)$. This in particular says that for such N, $t \in [\sigma, T]$, $\rho_j(t, \varphi^{(N)}(t))$ is uniformly bounded away from 0 for $j = 0, \ldots, J+1$ when $a > 0$ and for $j = 0, \ldots, J$ when $a < 0$. In the latter case, both $\rho_{J+1}(t, \varphi^{(N)}(t))$ and $\theta_{J+1}^{(N)}(t)$ are identically 0.

As $N \to \infty$, $\theta^{(N)}(t)$ converges to $\theta(t)$ and $\varphi^{(N)}(t)$ converges to $\varphi(t)$ for a.e. $t \in [0, T]$. Using that ρ is bounded away from zero and $\theta^{(N)}$ is bounded above, by the dominated convergence theorem

$$\lim_{N\to\infty}\int_\sigma^T R(\theta^{(N)}(t)\,\|\rho(t,\varphi^{(N)}(t)))dt=\int_\sigma^T R\left(\theta(t)\,\|\rho(t,\varphi(t))\right)dt.$$

Now choose $N<\infty$ large enough so that the integrals differ by less than $\varepsilon/2$. Let θ^* be defined as it was previously on $[0,\sigma]$, and set it equal to $\theta^{(N)}$ on $[\sigma,T]$. Let φ^* denote the corresponding occupancy path over $[0,T]$. Then

$$\begin{aligned}I(\varphi^*)&=\int_\sigma^T R(\theta^{(N)}(t)\,\|\rho(t,\varphi^{(N)}(t)))dt+\int_0^\sigma R\left(\theta^*(t)\,\|\rho(t,\varphi^*(t))\right)dt\\&\le\int_\sigma^T R\left(\theta(t)\,\|\rho(t,\varphi(t))\right)dt+\varepsilon/2+\varepsilon/2\\&\le I(\varphi)+\varepsilon.\end{aligned}$$

This completes the proof. □

7.2.4 *Laplace Lower Bound*

Theorem 7.11 *Define I by (7.6) for any of the occupancy models described in Sect. 7.2.1. If $F:\mathscr{C}([0,T]:\mathscr{P}(\Lambda))\to\mathbb{R}$ is bounded and continuous, then*

$$\limsup_{n\to\infty}-\frac{1}{n}\log E\exp\left[-nF(X^n)\right]\le\inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))}\left[I(\varphi)+F(\varphi)\right].$$

Proof According to (7.12), the theorem follows if

$$\limsup_{n\to\infty}\inf_{\{\bar\mu_i^n\}}\bar E\left[F\left(\bar X^n\right)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar\mu_i^n\,\|r_i^n\right)\right]\le\inf_{\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))}\left[I(\varphi)+F(\varphi)\right].$$

As was the case with Chap. 4, the main difficulty in the proof of the lower bound is that controls and controlled processes should be constructed so that the dominated convergence theorem can be used. Since vanishing transition probabilities can make relative entropy costs diverge some care is required, but the constructions of the last section will very carefully control the rates at which balls are put into urns of category i when r_i^n is small.

For any $\varphi\in\mathscr{C}([0,T]:\mathscr{P}(\Lambda))$ such that $I(\varphi)<\infty$, Lemma 7.7 and Theorem 7.10 imply that for any $\varepsilon>0$ there exists (φ^*,θ^*) with the properties described in Theorem 7.10. Since F is continuous on $\mathscr{C}([0,T]:\mathscr{P}(\Lambda))$, we only need to show that there exists a sequence of admissible controls $\left\{\bar\mu_i^n\right\}$ so that

$$\limsup_{n\to\infty}\bar E\left[F\left(\bar X^n\right)+\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor}R\left(\bar\mu_i^n\,\|r_i^n\right)\right]\le I(\varphi^*)+F(\varphi^*).$$

The latter inequality will follow if we can find a sequence of admissible $\left\{\bar{\mu}_i^n\right\}$ such that

$$\limsup_{n\to\infty} \bar{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor nT\rfloor} R\left(\bar{\mu}_i^n \,\|\, r_i^n\right)\right] \le I(\varphi^*), \tag{7.25}$$

and such that if $\bar{X}^n$ is the occupancy process constructed under $\left\{\bar{\mu}_i^n\right\}$ then for any small $b > 0$

$$\limsup_{n\to\infty} \bar{P}\left\{\left\|\bar{X}^n - \varphi^*\right\|_{\infty,T} > b\right\} = 0. \tag{7.26}$$

In other words, $\bar{X}^n$ converges to φ^* in probability.

To prove the desired inequalities (7.25) and (7.26) we need to construct the proper $\left\{\bar{\mu}_i^n\right\}$. Recall that $\left\{\bar{\mu}_i^n\right\}$ can depend in any measurable way on the "past," and so we could, in principle, use such information in constructing the controls. However, as seen previously for certain problems of this type we can construct the controls without reference to the controlled process (i.e., "open loop" controls). Let θ^* be the good control as described in Theorem 7.10. We know that θ^* is piecewise constant and pure up to time $\sigma > 0$. From property (c) in Theorem 7.10, we also know that before time σ, if $\theta_i^*(t) = 1$ then both $\rho_i(t, \varphi^*(t))$ and $\varphi_i^*(t)$ are greater than a fixed value $\zeta > 0$ (for all $i \le J+1$ when $a > 0$ and for all $i \le J$ when $a < 0$). Using part (b) of Theorem 7.10 we can also assume for the same value of ζ that both $\rho_i(t, \varphi^*(t))$ and $\varphi_i^*(t)$ are greater than ζ for all $t \in [\sigma, T]$ (and again for all $i \le J+1$ when $a > 0$ and for all $i \le J$ when $a < 0$).

Although the limit trajectory stays away from the boundary after time σ, there is no guarantee that the random process $\bar{X}^n$ is uniformly bounded away. In order to handle this possibility, we use a stopping time argument similar to one used in [109].

Let (l_n/n) be the minimum of the first time such that for some i, $\bar{X}_i^n(l_n/n) \le \zeta/2$ and $\theta_i^*(l_n/n) > 0$, and the fixed deterministic time $\lfloor nT \rfloor / n$. This is the first time the random process is close to the boundary, with the possibility of a large contribution to the total cost [note that when $\theta_i^*(l_n/n) = 0$ there is no contribution to the cost regardless of the value of $\bar{X}_i^n(l_n/n)$]. The control $\left\{\bar{\mu}_i^n\right\}$ is then defined by

$$\bar{\mu}_i^n(\{k\}) = \begin{cases} \theta_k^*(i/n) & \text{if } i \le l_n \\ \rho_k\left(i/n, \bar{X}^n(i/n)\right) & \text{if } i > l_n. \end{cases}$$

Prior to the stopping time, we use exactly what θ^* suggests, and after the stopping time we follow the law of large number trajectory (and therefore incur no additional cost).

Now we apply Lemma 7.2. Thus given any subsequence there is convergence along a further subsequence as indicated in the theorem, with limit $\left(\bar{X}, \bar{\mu}\right)$. Using the standard argument by contradiction, it will be enough to prove (7.25) and (7.26) for this convergent processes. Let $\tau^n = (l_n/n) \le T$. Note that because the applied controls are pure, the process $\bar{X}^n(t)$ is deterministic prior to σ, and also that prior to this time, the time derivatives of $\bar{X}^n(t)$ and $\varphi^*(t)$ are piecewise constant. In fact,

the two derivatives are identical except possibly on a bounded number of intervals each of length less than $1/n$ [the points where they may disagree are all located with within distance $1/n$ of the endpoints of the intervals of constancy of $\dot{\varphi}^*(t)$]. Thus for large n we cannot have $\tau^n < \sigma$. Since the range of τ^n is the bounded set $[0, T]$, we can also assume τ^n converges in distribution to a limit τ, and without loss we assume the convergence is along the same subsequence. Since $\tau^n \geq \sigma$ for large n we have $\tau \geq \sigma$ w.p.1.

Suppose that $\tau < T$. It is easy to check that the limit control processes w.p.1 satisfies, for a.e. $t \in [0, T]$,

$$\bar{\mu}(\{k\} \,|\, t) = \begin{cases} \theta_k^*(t) & \text{if } t \leq \tau \\ \rho_k\left(t, \bar{X}(t)\right) & \text{if } t > \tau \end{cases}.$$

Owing to the definition of τ^n, $\tau < T$ implies $\bar{X}_i(\tau) \leq \zeta/2$ for some $i \in \Lambda$ (although $\varphi_i^*(t) \geq \zeta$ when $t \in [\sigma, T]$). We use that $\bar{\mu}(\{k\} \,|t) = \theta_k^*(t)$ when $t \leq \tau$ and that $\theta^*(t)$ is deterministic. As shown in Theorem 7.2, $\left(\bar{X}, \bar{\mu}\right)$ satisfies (7.15) for $t \in [0, \tau]$. Thus for $t \in [0, \tau]$, $\bar{X}(t) = \varphi^*(t)$ w.p.1. This gives a contradiction since

$$\bar{X}_i(\tau) \leq \zeta/2 < \zeta \leq \varphi_i^*(\tau).$$

Therefore $\tau = T$, and thus for all $t \in [0, T]$, $\bar{X}(t) = \varphi^*(t)$ w.p.1. This also proves that the weak limit of the random processes $\bar{X}^n$ is indeed φ^*, which implies (7.26). To prove (7.25), we use the weak convergence, the continuity of the map $(x, y) \mapsto x \log(x/y)$ on $[0, \infty) \times (0, \infty)$ and the dominated convergence theorem to obtain

$$\limsup_{n\to\infty} \bar{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor nT \rfloor} R\left(\bar{\mu}_i^n \,\|r_i^n\right)\right] = \int_0^T R\left(\theta^*(t)||\rho(t, \varphi^*(t))\right) dt = I(\varphi^*).$$

This completes the proof. □

7.2.5 *Solution to Calculus of Variations Problems*

In the previous sections we identified the process level large deviation rate function (7.6) for a class of occupancy problems. The large deviation principle for the process at a given fixed time can then be expressed in terms of the solution to a calculus of variations problem. In most cases this calculus of variations problem will not have a closed form solution. However, for the class of occupancy models studied here it can be identified with the solution to a related finite dimensional minimization problem. This latter problem can be solved by the standard Lagrange multiplier method, which is easily implemented numerically. In this section we give the precise statement of this equivalence. We mention two results. The first gives the minimum of the rate function subject to a terminal constraint, and the second gives the minimum of the

sum of the rate function plus a cost that is affine in the terminal location. The explicit formulas generalize ones obtained in [109] for the special case of MB statistics. The techniques used are quite different, based as they are on dynamic programming and control theory rather than methods from the calculus of variations. When combined with methods for accelerated Monte Carlo as discuss in later chapters, these explicit solutions allow one to obtain not just large deviation approximations but also accurate approximations to nonasymptotic quantities. Proofs are not given, but interested readers can find the details in [266].

7.2.5.1 Problem Formulation

Suppose the current occupancy state is $x \in \mathscr{P}(\Lambda)$ and that t is the number of balls per urn among all categories. If $y_i, i = 0, 1, \ldots, J, J+1, \ldots$ are the fraction in category i, then $x_i = y_i$ for $i \leq J$ and

$$t = \sum_{k=0}^{\infty} k y_k.$$

Note that $t \geq \sum_{k=0}^{J+1} k x_k$.

In previous sections we considered the large deviation analysis for just the case of the initial condition where all urns are empty. To use dynamic programming, one must introduce the analogue of the rate function that is suitable for general initial times and states. The set of possible states for a given t [i.e., ones that can be reached starting from $(1, 0, \ldots, 0)$ at $t = 0$] depends on both t and a, which leads to the following definition.

Definition 7.12 Define $\mathscr{D}_a$, the **feasible domain** for the occupancy model with parameter a, as follows:

- when $a > 0$,

$$\mathscr{D}_a \doteq \left\{ (x,t) \in \mathscr{P}(\Lambda) \times [0,T) : x_{J+1} > 0 \text{ and } t \geq \sum_{i=0}^{J+1} i x_i \right\} \bigcup \left\{ (x,t) \in \mathscr{P}(\Lambda) \times [0,T) : x_{J+1} = 0 \text{ and } t = \sum_{i=0}^{J} i x_i \right\};$$

- and when $a < 0$ and $J = -a - 1$,

$$\mathscr{D}_a \doteq \left\{ (x,t) \in \mathscr{P}(\Lambda) \times [0,T) : t = \sum_{i=0}^{J+1} i x_i \right\}.$$

As before, when $a < 0$ we restrict to $T < -a$. In the first case the second set in the union reflects the fact that when $x_{J+1} = 0$ the number of balls thrown is exactly $\sum_{i=0}^{J} i x_i$, and similarly for the second case.

Consider a valid occupancy process $\varphi \in \mathscr{C}([t, T] : \mathscr{P}(\Lambda))$ with $\varphi(t) = x$ and $(x, t) \in \mathscr{D}_a$. Making the dependence on (x, t) explicit, the rate function $I(x, t; \varphi)$ for such paths can be written

$$I(x, t; \varphi) \doteq \int_t^T R(\theta(s) \| \rho(s, \varphi(s))) \, ds,$$

where

$$\varphi(s) = \varphi(t) + \int_t^s M\theta(r) dr$$

and

$$\rho_k(s, y) \doteq \frac{a + k}{a + s} y_k, \qquad k = 0, 1, \ldots, J, \qquad \rho_{J+1}(s, y) \doteq 1 - \sum_{k=0}^{J} \rho_k(s, y).$$

The relevant calculus of variations problem for a point in the feasible domain is

$$O(x, t; \omega) \doteq \inf_{\substack{\varphi \in \mathscr{C}([t,T]:\mathscr{P}(\Lambda)) \\ \varphi(t)=x, \varphi(T)=\omega}} I(x, t; \varphi). \tag{7.27}$$

The formula for the finite dimensional minimization problem requires some notation. For all $a \in \mathbb{R}$, $a \neq 0$ and $i \in \mathbb{N}$, let

$$\binom{a}{i} \doteq \frac{\prod_{j=0}^{i-1}(a - j)}{i!}$$

and $\binom{a}{0} = 1$. Note that if $a \in \mathbb{N}$ and $i > a$ then $\binom{a}{i} = 0$, and that if $a \notin \mathbb{N} \cup \{0\}$, then $\binom{a}{i} \neq 0$. We will use the fact that if $a \in \mathbb{R}$ and $|z| < 1$ then the binomial expansion

$$(1 + z)^{-a} = \sum_{i=0}^{\infty} \binom{-a}{i} z^i$$

is valid, and if $-a \in \mathbb{N}$ then the sum contains only a finite number of nonzero terms and is valid for all $z \in \mathbb{R}$.

For $i \in \mathbb{N} \cup \{0\}$ and $a > 0$, $s \geq 0$ or $a \in -\mathbb{N}$, $0 \leq s \leq -a$, define

$$Q_i^a(s) \doteq \left(-\frac{s}{a}\right)^i \binom{-a}{i} \left(1 + \frac{s}{a}\right)^{-a-i}.$$

When $a = 0$ we use the limiting values

$$Q_0^0(s) = 1, \;\; Q_i^0(s) = 0$$

for all $i \in \mathbb{N}$ and $s \geq 0$. One can check that $\{Q_i^a(s)\}_{i=0}^{\infty}$ is a probability vector for any choice of (a, s) as above.

Denote $\pi^k = \{\pi_0^k, \pi_1^k, \dots\}$ for all $0 \leq k \leq J+1$, where π_i^k represents the probability of throwing i additional balls into the kth category. Denote $\pi = (\pi^0, \pi^1, \dots, \pi^{J+1})$. For any given $x \in \mathscr{P}(\Lambda)$, we say $\pi = (\pi^0, \pi^1, \dots, \pi^J, \pi^{J+1}) \in \mathscr{F}(x, t; \omega, T)$ if

$$\sum_{j=0}^{\infty} \pi_j^k = 1, \quad 0 \leq k \leq J+1, \quad \sum_{k=0}^{J+1} x_k \sum_{j=0}^{\infty} j\pi_j^k = T - t, \tag{7.28}$$

and

$$\omega_i = \sum_{k=0}^{i} x_k \pi_{i-k}^k, \quad 0 \leq i \leq J, \qquad \omega_{J+1} = 1 - \sum_{k=0}^{J} \omega_k. \tag{7.29}$$

We will use $\omega \doteq x \times \pi$ as shorthand for the last display. Roughly speaking, if $\{x_k\}_{k=0}^{J+1}$ is the occupancy state at time instant t and π_i^k represents the probability of throwing i additional balls over the interval $[nt, nT]$ into the kth category, then ω_i gives the average fraction of category i urns at time nT.

A terminal point ω is **feasible** (for the given initial time and condition) if $\mathscr{F}(x, t; \omega, T)$ is not empty.

Now we are ready to state the theorem. For $s > 0$ let $P(s)$ denote the Poisson distribution with parameter s, and if $s = 0$ let $P(s)$ denote the probability measure on $\{0, 1, \dots\}$ with mass one on $\{0\}$. The proof of the representation can be found in [266].

Theorem 7.13 (Explicit Formula for the Rate Function) *Consider an initial condition $(x, t) \in \mathscr{D}_a$, and a feasible terminal condition ω. If $a \in (0, \infty)$, then for $x_{J+1} > 0$ let*

$$\tau(x, t) \doteq \frac{(t - \sum_{k=0}^{J} k x_k)}{x_{J+1}}$$

(so that $\tau(x, t)$ is the average number of balls per urn distributed in the $J+$ categories for the initial condition (x, t)) and if $x_{J+1} = 0$ let $\tau(x, t) = 0$. Then the quantity $O(x, t; \omega)$ defined in (7.27) has the representation

$$O(x, t; \omega) = \min_{\pi \in \mathscr{F}(x,t;\omega,T)} \left[\sum_{k=0}^{J} x_k R\left(\pi^k \,\middle\|\, Q^{a+k}\left(\frac{a+k}{a+t}(T-t)\right)\right) \right.$$
$$\left. + x_{J+1} R\left(\pi^{J+1} \,\middle\|\, Q^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t)\right)\right)\right].$$

If $a \in -\mathbb{N}$ with $J = -a - 1$ then $\tau(x, t) = J+1$, and

$$O(x,t;\omega) = \min_{\pi \in \mathscr{F}(x,t;\omega,T)} \left[\sum_{k=0}^{J+1} x_k R\left(\pi^k \left\| Q^{a+k}\left(\frac{a+k}{a+t}(T-t) \right) \right) \right].$$

In the final case of $a = \infty$, we have

$$O(x,t;\omega) = \min_{\pi \in \mathscr{F}(x,t;\omega,T)} \left[\sum_{k=0}^{J+1} x_k R\left(\pi^k \| P(T-t) \right) \right].$$

Although these minimization problems as stated appear to be infinite dimensional, they can in fact be reduced to finite dimensional problems. This is because if π^k is the minimizer, then π_j^k takes a prescribed form for $j > J$. In fact, all π_j^k can be represented in terms of no more than $J + 3$ Lagrange multipliers [119, 266].

Theorem 7.13 gives the minimal cost to move from one point in the feasible domain to another. For the construction of accelerated Monte Carlo schemes it is useful to know how to construct subsolutions to the related Hamilton-Jacobi-Bellman (HJB) equation with various terminal conditions. This can often be done by approximating general terminal conditions from below by a special class of terminal conditions, such as those involving affine costs (see the examples in Chap. 17). Such a result is stated in Proposition 7.14, and in fact Theorem 7.13 is shown to be a consequence of Proposition 7.14 by approximating the function equal to 0 when $x = \omega$ and ∞ elsewhere from below by affine functions.

7.2.5.2 The Hamilton-Jacobi-Bellman Equation

In this section we assume $a < \infty$, noting that the Maxwell-Boltzmann case ($a = \infty$) can easily be obtained as a limit. See [119, 266] for further discussion.

The calculus of variations problem (7.27) has a natural control interpretation, where $\theta(s)$ is the control, $\dot{\varphi}(s) = M\theta(s)$ are the dynamics, $R(\theta(s) \| \rho(s, \varphi(s)))$ is the running cost and $g(x) = \infty 1_{\{\omega\}^c}(x)$ is the terminal cost. It is expected that if we define

$$V(x,t) \doteq \inf_{\varphi \in \mathscr{C}([t,T]:\mathscr{P}(\Lambda)),\varphi(t)=x} \left[\int_t^T R(\theta(s)||\rho(s,\varphi(s)))\, ds + g(\varphi(T)) \right], \tag{7.30}$$

then $V(x,t)$ is a weak-sense solution [14] to the HJB equation

$$W_t(x,t) + \mathbb{H}(DW(x,t),x,t) = 0,$$

and terminal condition

$$W(x,T) = \infty 1_{\{\omega\}^c}(x).$$

Here the Hamiltonian $\mathbb{H}(p,x,t)$ is defined by

$$\mathbb{H}(p,x,t) \doteq \inf_{\theta \in \mathscr{P}(\Lambda)} [\langle p, M\theta \rangle + R(\theta \,\|\rho(t,x))]$$

and W_t and DW denote the partial derivative with respect to t and gradient in x, respectively. Note that by the representation formula Proposition 2.2, the infimum in the definition of $\mathbb{H}(p,x,t)$ can be evaluated, yielding

$$W_t(x,t) = \log\left(\sum_{k=0}^{J} x_k \left(\frac{a+k}{a+t}\right) e^{\left(W_{x_k}(x,t) - W_{x_{k+1}}(x,t)\right)} + \left(\frac{a+\tau(x,t)}{a+t}\right) x_{J+1}\right)$$

plus the terminal condition $W(x,T) = g(x)$, where $W_{x_k}(x,t)$ is the partial derivative and $\tau(x,t)$ is as in Theorem 7.13.

A class of problems that are of interest in applications are those with a terminal condition of the form

$$g(x) = \infty 1_{\mathscr{A}^c}(x),$$

where $\mathscr{A}$ is some convex set. Such terminal conditions usually yield only a weak-sense solution, and not a classical-sense C^1 solution to the HJB equation. However, as mentioned previously it is possible for the purposes of design of accelerated Monte Carlo to approximate these terminal conditions from below in terms of affine terminal conditions. In the next result we state a representation for the calculus of variations problem with affine terminal cost $g(\omega) = \langle l, \omega \rangle + b$. The representation turns out to be the unique classical sense solution to the corresponding PDE. To simplify, we first observe that W is a solution of just the PDE alone (i.e., without the terminal condition) if and only if $W + c$ is a solution for any real number c. Since ω is a probability vector, it suffices to prove the representation when $l_{J+1} = 0$ and $b = 0$. We also recall the definition (7.29).

Proposition 7.14 *Consider* $(x,t) \in \mathscr{D}_a$ *and* $g(\omega) = \langle l, \omega \rangle$, *where* $l \in \mathbb{R}^{J+2}$ *and* $l_{J+1} = 0$. *Define* V *by (7.30) and*

$$U(x,t) \doteq \min_{\pi \in \mathscr{F}(x,t;T)} \left[\sum_{k=0}^{J} x_k R\left(\pi^k \,\middle\|\, Q^{a+k}\left(\frac{a+k}{a+t}(T-t)\right)\right) + x_{J+1} R\left(\pi^{J+1} \,\middle\|\, Q^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t)\right)\right) + g(x \times \pi)\right]$$

where $\pi \in \mathscr{F}(x,t;T)$ *means that* π *satisfies the constraints in (7.28). Then* $V(x,t) = U(x,t)$.

7.3 Two Scale Recursive Markov Systems with Small Noise

In this section we consider a discrete time stochastic dynamical system in which there are two components to the state. One of the components evolves at a slower time scale then the other, and this scale separation is determined by the parameter

that also scales the size of the noise. Such systems include many models arising in queuing theory and communication systems [18, 35, 182], where they are called Markov-modulated processes.

We are interested in studying the large deviation behavior of the slow component (though one could also study the joint large deviation properties of the slow component and a time dependent empirical measure of the fast process). The main result of the section is Theorem 7.17, which establishes the LDP for the slow component. The proof, which is left to the reader, combines techniques from Chaps. 4 and 6. We begin by describing the model in precise terms.

7.3.1 Model and Assumptions

Let S be compact metric space and let $p(\xi, d\varsigma)$ be a probability transition kernel on S. We assume that the kernel satisfies the Feller and the transitivity properties from Chap. 6, namely Conditions 6.2 and 6.3. The fast component of the Markov chain will be governed by this kernel. The slow component is described through a stochastic kernel $\theta(dy|x, \xi)$ on $\mathbb{R}^d$ given $\mathbb{R}^d \times S$. We suppose as given a probability space that supports iid random vector fields $\{v_i(x, \xi), i \in \mathbb{N}_0, (x, \xi) \in \mathbb{R}^d \times S\}$, with the property that for any $(x, \xi) \in \mathbb{R}^d \times S$ $v_i(x, \xi)$ has distribution $\theta(\cdot|x, \xi)$. We also suppose as given an S-valued Markov chain $\{\Xi_i\}_{i\in\mathbb{N}_0}$ on this probability space with transition kernel $p(\xi, d\varsigma)$ and with $\Xi_0 = \xi_0 \in S$. The sequence $\{\Xi_i\}$ will be the fast component, and is independent of $\{v_i\}$. The stochastic process describing the evolution of the slow component is then given by

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n, \Xi_{i+1}), \quad X_0^n = x_0.$$

Thus $\{X_i^n\}$ is a stochastic dynamical system with small noise, though the distribution of the noise depends on both X_i^n and the modulating process Ξ_i. The evolution of X_i^n, being scaled by $1/n$, is slow relative to that of Ξ_i. As in Chap. 4 this discrete time process is interpolated into continuous time according to

$$X^n(t) = X_i^n + \left[X_{i+1}^n - X_i^n\right](nt - i), \quad t \in [i/n, i/n + 1/n].$$

We are interested in the large deviation properties of the sequence $\{X^n\}_{n\in\mathbb{N}}$ of $\mathscr{C}([0, 1] : \mathbb{R}^d)$-valued random variables.

We impose the following analogues of Conditions 4.3 and 4.7 from Chap. 4. For $(x, \xi) \in \mathbb{R}^d \times S$ and $\alpha \in \mathbb{R}^d$ define

$$H(x, \xi, \alpha) \doteq \log E e^{\langle \alpha, v_i(x,\xi)\rangle}.$$

Condition 7.15 (a) *For each* $\alpha \in \mathbb{R}^d$ $\sup_{(x,\xi)\in\mathbb{R}^d\times S} H(x, \xi, \alpha) < \infty$.

(b) *The mapping* $(x, \xi) \mapsto \theta(\cdot|x, \xi)$ *from* $\mathbb{R}^d \times S$ *to* $\mathscr{P}(\mathbb{R}^d)$ *is continuous in the topology of weak convergence.*

Condition 7.16 *For each* $(x, \xi) \in \mathbb{R}^d \times S$, *the convex hull of the support of* $\theta(\cdot|x, \xi)$ *is* $\mathbb{R}^d$.

7.3.2 Rate Function and the LDP

We next introduce the rate function for $\{X^n\}$. For $\mu \in \mathscr{P}(S)$ define $A(\mu)$ as in Sect. 6.3 [see 6.6]:

$$A(\mu) \doteq \left\{\gamma \in \mathscr{P}(S^2) : [\gamma]_1 = [\gamma]_2 = \mu\right\}.$$

Also, as in Chap. 6, given $\mu \in \mathscr{P}(S)$, let $(\mu \otimes p)(dx \times dy)$ denote the probability measure on S^2 given by $\mu(dx)p(x, dy)$. Let I_1 denote the rate function I in Theorem 6.6:

$$I_1(\mu) = \inf_{\gamma \in A(\mu)} R\left(\gamma \,\|\, \mu \otimes p\right), \ \mu \in \mathscr{P}(S).$$

Define $L : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty]$ by

$$L(x, \beta) \doteq \inf\left[\int_S R(\nu(\cdot|\xi)\|\theta(\cdot|x, \xi))\mu(d\xi) + I_1(\mu) : \int_{S\times\mathbb{R}^d} y\nu(dy|\xi)\mu(d\xi) = \beta\right],$$

where the infimum is over $\mu \in \mathscr{P}(S)$ and stochastic kernels ν on $\mathscr{P}(\mathbb{R}^d)$ given S. The definition of the local rate function involves two changes of distribution and the associated relative entropy costs. The first switches the distribution of transitions of $\{\Xi_i\}$ from $p(\xi, d\varsigma)$ to $q(\xi, d\varsigma)$, where $[\mu \otimes q]_2 = \mu$. Since X^n moves only a small distance over a small interval in continuous time, it is the invariant distribution μ of q which affects the evolution of the controlled analogue of X^n. Thus if we shift from the invariant distribution of p to μ, then we must pay a cost of $I_1(\mu)$ per unit time. Once this is done, as in Chap. 4 the distribution of the noises $v_i(x, \xi)$ can be perturbed away from $\theta(\cdot|x, \xi)$ to $\nu(\cdot|\xi)$, but one must pay a relative entropy cost. The overall cost to track a velocity β minimizes these two costs.

Recall that $\mathscr{AC}_{x_0}([0, 1] : \mathbb{R}^d)$ denotes the subset of $\mathscr{C}([0, 1] : \mathbb{R}^d)$ consisting of all absolutely continuous functions satisfying $\phi(0) = x_0$. The rate function for $\{X^n\}$ is given as follows. Let

$$I(\phi) = \int_0^1 L(\phi(s), \dot{\phi}(s))ds \quad \text{if } \phi \in \mathscr{AC}_{x_0}([0, 1] : \mathbb{R}^d),$$

and in all other cases $I(\phi) = \infty$.

The following theorem states the LDP for $\{X^n\}$.

Theorem 7.17 *Suppose that Conditions 6.2, 6.3, 7.15 and 7.16 are satisfied. Then I is a rate function and $\{X^n\}_{n\in\mathbb{N}}$ satisfies the Laplace principle on $\mathscr{C}([0,1]:\mathbb{R}^d)$ with rate function I, uniformly for initial conditions in compact sets.*

7.3.3 Extensions

We have considered the simplest form of a two scale system in discrete time, and in particular under assumptions such that a straightforward combination of the methods from Chaps. 4 and 6 can be applied to complete the proof. The model can in principle be extended in several directions, under various sets of additional assumptions. For example, as in Chap. 6 the compactness of S can be replaced by a condition on the existence of a suitable Lyapunov function. Likewise the condition on the support of the transition kernel $\theta(\cdot|x,\xi)$, Condition 7.16, can be replaced by a Lipschitz type condition of a similar form as Condition 4.8. Finally, for the model considered here the evolution of the fast component did not depend on the state of the slow variable. This condition can be relaxed to allow for a fully coupled system. See [42] for sufficient conditions in a continuous time setting and [94] for a discrete time system.

7.4 Notes

An overview of occupancy models and their applications can be found in [165]. The first paper to consider the large deviation properties of an occupancy model appears to be [109], which was motivated by the problem of sizing switches in optical communications. In [109] the LDP for the MB model is obtained, and the rate function exhibited in more-or-less explicit form. The arguments in Sect. 7.2 are based on those used in [266], though as in previous chapters the presentation here first studies the large deviation properties of an empirical measure and then obtains those for the process.

As was discussed in Sect. 7.2, the most difficult part of the analysis is in dealing with parts of the state space where rates go to zero, which produces singular behaviors in the local rate function. The are many other classes of models in applied probability where transition probabilities vanish (or in their continuous time analogues jump rates vanish), including models from queueing and related stochastic networks [231], chemical reaction networks, and random graphs [23]. A positive feature of this collection of problems (one that is emphasized in Sect. 7.2) is that the associated variational problems have explicit or nearly explicit solutions.

The main difficulties are typically in the proof of the large deviation lower bound, and the approach used in this chapter involves a careful analysis of the local rate function to construct controls that can be used to establish the lower bound. For the corresponding continuous time models, one can sometimes represent the process

as the solution to a stochastic differential equation driven by one or more Poisson random measures. In this case one might ask if the perspective of Sects. 3.1 and 3.3, which exploits the fact that the mapping from the noise model (Brownian motion or Poisson random measure) into the state variable is "nearly continuous" could be used. This turns out to be possible, as described for example in [23].

The second model of this chapter is a stochastic recursive system with two time scales. Models of this type appear in many different areas of application, and general references include [171, 259]. One of the first papers to consider the large deviation properties of processes of this general sort is Freidlin [138]. Continuous time analogues of such two time scale systems have also been well studied (see [42] and references therein). Related and very challenging problems involve systems where the averaging is with respect to an "environment" variable rather than time, e.g., a stochastic differential equation where the drift is itself random or periodic and ergodic in an appropriate sense. An example of how weak convergence methods can be used to account for such averaging in a relatively simple setting appears in [111].

Part III
Continuous Time Processes

The next part of the book is concerned with the development of useful representations for continuous time models and their application. In the setting of discrete time one could obtain very useful representations using the chain rule. This is no longer the case in continuous time, but it is possible to obtain representations of a form analogous to those in discrete time using the Radon–Nikodym Theorem on path space for one inequality and stochastic control arguments for the reverse inequality. The representations for infinite dimensional Brownian motion and Poisson random measures are the topic of Chap. 8.

For many models in continuous time, driving noises enter in a more regular fashion than for their discrete time counterparts. In particular, for stochastic ordinary and partial differential equations driven by Brownian and Poisson noise, these noises enter in an "affine" manner, which is meant to include both what is often referred to as "additive" noise (constant noise coefficient), and "multiplicative" noise (state-dependent noise coefficient). As a consequence, the mapping from the noise space into the state space often has more structure and regularity than in the corresponding discrete time setting. One can compare, for example, Markov processes described by SDEs with Markov processes of the type studied in Chap. 4. A consequence of the improved regularity is that one can identify conditions on the map that takes the noise process into the state process that are sufficient for large and moderate deviation principles to hold, and which are broadly applicable. This is the topic of Chap. 9, which proves large and moderate deviation properties for general mappings of Brownian motion and Poisson random measures.

The results of Chap. 9 are specialized to prove large and moderate deviation properties for finite dimensional systems in Chap. 10. The setting is that of a standard SDE model with regular (e.g., Lipschitz continuous) coefficients. Infinite dimensional systems driven by Brownian noise, including SPDE, are the topic of Chap. 11. Our previously published versions of these results have found wide use in small noise SPDE models with multiplicative noise, and a listing of some of the applications is given in the notes at the end of Chap. 9.

Chapter 12 presents another use of the representation for infinite dimensional Brownian motion, which is to the large deviation theory for small noise flows of diffeomorphisms. Also included is an application to Bayesian image reconstruction. Chapter 13 returns to finite dimensional models and considers those with special

features, and emphasizes problems with less regularity in the mapping that takes the noise into the state. Chapter 13 illustrates a useful aspect of the weak convergence approach, which is that by converting large deviation questions into questions of weak convergence, one can more easily understand exactly what regularity conditions are really needed. For related works that highlight this same feature, see [1, 2, 3]. An important class of problems covered in Chap. 13 is that of large and moderate deviations for certain pure jump processes that, when written in the form of a stochastic differential equation driven by a Poisson random measure, have discontinuous coefficients in the stochastic integral and are therefore not covered by Chap. 10.

Chapter 8
Representations for Continuous Time Processes

In previous chapters we developed and applied representations for the large deviation analysis of discrete time processes. The derivation of useful representations in this setting follows from a straightforward application of the chain rule. The only significant issue is to decide on the ordering used for the underlying "driving noises" when the chain rule is applied, since controls are allowed to depend on the "past," which is determined by this ordering.

In continuous time, the situation is both simpler and more complex. It is simpler in that most models in continuous time can be conveniently represented as systems driven by an exogenous noise process of either Gaussian or Poisson type. As we will see, useful representations hold in great generality for both types of noise. It is also more complex, in that the chain rule cannot be directly applied, and one must approximate and justify suitable limits to establish the representations. In the end, the representations take a form that is analogous to their discrete time counterparts, and we consider controls that are allowed to depend on the past, i.e., controls that are predictable with respect to a suitable filtration.[1]

This chapter consists of three sections, which present the representations for functionals of infinite dimensional Brownian motion, functionals of a Poisson random measure, and the combined case. The proofs given here differ from the first versions that appeared in [39, 45]. In particular, while the details are different, the approach to both models is very much the same.

[1] For special cases, one can consider the infimum of a smaller class (e.g., feedback controls), a result that is sometimes of interest.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_8

8.1 Representation for Infinite Dimensional Brownian Motion

The starting point of the proof of the representation is of course (2.1). Hence we will need to understand the form of $d\gamma/d\theta$ when θ is the measure induced by an infinite dimensional Brownian motion. Several formulations of infinite dimensional Brownian motion are commonly used. We focus for now on the formulation as a Hilbert space valued Wiener process,[2] and comment in Chap. 11 on how representations for other formulations follow easily from this one.

8.1.1 The Representation

Let $(\Omega, \mathscr{F}, P)$ be a probability space with a filtration $\{\mathscr{F}_t\}_{0\le t\le T}$ satisfying the usual conditions. We begin with the definition of a Hilbert space valued Wiener process. Let $(\mathscr{H}, \langle\cdot,\cdot\rangle)$ be a real separable Hilbert space. Let Λ be a symmetric strictly positive trace class operator on $\mathscr{H}$ (see Appendix E for definitions and terminology related to Hilbert spaces). This means that Λ is a bounded linear operator such that if $\{e_i\}_{i\in\mathbb{N}}$ is any complete orthonormal sequence (CONS) in $\mathscr{H}$, then for all $i, j \in \mathbb{N}$, we have $\langle e_i, \Lambda e_j\rangle = \langle e_j, \Lambda e_i\rangle$, $\langle e_i, \Lambda e_i\rangle > 0$, and $\sum_{i=1}^{\infty}\langle e_i, \Lambda e_i\rangle < \infty$.

Definition 8.1 An $\mathscr{H}$-valued continuous stochastic process $\{W(t)\}_{0\le t\le T}$ is called a Λ-Wiener process with respect to $\{\mathscr{F}_t\}_{0\le t\le T}$ if for every nonzero $h \in \mathscr{H}$, $\langle \Lambda h, h\rangle^{-1/2}\langle W(t), h\rangle$ is a standard one-dimensional $\mathscr{F}_t$-Wiener process (see Sect. 3.2).

Define $\mathscr{H}_0 \doteq \Lambda^{1/2}\mathscr{H}$. Then $\mathscr{H}_0$ is a Hilbert space with the inner product

$$\langle h, k\rangle_0 \doteq \left\langle \Lambda^{-1/2}h, \Lambda^{-1/2}k\right\rangle$$

for $h, k \in \mathscr{H}_0$. Denote the norms in $\mathscr{H}$ and $\mathscr{H}_0$ by $\|\cdot\|$ and $\|\cdot\|_0$ respectively. Since Λ is trace class, the identity mapping from $\mathscr{H}_0$ to $\mathscr{H}$ is Hilbert–Schmidt. This Hilbert–Schmidt embedding of $\mathscr{H}_0$ in $\mathscr{H}$ will play a central role in many of the arguments to follow. An important consequence of the embedding is that if v^n is a sequence in $\mathscr{H}_0$ such that $v^n \to 0$ weakly in $\mathscr{H}_0$, then $\|v^n\| \to 0$. For an exposition of stochastic calculus with respect to an $\mathscr{H}$ valued Wiener process, we refer to [69]. Other useful references are [197, 198, 252].

We first present and prove a representation that uses controls that are predictable with respect to the filtration generated by the Wiener process, and later, in Sect. 8.1.5, we extend this representation to controls that are predictable with respect to $\{\mathscr{F}_t\}$. Let $\{\mathscr{G}_t\}_{0\le t\le T}$ be the filtration generated by $\{W(t)\}_{0\le t\le T}$ augmented with all P-null sets in $\mathscr{F}$.

[2] We will use the terms "Brownian motion" and "Wiener process" interchangeably.

Definition 8.2 Given $0 \le a < b \le T$ and a bounded $\mathscr{F}_a$-measurable real random variable ξ, let $g : [0,T] \times \Omega \to \mathbb{R}$ be defined by $g(s,\omega) \doteq \xi(\omega)1_{(a,b]}(s), s \in [0,T]$, $\omega \in \Omega$. Denote by $\mathscr{P}\mathscr{F}$ the σ-field on $[0,T] \times \Omega$ generated by the collection of all such g. This σ-field is called the $\mathscr{F}_t$-predictable σ-field. For a Polish space $\mathscr{E}$, a $\mathscr{P}\mathscr{F}/\mathscr{B}(\mathscr{E})$-measurable map $v : [0,T] \times \Omega \to \mathscr{E}$ is referred to as an $\mathscr{E}$-valued $\mathscr{F}_t$-predictable process.

Define $\bar{\mathscr{A}}$ to be the class of $\mathscr{H}_0$-valued $\mathscr{F}_t$-predictable processes v that satisfy

$$P\left\{\int_0^T \|v(s)\|_0^2 ds < \infty\right\} = 1,$$

and let $\mathscr{A}$ denote the subset of those that are predictable with respect to $\{\mathscr{G}_t\}_{0\le t\le T}$. We refer to [69, Chap. 4] for the definition of stochastic integrals of elements of $\bar{\mathscr{A}}$ with respect to W. Let $\mathscr{L}^2([0,T] : \mathscr{H}_0)$ denote the Hilbert space of all measurable maps $u : [0,T] \to \mathscr{H}_0$ for which $\int_{[0,T]} \|u(s)\|_0^2 ds$ is finite together with the usual inner product, and for $M \in \mathbb{N}$, define

$$S_M \doteq \left\{u \in \mathscr{L}^2([0,T] : \mathscr{H}_0) : \int_0^T \|u(s)\|_0^2 ds \le M\right\}. \tag{8.1}$$

We endow S_M with the weak topology, which makes it a compact Polish space (cf. [93]). In particular, a sequence $\{v_n\} \subset S_M$ converges to $v \in S_M$ if $\int_0^T \langle v_n(s), h(s)\rangle_0 ds$ converges to $\int_0^T \langle v(s), h(s)\rangle_0 ds$ for all $h \in \mathscr{L}^2([0,T] : \mathscr{H}_0)$. Finally, let

$$\mathscr{A}_{b,M} \doteq \{v \in \mathscr{A} : v(\omega) \in S_M \ \theta\text{-}a.s.\}, \qquad \mathscr{A}_b \doteq \cup_{M\in\mathbb{N}} \mathscr{A}_{b,M}. \tag{8.2}$$

Let $\bar{\mathscr{A}}_{b,M}$ [resp. $\bar{\mathscr{A}}_b$] be defined exactly as $\mathscr{A}_{b,M}$ [resp. $\mathscr{A}_b$], except that $\{\mathscr{G}_t\}$ is replaced by $\{\mathscr{F}_t\}$.

We next state the main result of this section. Let E denote expectation with respect to P. Though in the theorem we take G to be a bounded function, it can be shown that the representation holds if G is bounded from above. The fact that the representation also holds with respect to the smaller class $\mathscr{A}_b \subset \mathscr{A}$ is quite convenient in applications, since these are in some sense very well behaved processes.

Theorem 8.3 *Let W be a Λ-Wiener process and let G be a bounded Borel measurable function mapping $\mathscr{C}([0,T] : \mathscr{H})$ into $\mathbb{R}$. Then*

$$-\log E\exp\{-G(W)\} = \inf_{v\in\mathscr{R}} E\left[\frac{1}{2}\int_0^T \|v(s)\|_0^2 ds + G\left(W + \int_0^{\cdot} v(s)ds\right)\right], \tag{8.3}$$

where $\mathscr{R}$ can be either $\mathscr{A}_b$, $\mathscr{A}$, $\bar{\mathscr{A}}_b$ or $\bar{\mathscr{A}}$.

Using (8.3), one can prove the following in an identical manner as Theorem 3.17, and we therefore omit the proof.

Theorem 8.4 *Let W and G be as in Theorem 8.3 and let $\delta > 0$. Then there exists $M < \infty$ depending on $\|G\|_\infty$ and δ such that for all $\varepsilon \in (0,1)$,*

$$-\varepsilon \log E \exp\left\{-\frac{1}{\varepsilon} G(\sqrt{\varepsilon} W)\right\}$$
$$\geq \inf_{v \in \mathscr{A}_{b,M}} E\left[\frac{1}{2}\int_0^T \|v(s)\|_0^2 ds + G\left(\sqrt{\varepsilon} W + \int_0^{\cdot} v(s) ds\right)\right] - \delta.$$

The rest of Sect. 8.1 is devoted to the proof of Theorem 8.3. After developing the needed preliminary results, the proof for $\mathscr{A}_b$ and $\mathscr{A}$ is given in Sects. 8.1.3 and 8.1.4, and the extension to $\bar{\mathscr{A}}_b$ and $\bar{\mathscr{A}}$ is completed in Sect. 8.1.5.

8.1.2 Preparatory Results

In this section we present several theorems and approximations that will be used in the proof of the representation. For use later on, some results are stated for the more general class of processes $\bar{\mathscr{A}}$. The following result follows from Theorem 10.14 of [69].

Theorem 8.5 (GIRSANOV) *Let $\psi \in \bar{\mathscr{A}}$ be such that*

$$E\left[\exp\left\{\int_0^T \langle \psi(s), dW(s)\rangle_0 - \frac{1}{2}\int_0^T \|\psi(s)\|_0^2 ds\right\}\right] = 1.$$

Then the process

$$\tilde{W}(t) \doteq W(t) - \int_0^t \psi(s) ds,$$

$t \in [0,T]$, is a Λ-Wiener process with respect to $\{\mathscr{F}_t\}$ on $(\Omega, \mathscr{F}, Q)$, where Q is the probability measure defined by

$$\frac{dQ}{dP} = \exp\left\{\int_0^T \langle \psi(s), dW(s)\rangle_0 - \frac{1}{2}\int_0^T \|\psi(s)\|_0^2 ds\right\}.$$

We record a result that will be used in proving tightness for a sequence of Hilbert space valued processes. Recall the topology on S_N introduced below (8.1).

Lemma 8.6 *Let $\{v^n\}_{n\in\mathbb{N}}$ be a sequence of elements of $\bar{\mathscr{A}}$. Assume that there is $M < \infty$ such that*

$$\sup_{n\in\mathbb{N}} \int_0^T \|v^n(s)\|_0^2 ds \leq M$$

a.s. Suppose further that $\{v^n\}$ converges in distribution to v as S_M-valued random variables. Then $\int_0^{\cdot} v^n(s) ds$ converges in distribution to $\int_0^{\cdot} v(s) ds$ in $\mathscr{C}([0,T] : \mathscr{H})$.

Proof It suffices to show that the map from S_M to $\mathscr{C}([0,T]:\mathscr{H})$ defined by $u \mapsto \int_0^{\cdot} u(s)ds$ is continuous. Let $\{\phi^n\}$ be a sequence in S_M that converges to ϕ. Let $\{e_j\}_{j\in\mathbb{N}}$ be a CONS of eigenvectors of Λ with corresponding eigenvalues $\{\lambda_j\}_{j\in\mathbb{N}}$. Then for every $t \in [0,T]$ and $i \in \mathbb{N}$,

$$\int_0^t \langle \phi^n(s) - \phi(s), e_i\rangle ds = \lambda_i \int_0^t \langle \phi^n(s) - \phi(s), e_i\rangle_0 ds,$$

and by assumption, the right side converges to 0 as $n \to \infty$. Also,

$$\left\| \int_0^t [\phi^n(s) - \phi(s)]ds \right\|^2 = \sum_{i=1}^{\infty} \left(\int_0^t \langle \phi^n(s) - \phi(s), e_i\rangle ds \right)^2.$$

By Hölder's inequality and $\Lambda^{1/2} e_i = \lambda_i e_i$,

$$\begin{aligned}
\left(\int_0^t \langle \phi^n(s) - \phi(s), e_i\rangle ds \right)^2 &\le \lambda_i T \int_0^T \|\Lambda^{-1/2}[\phi^n(s) - \phi(s)]\|^2 ds \\
&= \lambda_i T \int_0^T \|\phi^n(s) - \phi(s)\|_0^2 ds \\
&\le 4MT\lambda_i.
\end{aligned}$$

Since $\sum_{i=1}^{\infty} \lambda_i < \infty$, it follows from the dominated convergence theorem that for each $t \in [0,T]$, $\int_0^t \phi^n(s)ds$ converges to $\int_0^t \phi(s)ds$ in $\mathscr{H}$. To prove that this convergence is uniform on $[0,T]$, we need an equicontinuity estimate. This follows by noting that for $0 \le s \le t \le T$,

$$\begin{aligned}
\left\| \int_0^t \phi^n(r)dr - \int_0^s \phi^n(r)dr \right\| &\le \sqrt{t-s} \left(\int_0^T \|\phi^n(s)\|^2 ds \right)^{1/2} \\
&\le \sqrt{t-s}\, \|\Lambda\|^{1/2} \left(\int_0^T \|\Lambda^{-1/2}\phi^n(s)\|^2 ds \right)^{1/2} \\
&= \sqrt{t-s}\, \|\Lambda\|^{1/2} \left(\int_0^T \|\phi^n(s)\|_0^2 ds \right)^{1/2} \\
&\le \sqrt{t-s}\, \|\Lambda\|^{1/2} M^{1/2},
\end{aligned}$$

where $\|\Lambda\| \doteq \sup_{h\in\mathscr{H}:\|h\|=1} \|\Lambda h\|$ is the operator norm. □

Before turning to the proof of Theorem 8.3, we state one last result. A process $v \in \mathscr{A}$ is called **simple** if there exist $k \in \mathbb{N}$, $0 = t_1 \le \cdots \le t_{k+1} = T$ and $N \in \mathbb{N}$ such that

$$v(s,\omega) \doteq \sum_{j=1}^{k} X_j(\omega) 1_{(t_j, t_{j+1}]}(s),$$

where the X_j are $\mathscr{H}_0$-valued $\mathscr{G}_{t_j}$-measurable random variables satisfying $\|X_j(\omega)\|_0 \leq N$ for all $j \in \{1, \ldots, k\}$. Let $\mathscr{A}_s$ denote the collection of simple processes, and note that $\mathscr{A}_s \subset \mathscr{A}_b$. Given any $v \in \mathscr{A}_s$, it is straightforward that

$$E\left[\exp\left\{\int_0^T \langle v(s), dW(s)\rangle_0 - \frac{1}{2}\int_0^T \|v(s)\|_0^2 ds\right\}\right] = 1,$$

and thus by Theorem 8.5, the process

$$W^v(t) \doteq W(t) - \int_0^t v(s)ds,$$

$t \in [0, T]$, is a Λ-Wiener process with respect to $\{\mathscr{G}_t\}$ on $(\Omega, \mathscr{F}, Q^v)$, where Q^v is the probability measure defined by

$$\frac{dQ^v}{dP} = \exp\left\{\int_0^T \langle v(s), dW(s)\rangle_0 - \frac{1}{2}\int_0^T \|v(s)\|_0^2 ds\right\}.$$

Let E^v denote integration with respect to Q^v.

Lemma 8.7 *For every $v \in \mathscr{A}_s$, there is $\tilde{v} \in \mathscr{A}_s$ such that $(W^{\tilde{v}}, \tilde{v})$ has the same distribution under $Q^{\tilde{v}}$ as (W, v) does under P.*

Proof Let v be simple and of the form

$$v(s, \omega) \doteq \sum_{j=1}^{k} X_j(\omega) 1_{(t_j, t_{j+1}]}(s),$$

where $k \in \mathbb{N}$, $0 = t_1 \leq \cdots \leq t_{k+1} = T$ and X_j are $\mathscr{H}_0$-valued $\mathscr{G}_{t_j}$-measurable random variables satisfying $\|X_j(\omega)\|_0 \leq N$ for all $j \in \{0, \ldots, k\}$ and some $N \in \mathbb{N}$. New random variables $\bar{X}_j$, $j \in \{0, \ldots, k\}$, are defined as follows. Since $X_1(\omega)$ is $\mathscr{G}_0$-measurable, there exists measurable $G_1 : \mathscr{H}_0 \to \mathscr{H}_0$ such that $X_1(\omega) = G_1(W(0, \omega))$ a.s. Let $\bar{X}_1 \doteq G_1(W(0)) = X_1$. For $j \in \{2, \ldots, k\}$, there are measurable $G_j : \mathscr{C}([0, t_j] : \mathscr{H}_0) \to \mathscr{H}_0$ such that $X_j(\omega) = G_j(W(t, \omega), 0 \leq t \leq t_j)$ a.s. We can also consider G_j as a mapping $\mathscr{C}([0, T] : \mathscr{H}_0) \to \mathscr{H}_0$, which depends on $w \in \mathscr{C}([0, T] : \mathscr{H}_0)$ only though the restriction to $[0, t_j]$, and we do so with the notation $G_j(w)$. We then recursively define

$$\bar{X}_j \doteq G_j\left(W(\cdot) - \int_0^{\cdot} \sum_{i=1}^{j-1} \bar{X}_i 1_{(t_i, t_{i+1}]}(s)ds\right).$$

By construction, each $\bar{X}_j$ is $\mathscr{G}_{t_j}$-measurable and satisfies $\|\bar{X}_j(\omega)\|_0 \leq N$ for a.e. ω. Now let

$$\tilde{v}(s,\omega) \doteq \sum_{j=1}^{k} \bar{X}_j(\omega) 1_{(t_j,t_{j+1}]}(s),$$

and note that

$$\tilde{v}(s) \doteq \sum_{j=1}^{k} G_j\left(W(\cdot) - \int_0^{\cdot} \tilde{v}(s)ds\right) 1_{(t_j,t_{j+1}]}(s). \tag{8.4}$$

By Theorem 8.5, $W(t) - \int_0^t \tilde{v}(s)ds$ is a Λ-Wiener process under $Q^{\tilde{v}}$. Since $\tilde{v}$ has the form given in (8.4), it follows that $(W^{\tilde{v}}, \tilde{v})$ has the same distribution under $Q^{\tilde{v}}$ as (W, v) does under P. □

8.1.3 *Proof of the Upper Bound in the Representation*

In this subsection we prove

$$-\log E\exp\{-G(W)\} \le \inf_{v\in\mathscr{A}} E\left[\frac{1}{2}\int_0^T \|v(s)\|_0^2 ds + G\left(W + \int_0^{\cdot} v(s)ds\right)\right].$$

Note that this automatically gives the corresponding bound for the smaller class $\mathscr{A}_b$ in (8.3). The proof is in two steps.

Step 1. Simple v. According to (2.1), for every probability measure Q on $(\Omega, \mathscr{F})$,

$$-\log E\exp\{-G(W)\} \le R(Q\,\|\,P) + \int_\Omega G(W)dQ. \tag{8.5}$$

If $v \in \mathscr{A}_s$, then by Lemma 8.7 there is $\tilde{v} \in \mathscr{A}_s$ such that the distribution of (W, v) under P is the same as that of $(W^{\tilde{v}}, \tilde{v})$ under $Q^{\tilde{v}}$. Since $\tilde{v}$ is bounded, it follows from Theorem 8.5 that

$$\begin{aligned}
R(Q^{\tilde{v}}\,\|\,P) &= E^{\tilde{v}}\left[\int_0^T \langle \tilde{v}(s), dW(s)\rangle_0 - \frac{1}{2}\int_0^T \|\tilde{v}(s)\|_0^2 ds\right] \\
&= E^{\tilde{v}}\left[\int_0^T \left\langle \tilde{v}(s), dW^{\tilde{v}}(s)\right\rangle_0 + \frac{1}{2}\int_0^T \|\tilde{v}(s)\|_0^2 ds\right] \\
&= E^{\tilde{v}}\left[\frac{1}{2}\int_0^T \|\tilde{v}(s)\|_0^2 ds\right] \\
&= E\left[\frac{1}{2}\int_0^T \|v(s)\|_0^2 ds\right].
\end{aligned}$$

Taking $Q = Q^{\tilde{v}}$ in (8.5) together with $E^{\tilde{v}}G(W) = E^{\tilde{v}}G(W^{\tilde{v}} + \int_0^{\cdot} \tilde{v}(s)ds) = EG(W + \int_0^{\cdot} v(s)ds)$ gives

$$-\log E\exp\{-G(W)\} \le E\left[\frac{1}{2}\int_0^T \|v(s)\|_0^2 ds + G\left(W + \int_0^{\cdot} v(s)ds\right)\right]. \quad (8.6)$$

Step 2. General v. Next consider $v \in \mathscr{A}$. We can assume without loss of generality that $E[\int_0^T \|v(s)\|_0^2 ds] < \infty$. Then (see, for example, [159, Lemma II.1.1]) there is a sequence $\{v_n\} \subset \mathscr{A}_s$ such that

$$E\int_0^T \|v_n(s) - v(s)\|_0^2 ds \to 0. \quad (8.7)$$

In particular, $N \doteq \sup_{n\in\mathbb{N}} E\int_0^T \|v_n(s)\|_0^2 ds < \infty$. From Step 1, for all n,

$$-\log E\exp\{-G(W)\} \le E\left[\frac{1}{2}\int_0^T \|v_n(s)\|_0^2 ds + G\left(W + \int_0^{\cdot} v_n(s)ds\right)\right]. \quad (8.8)$$

We would like to apply Lemma 2.5, where μ_n and θ are the distributions induced by $W + \int_0^{\cdot} v_n(s)ds$ and W under P, respectively. Since μ_n is also the distribution induced by W under $Q^{\tilde{v}_n}$, part (f) of Lemma 2.4 implies

$$R(\mu_n \,\|\theta) \le R(Q^{\tilde{v}_n}\,\|P) = E^{\tilde{v}_n}\left[\frac{1}{2}\int_0^T \|\tilde{v}_n\|_0^2 ds\right] = E\left[\frac{1}{2}\int_0^T \|v_n\|_0^2 ds\right].$$

Thus $\sup_n R(\mu_n\,\|\theta) \le N/2$. From (8.7), it follows that

$$E\sup_{0\le t\le T}\left\|\int_0^t v_n(s)ds - \int_0^t v(s)ds\right\|^2 \le T\|\Lambda\|_{\mathrm{op}}^2 E\int_0^T \|v_n(s) - v(s)\|_0^2 ds \to 0,$$

and therefore μ_n converges weakly to μ, where μ is the distribution of $W + \int_0^{\cdot} v(s)ds$. Since G is bounded and measurable, we now obtain (8.6) using Lemma 2.5 and sending $n \to \infty$ in (8.8). □

8.1.4 *Proof of the Lower Bound in the Representation*

In this subsection we prove

$$-\log E\exp\{-G(W)\} \ge \inf_{v\in\mathscr{A}_b} E\left[\frac{1}{2}\int_0^T \|v(s)\|_0^2 ds + G\left(W + \int_0^{\cdot} v(s)ds\right)\right]. \quad (8.9)$$

This automatically gives the corresponding bound for the larger class $\mathscr{A}$ in (8.3). The proof is in two steps.

Step 1. G of a particular form. We first consider G of a special form. Recall that $\{e_n\}_{n\in\mathbb{N}}$ denotes a CONS in $\mathscr{H}$. Let $K, N \in \mathbb{N}$ be arbitrary, and consider any

collection $0 = t_1 < t_2 < \cdots < T_K = T$. Let $h : \mathbb{R}^{KN} \to \mathbb{R}$ have compact support and continuous derivatives of all orders. Then G is of the form

$$G(W) = h(w(t_1), w(t_2) - w(t_1), \ldots, w(t_K) - w(t_{K-1})), \tag{8.10}$$

where for $0 \le t \le T$,

$$w(t) = (\lambda_1^{-1/2} \langle e_1, W(t) \rangle, \ldots, \lambda_N^{-1/2} \langle e_N, W(t) \rangle). \tag{8.11}$$

Note that $\{w(t)\}_{0 \le t \le T}$ is an N-dimensional standard $\mathscr{G}_t$-Wiener process. Using methods from stochastic control theory, we will construct a process $v \in \mathscr{A}_b$ that gives equality in (8.9). The following lemma, whose proof is omitted, follows by classical and elementary stochastic control arguments that apply when there is a smooth value function (see Sect. VI.2 of [134]).

Lemma 8.8 *Let $g : \mathbb{R}^m \times \mathbb{R}^N \to \mathbb{R}$ have compact support and continuous derivatives of all orders. Let $\{w(t)\}_{0 \le t \le T}$ be an N-dimensional standard Brownian motion, and let $V : [0, T] \times \mathbb{R}^m \times \mathbb{R}^N \to \mathbb{R}$ be defined by*

$$V(t, z, x) \doteq -\log E e^{-g(z, x + w(T-t))}.$$

Then the following hold.

(a) *For all $(t, x) \in [0, T] \times \mathbb{R}^N$, $z \mapsto V(t, z, x)$ has compact support and derivatives of all orders that are continuous functions of (t, z, x).*

(b) *For all $(t, z) \in [0, T] \times \mathbb{R}^m$, $x \mapsto V(t, z, x)$ has derivatives of all orders that are continuous and bounded functions of (t, z, x).*

(c) *For $z \in \mathbb{R}^m$, let $\{X(z, t)\}_{0 \le t \le T}$ be the unique solution of*

$$X(z, t) = -\int_0^t D_x V(s, z, X(z, s)) ds + w(t), \ t \in [0, T].$$

Then with $u(t) = -D_x V(t, z, X(z, t))$ for $t \in [0, T]$,

$$-\log E \exp\{-g(z, w(T))\} = E\left[\frac{1}{2}\int_0^T \|u\|^2 ds + g\left(z, w + \int_0^{\cdot} u ds\right)\right]. \tag{8.12}$$

Remark 8.9 For the proof of Lemma 8.8, one starts with the linear partial differential equation (PDE) for which $(t, z, x) \mapsto E e^{-g(z, x + w(T-t))}$ is a classical-sense solution. From this, one obtains the nonlinear PDE (Hamilton–Jacobi–Bellman equation) for which V is a classical-sense solution. As such, V also has an interpretation as the minimal cost in a stochastic optimal control problem. Using a classical verification argument [134], it is straightforward to show that u as defined in the lemma is the optimal control, and the right-hand side of (8.12) is the minimal cost starting from $x = 0$, which establishes (8.12).

Now for $j = 1, \ldots, K$, define $V_j : \mathbb{R}^{jN} \to \mathbb{R}$ as follows: $V_K = h$ and

$$V_j(z_j) = -\log E e^{-V_{j+1}(z_j, w(t_{j+1}) - w(t_j))}, \ z_j \in \mathbb{R}^{jN}, \ j = 1, \ldots, K-1.$$

By successive conditioning, it is easily checked that

$$V_0 \doteq -\log E e^{-V_1(w(t_1) - w(t_0))} = -\log E e^{-G(W)},$$

where G is as in (8.10). From part (a) of Lemma 8.8 it follows that for all $j = 1, \ldots, K$, V_j has continuous and bounded derivatives of all orders. For $j = 1, \ldots, K$, let $Z_j = (w(t_1), w(t_2) - w(t_1), \ldots, w(t_j) - w(t_{j-1}))$, and note that Z_j is an $\mathbb{R}^{jN}$-valued random variable. For $z_j \in \mathbb{R}^{jN}$, let $\{Y(z_j, t)\}_{t \in [t_j, t_{j+1}]}$, $j = 1, \ldots, K-1$, be the unique solution of

$$Y(z_j, t) = -\int_{t_j}^{t} D_x V_{j+1}(s, z_j, Y(z_j, s))ds + w(t) - w(t_j), \ t \in [t_j, t_{j+1}].$$

The existence and uniqueness of the solution is a consequence of the smoothness property of V_{j+1} noted earlier. Now define

$$u(t) = -D_x V_{j+1}(t, Z_j, Y(Z_j, t)), \ t \in [t_j, t_{j+1}), \ j = 0, \ldots, K-1.$$

Then by a straightforward recursive argument using Lemma 8.8, we see that

$$-\log E e^{-G(W)} = E\left[\frac{1}{2}\int_0^T \|u(s)\|^2 ds + h\left(w(t_1) + \int_0^{t_1} u(s)ds, \ldots \right.\right.$$
$$\left.\left., w(t_K) - w(t_{K-1}) + \int_{t_{K-1}}^{t_K} u(s)ds\right)\right].$$

Let $v(s) \doteq \sum_{i=1}^{N} \lambda_i^{1/2} u_i(s) e_i$, $s \in [0, T]$. Then $v \in \mathscr{A}_b$, and by (8.10) and (8.11),

$$-\log E \exp\{-G(W)\} = E\left[\frac{1}{2}\int_0^T \|v(s)\|_0^2 ds + G\left(W + \int_0^{\cdot} v(s)ds\right)\right].$$

Thus we have proved (8.9) for all G of the form (8.10)–(8.11).

Step 2. G that is bounded and measurable. Now suppose that G is simply bounded and measurable. We claim that there exist functions $\{G_n\}_{n \in \mathbb{N}}$ such that for each n, G_n is of the form assumed in Step 1, $\|G_n\|_\infty \le \|G\|_\infty$, and $G_n \to G$ a.s. with respect to θ. This can be seen most easily by considering the approximation in stages. We note that each of the following classes admits an approximation of this form relative to elements of the preceding class, save of course the first:

- G bounded and measurable;
- G bounded and continuous;

- $G(W) = H(W(t_1), W(t_2), \ldots, W(t_K))$, where $H : \mathscr{H}^K \to \mathbb{R}$ is continuous and bounded and $K \in \mathbb{N}$ and $0 = t_1 < t_2 < \cdots < T_K = T$ are arbitrary;
- G of the form (8.10)–(8.11) where h is a bounded and continuous function from $\mathbb{R}^{NK} \to \mathbb{R}$, $K, N \in \mathbb{N}$ and $0 = t_1 < t_2 < \cdots < t_K = T$ are arbitrary;
- G as above and in addition, h has compact support;
- G as above and in addition, h has continuous and bounded derivatives of all orders.

All of these approximations follow by standard arguments. The first approximation statement (i.e., a bounded measurable G can be approximated by a bounded continuous G) is the conclusion of a result due to Doob and presented in the appendix as Theorem E.4. For the second statement we use the martingale convergence theorem, which states that if $\{\mathscr{F}_n\}$ is a filtration increasing to the σ-field $\mathscr{F}_\infty$ and X is an integrable $\mathscr{F}_\infty$-measurable random variable, then $E[X \mid \mathscr{F}_n]$ converges to X a.s. Consider a sequence of partitions $\pi_n = \{0 = t_1^n < t_2^n < \cdots < t_{K_n}^n = T\}$ such that $\pi_n \subset \pi_{n+1}$ and $|\pi_n| \doteq \max_{1 \le j \le K_n - 1}(t_{j+1}^n - t_j^n) \to 0$ as $n \to \infty$. Let $\mathscr{F}_n \doteq \sigma\{W(t_i^n), 1 \le i \le K_n\}$. Clearly, $\mathscr{F}_n$ is a filtration and G is $\mathscr{F}_\infty \doteq \sigma(\cup_{n \ge 1} \mathscr{F}_n)$-measurable. Thus by the martingale convergence theorem, $G_n = E[G \mid \mathscr{F}_n]$ converges a.s. to G. Clearly, $\|G_n\|_\infty \le \|G\|_\infty$ a.s. The second approximation statement now follows from another application of Theorem E.4 if G_n is not continuous. The proof of the third approximation statement is similar but uses the filtration

$$\mathscr{F}_n \doteq \sigma\{\langle W(t_i), e_j \rangle, j = 1, \ldots, n, \ i = 1, \ldots, K\},$$

where $\{e_j\}_{j \in \mathbb{N}}$ is a CONS in $\mathscr{H}$. The fourth approximation statement involves replacing h in (8.10)–(8.11) by $h\psi_n$ in defining G_n, where ψ_n is a continuous function with values in $[0, 1]$ such that $\psi_n(x) = 1$ when x is in a ball of radius n and $\psi_n(x) = 0$ outside a ball of radius $n + 1$. Finally, the last statement follows by replacing h with $h * \eta_n$ in (8.10)–(8.11), where $\eta_n(x) = n^{-NK}\eta(nx)$, $x \in \mathbb{R}^{NK}$,

$$\eta(x) \doteq c \exp\left\{-\frac{1}{1 - |x|^2}\right\} 1_{\{|x|<1\}},$$

and c is the normalizing constant such that $\int \eta(x)dx = 1$.

With the claim verified, we now complete the lower bound. With each $n \in \mathbb{N}$ we can associate $v_n \in \mathscr{A}_b$ such that

$$-\log E \exp\{-G_n(W)\} = E\left[\frac{1}{2}\int_0^T \|v_n(s)\|_0^2 ds + G_n\left(W + \int_0^{\cdot} v_n(s)ds\right)\right].$$

As in the proof of the upper bound, if μ_n is the distribution induced by $W + \int_0^{\cdot} v_n(s)ds$, then

$$R(\mu_n \,\|\theta) \le E\left[\frac{1}{2}\int_0^T \|v_n(s)\|_0^2 ds\right] \le 2\,\|G\|_\infty,$$

where the last inequality is valid because $\|G_n\|_\infty \le \|G\|_\infty$. Thus $\{\mu_n\}$ is tight, and from part (b) of Lemma 2.5, we have

$$\lim_{n\to\infty} E\left|G_n\left(W+\int_0^{\cdot} v_n(s)ds\right) - G\left(W+\int_0^{\cdot} v_n(s)ds\right)\right| = 0.$$

By the dominated convergence theorem,

$$\lim_{n\to\infty} |\log E\exp\{-G_n(W)\} - \log E\exp\{-G(W)\}| = 0.$$

Therefore, given $\varepsilon > 0$, we can find $n \in \mathbb{N}$ such that

$$\begin{aligned} -\log Ee^{-G(W)} &\ge -\log Ee^{-G_n(W)} - \varepsilon \\ &= E\left[\frac{1}{2}\int_0^T \|v_n(s)\|_0^2 ds + G_n\left(W+\int_0^{\cdot} v_n(s)ds\right)\right] - \varepsilon \\ &\ge E\left[\frac{1}{2}\int_0^T \|v_n(s)\|_0^2 ds + G\left(W+\int_0^{\cdot} v_n(s)ds\right)\right] - 2\varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary and $v_n \in \mathscr{A}_b$, we have (8.9), completing the proof. □

8.1.5 Representation with Respect to a General Filtration

We now return to the issue of whether the representation holds when $\{\mathscr{G}_t\}_{0\le t\le T}$, the filtration generated by the Wiener process, is replaced by any filtration $\{\mathscr{F}_t\}_{0\le t\le T}$ that satisfies the usual conditions and such that W is a Λ-Wiener process with respect to this larger filtration.

We will make use of the following lemma on measurable selections.

Lemma 8.10 *Let $\mathscr{E}_1, \mathscr{E}_2$ be Polish spaces and let $g : \mathscr{E}_1 \times \mathscr{E}_2 \to \mathbb{R}$ be a bounded continuous function. Let K be a compact set in $\mathscr{E}_2$. For each $x \in \mathscr{E}_1$, define*

$$\Gamma_x \doteq \left\{y \in K : \inf_{y_0\in K} g(x, y_0) = g(x, y)\right\}.$$

Then there exists a Borel measurable function $g_1 : \mathscr{E}_1 \to \mathscr{E}_2$ such that $g_1(x) \in \Gamma_x$ for all $x \in \mathscr{E}_1$.

Proof Let x_n be a sequence in $\mathscr{E}_1$ converging to $\bar{x}$. For each $n \in \mathbb{N}$, let $y_n \in \Gamma_{x_n}$. In view of Corollary E.3, it suffices to show that $\{y_n\}$ has a limit point in $\Gamma_{\bar{x}}$. Let $\bar{y}$ be a limit point of $\{y_n\}$. For each n, $g(x_n, y_n) - \inf_{y_0\in K} g(x_n, y_0)$ equals zero. Since the map $(x, y) \mapsto g(x, y) - \inf_{y_0\in K} g(x, y_0)$ is continuous, letting $n \to \infty$ shows that $\bar{y} \in \Gamma_x$. □

Recall that $\bar{\mathscr{A}}$ was defined exactly as $\mathscr{A}$, except with $\{\mathscr{G}_t\}$ replaced by $\{\mathscr{F}_t\}$. The only issue to check is whether the upper bound

$$-\log E\exp\{-G(W)\} \le \inf_{v\in\bar{\mathscr{A}}} E\left[\frac{1}{2}\int_0^T \|v(s)\|_0^2 ds + G\left(W + \int_0^{\cdot} v(s)ds\right)\right] \tag{8.13}$$

continues to hold.

The only place where the structure of $\{\mathscr{G}_t\}_{0\le t\le T}$ is used in Sect. 8.1.3 is in the proof of (8.6), where we appeal to Lemma 8.7 to argue that if $v \in \mathscr{A}_s$, then there is $\tilde{v} \in \mathscr{A}_s$ such that the distribution of (W, v) under P is the same as that of $(W^{\tilde{v}}, \tilde{v})$ under $Q^{\tilde{v}}$. We can reduce to that case if we show that given $\varepsilon > 0$ and any control v that is simple with respect to $\{\mathscr{F}_t\}_{0\le t\le T}$, there is a $\bar{v} \in \mathscr{A}_s$ such that

$$\begin{aligned} &E\left[\frac{1}{2}\int_0^T \|\bar{v}(s)\|_0^2 ds + G\left(W + \int_0^{\cdot} \bar{v}(s)ds\right)\right] \\ &\le E\left[\frac{1}{2}\int_0^T \|v(s)\|_0^2 ds + G\left(W + \int_0^{\cdot} v(s)ds\right)\right] + \varepsilon. \end{aligned} \tag{8.14}$$

For simplicity, we consider the case $v(s) = 0$ for $s \in [0, t]$ and $v(s) = X$ for $s \in (t, T]$, where X is $\mathscr{F}_t$-measurable, and also assume that $\|X\|_0 \le M < \infty$ a.s. The generalization to the finite collection of random variables that appear in a simple control is straightforward (see [39]). For the moment, we also assume that G is continuous as well as bounded. Consider the mapping

$$g(\phi, x) = E\left[\frac{(T-t)}{2}\|x\|_0^2 + G\left(\phi^B + \int_0^{\cdot} 1_{[t,T]}(s)x ds\right)\right],$$

where $\phi \in \mathscr{C}([0, t] : \mathscr{H}_0)$, $x \in \{x \in \mathscr{H}_0 : \|x\|_0 \le M\}$, and

$$\phi^B(s) = \begin{cases} \phi(s), & s \in [0, t], \\ \phi(t) + B(s-t) - B(0), & s \in [t, T], \end{cases}$$

with B a Λ-Wiener process.

Note that g is bounded, and that by the dominated convergence theorem, it is also continuous in (ϕ, x). Consider the $\mathscr{C}([0, t] : \mathscr{H}_0)$-valued random variable $Z \doteq \{W(s)\}_{0\le s\le t}$. Then

$$E\left[\int_0^T \frac{1}{2}\|v(s)\|_0^2 + G\left(W + \int_0^{\cdot} v(s)ds\right)\right] = E[g(Z, X)].$$

Since a single probability measure on a Polish space is tight, there is a compact subset K_0 of $\mathscr{H}_0$ such that $P\{X \in K_0^c\} \le \varepsilon/(2\|g\|_\infty + 1)$. Then

$$E[g(Z, X)] \ge E[g(Z, X)1_{K_0}(X)] - \frac{\varepsilon}{2}.$$

Now we apply Lemma 8.10 with $\mathscr{E}_1 = \mathscr{C}([0,t] : \mathscr{H}_0), \mathscr{E}_2 = \mathscr{H}_0$, and $K = K_0 \cap \{x \in \mathscr{H}_0 : \|x\|_0 \le M\}$. Then there is a measurable map $g_1 : \mathscr{C}([0,t] : \mathscr{H}_0) \to K$ such that with $\bar{X} \doteq g_1(Z)$,

$$\begin{aligned} E[g(Z,X)] &\ge E[g(Z,X)1_{K_0}(X)] - \frac{\varepsilon}{2} \\ &\ge E[g(Z,g_1(Z))1_{K_0}(X)] - \frac{\varepsilon}{2} \\ &\ge E[g(Z,g_1(Z))] - \varepsilon \\ &= E\left[\frac{[T-t]}{2}\left\|\bar{X}\right\|_0^2 + G\left(W + \int_0^{\cdot} 1_{[t,T]}(s)\bar{X}ds\right)\right] - \varepsilon. \end{aligned}$$

Letting $\bar{v}(s) = 0$ for $s \in [0,t)$ and $\bar{v}(s) = \bar{X}$ for $s \in [t,T]$, we now have that

$$\begin{aligned} &E\left[\int_0^T \frac{1}{2}\|v(s)\|_0^2 + G\left(W + \int_0^{\cdot} v(s)ds\right)\right] \\ &\ge E\left[\int_0^T \frac{1}{2}\|\bar{v}(s)\|_0^2 + G\left(W + \int_0^{\cdot} \bar{v}(s)ds\right)\right] - \varepsilon. \end{aligned} \tag{8.15}$$

This completes the argument for the case that G is continuous.

Finally, we remove the assumption that G is continuous. Let v take the same form as previously, and suppose that G is bounded and measurable. It then follows from Theorem E.4 that there are bounded and continuous G_j that converge to G as $j \to \infty$ almost surely with respect to the distribution of W and that have the same uniform bound as G. Thus by the dominated convergence theorem, given $\varepsilon > 0$, we have for all sufficiently large $j \in \mathbb{N}$ that

$$E\left[G_j\left(W + \int_0^{\cdot} v(s)ds\right)\right] \le E\left[G\left(W + \int_0^{\cdot} v(s)ds\right)\right] + \frac{\varepsilon}{2}.$$

We have shown that there is $\bar{v}_j \in \mathscr{A}_s$ such that (8.15) holds with G replaced by G_j. Since $\sup_j R(\mu_j \,\|\theta) \le TM^2/2$, where μ_j is the probability distribution of $W + \int_0^{\cdot} \bar{v}_j(s)ds$, an application of Lemma 2.5 shows that for sufficiently large $j \in \mathbb{N}$,

$$E\left[G\left(W + \int_0^{\cdot} \bar{v}_j(s)ds\right)\right] \le E\left[G_j\left(W + \int_0^{\cdot} \bar{v}_j(s)ds\right)\right] + \frac{\varepsilon}{2}.$$

Thus for j that satisfy the last two displays, we have

$$\begin{aligned} &E\left[\int_0^T \frac{1}{2}\left\|\bar{v}_j(s)\right\|_0^2 + G\left(W + \int_0^{\cdot} \bar{v}_j(s)ds\right)\right] \\ &\le E\left[\int_0^T \frac{1}{2}\left\|\bar{v}_j(s)\right\|_0^2 + G_j\left(W + \int_0^{\cdot} \bar{v}_j(s)ds\right)\right] + \frac{\varepsilon}{2} \end{aligned}$$

$$\leq E\left[\int_0^T \frac{1}{2}\|v(s)\|_0^2 + G_j\left(W + \int_0^{\cdot} v(s)ds\right)\right] + \frac{\varepsilon}{2}$$
$$\leq E\left[\int_0^T \frac{1}{2}\|v(s)\|_0^2 + G\left(W + \int_0^{\cdot} v(s)ds\right)\right] + \varepsilon.$$

Since $\bar{v}_j \in \mathscr{A}_s$, we have (8.14), and the desired upper bound (8.13) follows. □

8.2 Representation for Poisson Random Measure

In this section we present the analogous representations for functionals of a Poisson random measure (PRM), the other important driving noise in continuous time. In contrast to the case of a Wiener process, for PRM it is convenient and in some sense necessary to enlarge the underlying probability space. The enlargement is needed to define a very general class of controlled Poisson random measures. An alternative approach that has been considered is to dilate time, thereby increasing or decreasing rates, but this method of producing a controlled PRM does not allow for a representation general enough. For further discussion, see [45, p. 726].

8.2.1 The Representation

For a locally compact Polish space $\mathscr{S}$, we denote by $\Sigma(\mathscr{S})$ the space of all measures ν on $(\mathscr{S}, \mathscr{B}(\mathscr{S}))$ satisfying $\nu(K) < \infty$ for every compact $K \subset \mathscr{S}$. We endow $\Sigma(\mathscr{S})$ with the vague topology, namely the weakest topology such that for every $f \in \mathscr{C}_c(\mathscr{S})$ (the space of real continuous functions on $\mathscr{S}$ with compact support), the function $\nu \mapsto \langle f, \nu\rangle = \int_{\mathscr{S}} f(u)\,\nu(du)$, $\nu \in \Sigma(\mathscr{S})$ is continuous. This topology can be metrized such that $\Sigma(\mathscr{S})$ is a Polish space. For details, see Sect. A.4.1.

Definition 8.11 Fix $T \in (0, \infty)$, let $\mathscr{X}$ be a locally compact Polish space, and let $\mathscr{X}_T = [0, T] \times \mathscr{X}$. Let $(\Omega, \mathscr{F}, P)$ be a probability space with a filtration $\{\mathscr{F}_t\}_{0\leq t\leq T}$. Consider any measure $\nu \in \Sigma(\mathscr{X})$ and let $\nu_T = \lambda_T \times \nu$, where λ_T is Lebesgue measure on $[0, T]$. Then an $\mathscr{F}_t$-Poisson random measure with intensity measure ν_T is a measurable mapping N from Ω into $\Sigma(\mathscr{X}_T)$ such that the following properties hold.

- For every $t \in [0, T]$ and every Borel subset $A \subset [0, t] \times \mathscr{X}$, $N(A)$ is $\mathscr{F}_t$-measurable.
- For every $t \in [0, T]$ and every Borel subset $A \subset (t, T] \times \mathscr{X}$, $N(A)$ is independent of $\mathscr{F}_t$.
- If $k \in \mathbb{N}$ and $A_i \in \mathscr{B}(\mathscr{X}_T)$, $i = 1, \ldots, k$, are such that $A_i \cap A_j = \emptyset$ for $i \neq j$ and $\nu_T(A_i) < \infty$, then $N(A_1), \ldots, N(A_k)$ are mutually independent Poisson random variables with parameters $\nu_T(A_1), \ldots, \nu_T(A_k)$.

As with the case of functionals of a Wiener process, it is convenient to first discuss representations with respect to the canonical filtration. Thus we let $\mathbb{M} \doteq \Sigma(\mathscr{X}_T)$ and let P denote the unique probability measure on $(\mathbb{M}, \mathscr{B}(\mathbb{M}))$ under which the canonical map, $N : \mathbb{M} \to \mathbb{M}$, $N(m) \doteq m$, is a Poisson random measure with intensity measure ν_T. With applications to large deviations in mind, we also consider, for $\theta > 0$, the analogous probability measures P_θ on $(\mathbb{M}, \mathscr{B}(\mathbb{M}))$ under which N is a Poisson random measure with intensity $\theta\nu_T$. (In contrast to the case of a Wiener process, there is no simple transformation of a PRM with intensity ν_T that produces a PRM with intensity $\theta\nu_T$, $\theta \neq 1$.) The corresponding expectation operators will be denoted by E and E_θ, respectively. At the end of this section we state the representation for a general filtration.

We will obtain representations for $-\log E_\theta \exp\{-G(N)\}$, where $G \in \mathscr{M}_b(\mathbb{M})$, in terms of a "controlled" Poisson random measure constructed on a larger space. We now describe this construction. Let $\mathscr{Y} \doteq \mathscr{X} \times [0, \infty)$ and $\mathscr{Y}_T \doteq [0, T] \times \mathscr{Y}$. Let $\bar{\mathbb{M}} \doteq \Sigma(\mathscr{Y}_T)$ and let $\bar{P}$ be the unique probability measure on $(\bar{\mathbb{M}}, \mathscr{B}(\bar{\mathbb{M}}))$ such that the canonical map, $\bar{N} : \bar{\mathbb{M}} \to \bar{\mathbb{M}}$, $\bar{N}(m) \doteq m$, is a Poisson random measure with intensity measure $\bar{\nu}_T \doteq \lambda_T \times \nu \times \lambda_\infty$, where λ_∞ is Lebesgue measure on $[0, \infty)$. The corresponding expectation operator will be denoted by $\bar{E}$. The control will act through this additional component of the underlying point space.

Let $\mathscr{G}_t$ denote the augmentation of $\sigma\left\{\bar{N}((0, s] \times A) : 0 \le s \le t, A \in \mathscr{B}(\mathscr{Y})\right\}$ with all $\bar{P}$ null sets in $\mathscr{B}(\bar{\mathbb{M}})$, and denote by $\mathscr{P}\mathscr{F}$ the predictable σ-field on $[0, T] \times \bar{\mathbb{M}}$ with the filtration $\{\mathscr{G}_t\}_{0 \le t \le T}$ on $(\bar{\mathbb{M}}, \mathscr{B}(\bar{\mathbb{M}}))$. Let $\mathscr{A}$ be the class of all maps $\varphi : \mathscr{X}_T \times \bar{\mathbb{M}} \to [0, \infty)$ that are $(\mathscr{P}\mathscr{F} \otimes \mathscr{B}(\mathscr{X}))\backslash\mathscr{B}[0, \infty)$ measurable. [Note that there is a slight inconsistency in the notation, since $\mathscr{P}\mathscr{F}$ concerns t, ω, while $\mathscr{B}(\mathscr{X})$ concerns x, but we write them in the order (t, x, ω).] Since $\bar{\mathbb{M}}$ is the underlying probability space, following standard convention, we will at times suppress the dependence of $\varphi(t, x, \omega)$ on ω, $(t, x, \omega) \in \mathscr{X}_T \times \bar{\mathbb{M}}$, and merely write $\varphi(t, x)$. For $\varphi \in \mathscr{A}$, define a counting process N^φ on $\mathscr{X}_T$ by setting

$$N^\varphi((0, t] \times U) \doteq \int_{(0,t]\times U} \int_{(0,\infty)} 1_{[0,\varphi(s,x)]}(r) \bar{N}(ds \times dx \times dr) \tag{8.16}$$

for all $t \in [0, T]$, $U \in \mathscr{B}(\mathscr{X})$. Here N^φ is to be thought of as a controlled random measure, with $\varphi(s, x)$ selecting the intensity for the points at location x and time s, in a possibly random but nonanticipating way. Figure 8.1 illustrates how, for some particular value x, the control modulates the jump rate by "thinning", i.e., keeping only the jumps corresponding to atoms of $\bar{N}$ that lie below $\varphi(t, x)$ at time t.

When $\varphi(s, x, \omega) = \theta$ for all $(s, x, \omega) \in \mathscr{X}_T \times \bar{\mathbb{M}}$ and some $\theta > 0$, we write N^φ as N^θ. Note that N^θ has the same distribution on $\bar{\mathbb{M}}$ with respect to $\bar{P}$ as N has on $\mathbb{M}$ with respect to P_θ. Therefore, N^θ plays the role of N on $\bar{\mathbb{M}}$.

Define $\ell : [0, \infty) \to [0, \infty)$ by

$$\ell(r) \doteq r \log r - r + 1, \ r \in [0, \infty),$$

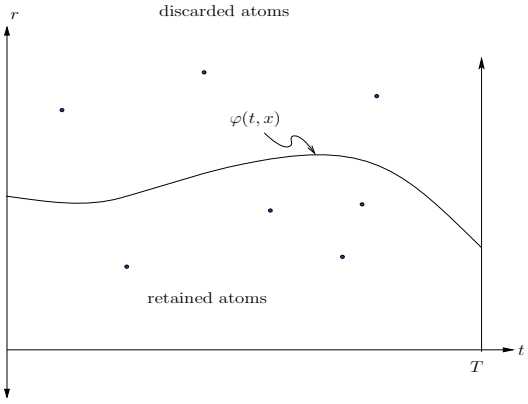

Fig. 8.1 Thinning by the control for a particular x

with the convention that $0 \log 0 = 0$. As is well known, ℓ is the local rate function (see Sect. 4.3 for this terminology) for a scaled standard Poisson process, and so it is not surprising that it plays a key role in our analysis. For $\varphi \in \mathscr{A}$, define a $[0, \infty]$-valued random variable $L_T(\varphi)$ by

$$L_T(\varphi)(\omega) \doteq \int_{\mathscr{X}_T} \ell(\varphi(t,x,\omega))\, \nu_T(dt \times dx), \ \ \omega \in \bar{\mathbb{M}}. \tag{8.17}$$

As in the setting of a Wiener process, it is convenient that the representation hold with a more restrictive class of controls. Let $\{K_n\}_{n\in\mathbb{N}}$ be an increasing sequence of compact subsets of $\mathscr{X}$ such that $\cup_{n=1}^{\infty} K_n = \mathscr{X}$. For each $M \in (0,\infty)$, let

$$\begin{aligned}\mathscr{A}_{b,M} \doteq \{\varphi \in \mathscr{A} : L_T(\varphi) \le M \text{ a.e. and for some } n \in \mathbb{N},\ n \ge \varphi(t,x,\omega) \ge 1/n \\ \text{and } \varphi(t,x,\omega) = 1 \text{ if } x \in K_n^c, \text{ for all } (t,\omega) \in [0,T] \times \bar{\mathbb{M}}\},\end{aligned} \tag{8.18}$$

and let

$$\mathscr{A}_b \doteq \cup_{M=1}^{\infty} \mathscr{A}_{b,M}. \tag{8.19}$$

As before, we let $\bar{\mathscr{A}}_{b,M}$, $\bar{\mathscr{A}}$, and $\bar{\mathscr{A}}_b$ denote the analogous spaces of controls when the canonical filtration $\{\mathscr{G}_t\}_{0\le t\le T}$ is replaced by a filtration $\{\mathscr{F}_t\}_{0\le t\le T}$ with the property that $\bar{N}$ is an $\mathscr{F}_t$-PRM with the same intensity.

The following is the representation theorem for PRM. The first equality holds because N under E_θ has the same distribution as N^θ under $\bar{E}$, as was discussed below (8.16).

Theorem 8.12 *Let $G \in \mathscr{M}_b(\mathbb{M})$. Then for $\theta > 0$,*

$$\begin{aligned}-\log E_\theta \exp\{-G(N)\} &= -\log \bar{E} \exp\{-G(N^\theta)\} \\ &= \inf_{\varphi\in\mathscr{R}} \bar{E}\left[\theta L_T(\varphi) + G(N^{\theta\varphi})\right],\end{aligned} \tag{8.20}$$

where $\mathscr{R}$ can be either $\mathscr{A}_b$, $\mathscr{A}$, $\bar{\mathscr{A}}_b$ or $\bar{\mathscr{A}}$.

The following is the analogue of Theorem 8.4 for the Poisson noise case.

Theorem 8.13 *Let $G \in \mathscr{M}_b(\mathbb{M})$ and let $\delta > 0$. Then there exist $M < \infty$ depending on $\|G\|_\infty$ and δ such that for all $\varepsilon \in (0,1)$,*

$$-\varepsilon \log \bar{E} \exp\left\{-\frac{1}{\varepsilon} G(\varepsilon N^{1/\varepsilon})\right\} \geq \inf_{\varphi \in \mathscr{A}_{b,M}} \bar{E}\left[L_T(\varphi) + G(\varepsilon N^{\varphi/\varepsilon})\right] - \delta.$$

Proof Let $\delta > 0$. Using Theorem 8.12 with $\mathscr{R} = \mathscr{A}_b$, we can find, for each $\varepsilon \in (0,1)$, $\tilde{\varphi}^\varepsilon \in \mathscr{A}_b$ such that

$$-\varepsilon \log \bar{E} \exp\left\{-\frac{1}{\varepsilon} G(\varepsilon N^{1/\varepsilon})\right\} \geq \bar{E}\left[L_T(\tilde{\varphi}^\varepsilon) + G(\varepsilon N^{\tilde{\varphi}^\varepsilon/\varepsilon})\right] - \delta/2.$$

From the boundedness of G, we obtain

$$\sup_{\varepsilon \in (0,1)} \bar{E}\left[L_T(\tilde{\varphi}^\varepsilon)\right] \leq C_G \doteq (2\|G\|_\infty + 1).$$

For $M \in \mathbb{N}$, let

$$\tau_M^\varepsilon(\omega) = \inf\left[t \in [0,T] : \int_{[0,t]\times\mathscr{X}} \ell(\tilde{\varphi}^\varepsilon(s,x,\omega))\, \nu_T(ds \times dx) \geq M\right] \wedge T.$$

Note that

$$\varphi^\varepsilon(s,x) \doteq 1 + (\tilde{\varphi}^\varepsilon(s,x) - 1)\, 1_{[0,\tau_M^\varepsilon]}(s), \ \ (s,x) \in \mathscr{X}_T$$

is an element of $\mathscr{A}_{b,M}$. Also,

$$\begin{aligned}
&\bar{E}\left[L_T(\tilde{\varphi}^\varepsilon) + G(\varepsilon N^{\tilde{\varphi}^\varepsilon/\varepsilon})\right] \\
&\quad \geq \bar{E}\left[L_T(\varphi^\varepsilon) + G(\varepsilon N^{\varphi^\varepsilon/\varepsilon})\right] + \bar{E}\left[G(\varepsilon N^{\tilde{\varphi}^\varepsilon/\varepsilon}) - G(\varepsilon N^{\varphi^\varepsilon/\varepsilon})\right].
\end{aligned}$$

By Chebyshev's inequality,

$$\bar{E}\left|G(\varepsilon N^{\tilde{\varphi}^\varepsilon/\varepsilon}) - G(\varepsilon N^{\varphi^\varepsilon/\varepsilon})\right| \leq 2\|G\|_\infty \bar{P}\left\{\tau_M^\varepsilon < T\right\} \leq 2\|G\|_\infty \frac{C_G}{M}.$$

Let $M = (2\|G\|_\infty C_G + 1)/\delta$. Then for all $\varepsilon \in (0,1)$,

$$-\varepsilon \log \bar{E} \exp\left\{-\frac{1}{\varepsilon} G(\varepsilon N^{1/\varepsilon})\right\} \geq \bar{E}\left[L_T(\varphi^\varepsilon) + G(\varepsilon N^{\varphi^\varepsilon/\varepsilon})\right] - \delta,$$

as desired. □

Remark 8.14 We note that Theorem 3.23 is a special case of Theorems 8.12 and 8.13. To see this, consider the case $\mathscr{X} \doteq \{0\}$ and $\nu \doteq \delta_0$. Define $\gamma : \mathbb{M} \to \mathscr{D}([0,T] : \mathbb{R})$ by $\gamma(m)(t) \doteq m((0,t] \times \{0\})$, $t \in [0,T]$, $m \in \mathbb{M}$. Then γ is a Borel measurable map, and thus by Theorem 8.12, for every bounded Borel measurable function G mapping $\mathscr{D}([0,T] : \mathbb{R})$ to $\mathbb{R}$ and $\theta \in (0,\infty)$, we have

$$-\log \bar{E} \exp\{-G \circ \gamma(N^\theta)\} = \inf_{\varphi \in \mathscr{A}} \bar{E}\left[\theta L_T(\varphi) + G \circ \gamma(N^{\theta\varphi})\right]. \tag{8.21}$$

Recalling the definition of $\mathscr{X}$ and ν, any $\varphi \in \mathscr{A}$ can be identified with a nonnegative predictable process, and $\gamma(N^{\theta\varphi})$ is a controlled Poisson process in the sense of Sect. 3.3; in particular, $\gamma(N^\theta)$ is a Poisson process with rate θ (denoted in Sect. 3.3 by N^θ). Thus the first representation in Theorem 3.23 follows readily from (8.21) with $(\bar{\Omega}, \bar{\mathscr{F}}, \bar{P}, \{\bar{\mathscr{F}}_t\}) = (\bar{M}, \mathscr{B}(\bar{M}), \bar{P}, \{\mathscr{G}_t\})$. The second representation in Theorem 3.23 follows similarly from Theorem 8.13.

The rest of this section is devoted to the proof of Theorem 8.12. For notational convenience we provide details only for the case $\theta = 1$. The general case is treated similarly. As in the case of Brownian motion, we first consider $\mathscr{A}_b$ and $\mathscr{A}$, and then extend to $\bar{\mathscr{A}}_b$ and $\bar{\mathscr{A}}$.

8.2.2 *Preparatory Results*

Recall that $\mathscr{P}\mathscr{F}$ denotes the predictable σ-field associated with the augmented PRM $\bar{N}$, and that $\mathscr{Y} = \mathscr{X} \times [0,\infty)$. A class of processes that will be used as test functions is defined as follows. Let $\hat{A}_b$ be the set of all $(\mathscr{P}\mathscr{F} \otimes \mathscr{B}(\mathscr{Y}))\backslash\mathscr{B}(\mathbb{R})$-measurable maps $\vartheta : \mathscr{Y}_T \times \bar{\mathbb{M}} \to \mathbb{R}$ that are bounded and such that for some compact $K \subset \mathscr{Y}$, $\vartheta(s,x,r,\omega) = 0$ whenever $(x,r) \in K^c$. Once again ω will usually be suppressed in the notation. The following result is standard (see, e.g., Theorem III.3.24 of [161]), and the analogue with $\mathscr{A}_b$ replaced by $\bar{\mathscr{A}}_b$ and $\mathscr{G}_t$ by $\mathscr{F}_t$ also holds. Let N_c^1 be the compensated version of N^1, which is defined by $N_c^1(A) \doteq N^1(A) - \nu_T(A)$ for all $A \in \mathscr{B}(\mathscr{X}_T)$ such that $\nu_T(A) < \infty$.

Theorem 8.15 (GIRSANOV) *Let $\varphi \in \mathscr{A}_b$. Then*

$$\begin{aligned}\mathscr{E}^\varphi(t) &\doteq \exp\left\{\int_{(0,t]\times\mathscr{X}} \log(\varphi(s,x)) N_c^1(ds \times dx) \right. \\ &\qquad \left. + \int_{(0,t]\times\mathscr{X}} (\log(\varphi(s,x)) - \varphi(s,x) + 1)\, \nu_T(ds \times dx)\right\} \\ &= \exp\left\{\int_{(0,t]\times\mathscr{X}\times[0,1]} \log(\varphi(s,x)) N(ds \times dx \times dr) \right. \\ &\qquad \left. + \int_{(0,t]\times\mathscr{X}\times[0,1]} (-\varphi(s,x) + 1)\, \bar{\nu}_T(ds \times dx \times dr)\right\}\end{aligned} \tag{8.22}$$

is a $\mathscr{G}_t$-martingale. Define a probability measure $\bar{Q}^\varphi$ on $\bar{\mathbb{M}}$ by

$$\bar{Q}^\varphi(H) = \int_H \mathscr{E}^\varphi(T) d\bar{P} \text{ for } H \in \mathscr{B}(\bar{\mathbb{M}}),$$

and let $\bar{E}^\varphi$ denote integration with respect to $\bar{Q}^\varphi$. Then for every $\vartheta \in \hat{A}_b$,

$$\begin{aligned}&\bar{E}^\varphi \int_{\mathscr{Y}_T} \vartheta(s,x,r)\bar{N}(ds \times dx \times dr)\\ &\quad = \bar{E}^\varphi \int_{\mathscr{Y}_T} \vartheta(s,x,r)\left[\varphi(s,x)1_{(0,1]}(r) + 1_{(1,\infty)}(r)\right]\bar{\nu}_T(ds \times dx \times dr).\end{aligned}$$

The last statement in the lemma says that under $\bar{Q}^\varphi$, $\bar{N}$ is a random counting measure with compensator $\left[\varphi(s,x)1_{(0,1]}(r) + 1_{(1,\infty)}(r)\right]\bar{\nu}_T(ds \times dx \times dr)$.

Recall that $\mathscr{X} = \cup_{n=1}^\infty K_n$ for increasing compact sets K_n. A process $\varphi \in \mathscr{A}_{b,M}$ is in the set $\mathscr{A}_{s,M}$ if the following holds. There exist $n, \ell, n_1, \ldots, n_\ell \in \mathbb{N}$; a partition $0 = t_0 < t_1 < \cdots < t_\ell = T$; for each $i = 1, \ldots, \ell$ a disjoint measurable partition E_{ij} of K_n, $j = 1, \ldots, n_i$; $\mathscr{G}_{t_{i-1}}$-measurable random variables X_{ij}, $i = 1, \ldots, \ell$, $j = 1, \ldots, n_i$, such that $1/n \le X_{ij} \le n$; and

$$\varphi(t,x,\bar{m}) = 1_{\{0\}}(t) + \sum_{i=1}^{\ell}\sum_{j=1}^{n_i} 1_{(t_{i-1},t_i]}(t)X_{ij}(\bar{m})1_{E_{ij}}(x) + 1_{K_n^c}(x)1_{(0,T]}(t). \quad (8.23)$$

We let $\mathscr{A}_s \doteq \cup_{M=1}^\infty \mathscr{A}_{s,M}$ and refer to elements in $\mathscr{A}_s$ as *simple* processes.

Lemma 8.16 *Let $\varphi \in \mathscr{A}_b$. Then there exists a sequence of processes $\varphi_k \in \mathscr{A}_s$ with the following properties.*

(a) *N^{φ_k} converges in distribution to N^φ as $k \to \infty$.*

(b) *$\bar{E}|L_T(\varphi_k) - L_T(\varphi)| \to 0$ and $\bar{E}|\mathscr{E}^{\varphi_k}(T) - \mathscr{E}^\varphi(T)| \to 0$, as $k \to \infty$.*

Proof We first construct processes φ_k that satisfy parts (a) and (b) of the lemma but that instead of being simple are continuous in t. Since $\varphi \in \mathscr{A}_b$, we have for some $n \in \mathbb{N}$ that $n \ge \varphi(t,x,\omega) \ge 1/n$ and $\varphi(t,x,\omega) = 1$ if $x \in K_n^c$ for all $(t,\omega) \in [0,T] \times \bar{\mathbb{M}}$. For $k \in \mathbb{N}$, define

$$\varphi_k(t,x,\omega) = k\left(\frac{1}{k} - t\right)^+ + k\int_{(t-\frac{1}{k})^+}^t \varphi(s,x,\omega)ds, \ \ (t,x,\omega) \in \mathscr{X}_T \times \bar{\mathbb{M}}.$$

An application of Lusin's theorem gives that for $\nu \times \bar{P}$-a.e. (x,ω), as $k \to \infty$,

$$\begin{aligned}&\int_{[0,T]} |\varphi_k(t,x,\omega) - \varphi(t,x,\omega)|dt \to 0\\ &\int_{[0,T]} |\ell(\varphi_k(t,x,\omega)) - \ell(\varphi(t,x,\omega))|dt \to 0.\end{aligned} \quad (8.24)$$

In particular, $\varphi_k \in \mathscr{A}_b$ for every k and $\bar{E}|L_T(\varphi_k) - L_T(\varphi)| \to 0$, as $k \to \infty$. It follows from (D.6) and the definition of the controlled PRM that for $g \in \mathscr{C}_c(\mathscr{X}_T)$,

$$\begin{aligned}
&\bar{E}\,|\langle g, N^{\varphi_k}\rangle - \langle g, N^{\varphi}\rangle| \\
&\quad \le \bar{E}\int_{\mathscr{Y}_T} |g(s,x)||1_{[0,\varphi_k(s,x,\omega)]}(r) - 1_{[0,\varphi(s,x,\omega)]}(r)|\,\bar{\nu}_T(ds\times dx\times dr) \\
&\quad \le \|g\|_\infty\,\bar{E}\int_{[0,T]\times K_n} |\varphi_k(s,x,\omega) - \varphi(s,x,\omega)|\,\nu_T(ds\times dx\,).
\end{aligned}$$

Using (8.24), $\nu(K_n) < \infty$, and the uniform bounds on φ_k and φ shows that the last quantity approaches 0 as $k \to \infty$, and hence by Lemma A.10, $N^{\varphi_k} \Rightarrow N^{\varphi}$.

Next we consider the convergence of $\mathscr{E}^{\varphi_k}(T)$ in $\mathscr{L}^1(\bar{P})$. By Scheffe's lemma [249, Sect. 5.11], if $f_k(\omega)$ and $f(\omega)$ are densities with respect to $\bar{P}$ such that $f_k \to f$ in probability, then the convergence is also in $\mathscr{L}^1(\bar{P})$. Thus it suffices to show that

$$\mathscr{E}^{\varphi_k}(T) \to \mathscr{E}^{\varphi}(T) \text{ in } \bar{P}\text{-probability.} \tag{8.25}$$

For this, it is enough to show [see (8.22)] that

$$\int_{\mathscr{X}_T} (1-\varphi_k(s,x))\,\nu_T(ds\times dx\,) \to \int_{\mathscr{X}_T} (1-\varphi(s,x))\,\nu_T(ds\times dx\,)$$

and since $N^1 = N^1_c + \nu_T$, that

$$\int_{\mathscr{X}_T} \log(\varphi_k(s,x))N^1(ds\times dx) \to \int_{\mathscr{X}_T} \log(\varphi(s,x))N^1(ds\times dx)$$

in probability as $k \to \infty$. The first convergence is immediate from (8.24), the uniform bounds on $\varphi_k, \varphi, \nu(K_n) < \infty$, and the fact that $1-\varphi_k(s,x) = 1-\varphi(s,x) = 0$ for $x \notin K_n$. The second convergence follows similarly on noting that $\varphi_k(s,x) \wedge \varphi(s,x) \ge 1/n$ implies

$$|\log(\varphi_k(s,x)) - \log(\varphi(s,x))| \le n|\varphi_k(s,x) - \varphi(s,x)|.$$

This proves (8.25) and so $\bar{E}|\mathscr{E}^{\varphi_k}(T) - \mathscr{E}^{\varphi}(T)| \to 0$ as $k \to \infty$. This completes the construction of φ_k that satisfy parts (a) and (b) of the lemma.

Next we show that the processes can be assumed to be simple. Note that by construction, $t \mapsto \varphi_k(t,x,\omega)$ is continuous for $\nu \times \bar{P}$-a.e. (x,ω). Consider any φ_k as constructed previously, and to simplify the notation, drop the k subscript. Two more levels of approximation will be used, and indexed by q and r. Thus for the fixed φ and $q \in \mathbb{N}$, define

$$\varphi_q(t,x,\omega) = \sum_{m=0}^{\lfloor qT\rfloor} \varphi\left(\frac{m}{q}, x, \omega\right) 1_{(\frac{m}{q}, \frac{m+1}{q}]}(t), \quad (t,x,\omega) \in \mathscr{X}_T \times \bar{\mathbb{M}}.$$

It is easily checked that (8.24) is satisfied as $q \to \infty$, and so arguing as above, the sequence $\{\varphi_q\}$ satisfies parts (a) and (b) of the lemma. Note that for fixed q and m, $g(x,\omega) = \varphi(m/q, x, \omega)$ is a $\mathscr{B}(\mathscr{X})\otimes\bar{\mathscr{F}}_{m/q}$-measurable map with values in $[1/n, n]$ and $g(x,\omega) = 1$ for $x \in K_n^c$. By a standard approximation procedure one can find $\mathscr{B}(\mathscr{X})\otimes\bar{\mathscr{F}}_{m/q}$-measurable maps g_r, $r \in \mathbb{N}$ with the following properties:

$$g_r(x,\omega) = \sum_{j=1}^{a(r)} c_j^r(\omega) 1_{E_j^r}(x) \text{ for } x \in K_n,$$

where for each r, $a(n) \in \mathbb{N}$, $\{E_j^r\}_{j=1,\dots,a(r)}$ is some measurable partition of K_n, and for all j, r, $c_j^r(\omega) \in [1/n, n]$ a.s.; $g_r(x,\omega) = 1$ for $x \in K_n^c$; $g_r \to g$ a.s. $\nu \times \bar{P}$. If we make such an approximation for each m and label the process obtained when the g's are replaced by g_r's as φ_q^r, then $\varphi_q^r \in \mathscr{A}_s$. Hence by the triangle inequality we can find a sequence $\{\varphi_k\} \subset \mathscr{A}_s$ such that (a) and (b) hold. □

A last result is needed before we can prove the main theorem. As in the case of the Wiener process, we need to know that controls under the original probability space can be replicated on a new space. The proof of the lemma, which uses an elementary but detailed argument, is given at the end of the chapter.

Lemma 8.17 *For every $\varphi \in \mathscr{A}_s$ there is $\hat{\varphi} \in \mathscr{A}_s$ such that $(L_T(\hat{\varphi}), N^1)$ has the same distribution under $\bar{Q}^{\hat{\varphi}}$ as $(L_T(\varphi), N^\varphi)$ does under $\bar{P}$.*

We now proceed to the proof of Theorem 8.12. We will provide the proof only for the case that $\mathscr{R}$ is $\mathscr{A}$ or $\mathscr{A}_b$. The general filtration setting, i.e., when $\mathscr{R} = \bar{\mathscr{A}}, \bar{\mathscr{A}}_b$, can be treated as in Sect. 8.1.5.

8.2.3 *Proof of the Upper Bound in the Representation*

In this subsection we prove (recall that we present the argument only for $\theta = 1$)

$$\begin{aligned} -\log E_1 \exp\{-G(N)\} &= -\log \bar{E} \exp\{-G(N^1)\} \\ &\le \inf_{\varphi\in\mathscr{A}} \bar{E}\left[L_T(\varphi) + G(N^\varphi)\right]. \end{aligned} \tag{8.26}$$

Note that this automatically gives the corresponding bound for the smaller class $\mathscr{A}_b$ in (8.20).

The proof parallels that of the case of a Wiener process. Let $h : \bar{\mathbb{M}} \to \mathbb{M}$ be defined by

$$h(\bar{m})((0,t]\times U) = \int_{(0,t]\times U\times(0,\infty)} 1_{[0,1]}(r)\bar{m}(ds\times dx\times dr)$$

for $t \in [0,T]$, $U \in \mathscr{B}(\mathscr{X})$. Thus $N^1 = h(\bar{N})$. Recalling (2.1), we have

$$
\begin{aligned}
-\log \bar{E}\exp\{-G(N^1)\} &= -\log \int_{\bar{\mathbb{M}}} \exp\{-G(h(\bar{m}))\}\bar{P}(d\bar{m}) \\
&= \inf_{\bar{Q}\in\mathscr{P}(\bar{\mathbb{M}})} \left[R\left(\bar{Q} \,\|\, \bar{P}\right) + \int_{\bar{\mathbb{M}}} G(h(\bar{m}))\bar{Q}(d\bar{m}) \right]. \qquad (8.27)
\end{aligned}
$$

We begin by evaluating $R\left(\bar{Q}^{\varphi} \,\|\, \bar{P}\right)$ for $\varphi \in \mathscr{A}_b$. By Theorem 8.15, $\{\mathscr{E}^{\varphi}(t)\}$ [defined in (8.22)] is an $\mathscr{G}_t$-martingale, and under $\bar{Q}^{\varphi}$, $\bar{N}$, it is a random counting measure with compensator $\left[\varphi(s,x)1_{(0,1]}(r) + 1_{(1,\infty)}(r)\right]\bar{\nu}_T(ds \times dx \times dr)$. It follows from the definition of relative entropy and L_T in (8.17) that

$$
\begin{aligned}
&R\left(\bar{Q}^{\varphi} \,\|\, \bar{P}\right) \\
&\quad = \bar{E}^{\varphi}\left[\int_{\mathscr{X}_T} \log(\varphi(s,x)) N_c^1(ds \times dx) \right. \\
&\qquad\qquad \left. + \int_{\mathscr{X}_T} (\log(\varphi(s,x)) - \varphi(s,x) + 1)\, \nu_T(ds \times dx) \right] \\
&\quad = \bar{E}^{\varphi}\left[\int_{\mathscr{X}_T} \log(\varphi(s,x)) N^1(ds \times dx) + \int_{\mathscr{X}_T} (-\varphi(s,x) + 1)\, \nu_T(ds \times dx) \right] \\
&\quad = \bar{E}^{\varphi}\left[\int_{\mathscr{X}_T} (\varphi(s,x)\log(\varphi(s,x)) - \varphi(s,x) + 1)\, \nu_T(ds \times dx) \right] \\
&\quad = \bar{E}^{\varphi} L_T(\varphi). \qquad (8.28)
\end{aligned}
$$

Thus by (8.27), for $\varphi \in \mathscr{A}_b$, we have

$$
\begin{aligned}
-\log \bar{E}\exp\{-G(N^1)\} &\le \left[R\left(\bar{Q}^{\varphi} \,\|\, \bar{P}\right) + \int_{\bar{\mathbb{M}}} G(h(\bar{m}))\bar{Q}^{\varphi}(d\bar{m}) \right] \qquad (8.29) \\
&= \bar{E}^{\varphi}\left[L_T(\varphi) + G(N^1) \right].
\end{aligned}
$$

The rest of the proof is in three steps.

Step 1. Simple φ. Suppose one is given $\varphi \in \mathscr{A}_s$. According to Lemma 8.17, one can find $\tilde{\varphi}$ that is $\mathscr{G}_t$-predictable and simple and such that $(\tilde{\varphi}, N^1)$ under $\bar{Q}^{\tilde{\varphi}}$ has the same distribution as (φ, N^{φ}) under $\bar{P}$. This implies

$$
\bar{E}^{\tilde{\varphi}}\left[L_T(\tilde{\varphi}) + G(N^1) \right] = \bar{E}\left[L_T(\varphi) + G(N^{\varphi}) \right],
$$

and thus the desired inequality follows directly from (8.29).

Step 2. Bounded φ. Given $\varphi \in \mathscr{A}_b$, let $\varphi_k \in \mathscr{A}_s$ be the sequence constructed in Lemma 8.16. By Step 1, for every $k \in \mathbb{N}$,

$$
-\log \bar{E}\exp\{-G(N^1)\} \le \bar{E}\left[L_T(\varphi_k) + G(N^{\varphi_k}) \right]. \qquad (8.30)
$$

From Lemma 8.16, under $\bar{P}$, we have $N^{\varphi_k} \Rightarrow N^{\varphi}$, and $\bar{E}\left[L_T(\varphi_k)\right] \to \bar{E}\left[L_T(\varphi)\right]$. However, G is not assumed continuous, and so we cannot simply pass to the limit

in the last display. Instead, we will apply Lemma 2.5, which requires bounds on relative entropies. The first and the last equalities in the following display follow from Lemma 8.17, the second equality is a consequence of (8.28), and the inequality follows from the fact that relative entropy can only decrease when one is considering measures induced by the same mapping (in this case the random variable N^1) [see part (f) of Lemma 2.4]:

$$\begin{aligned} R(\bar{P} \circ (N^{\varphi_k})^{-1} \,\|\, \bar{P} \circ (N^1)^{-1}) &= R(\bar{Q}^{\tilde{\varphi}_k} \circ (N^1)^{-1} \,\|\, \bar{P} \circ (N^1)^{-1}) \\ &\le R(\bar{Q}^{\tilde{\varphi}_k} \,\|\, \bar{P}) \\ &= \bar{E}^{\tilde{\varphi}_k} \left[L_T(\tilde{\varphi}_k) \right] \\ &= \bar{E} \left[L_T(\varphi_k) \right]. \end{aligned} \tag{8.31}$$

Since $\bar{E}\left[L_T(\varphi_k)\right] \to \bar{E}\left[L_T(\varphi)\right] < \infty$, the relative entropies in (8.31) are uniformly bounded in k. By Lemma 8.16 we can pass to the limit in (8.30) and obtain (8.26) when $\mathscr{A}$ is replaced by $\mathscr{A}_b$. For future use, note that the lower semicontinuity of relative entropy and (8.31) imply

$$R(\bar{P} \circ (N^{\varphi})^{-1} \,\|\, \bar{P} \circ (N^1)^{-1}) \le \bar{E}\left[L_T(\varphi)\right] \text{ for } \varphi \in \mathscr{A}_b.$$

Step 3. General φ. For $\varphi \in \mathscr{A}$, define

$$\varphi_n(t,x,\omega) = \begin{cases} [\varphi(t,x,\omega) \vee (1/n)] \wedge n, & x \in K_n, \\ 1 & \text{otherwise.} \end{cases}$$

Note that $\varphi_n \in \mathscr{A}_b$, and so (8.30) holds with φ_k replaced by φ_n. Since the definition of φ_n implies that $\ell(\varphi_n(x,t,\omega))$ is nondecreasing in n, by the monotone convergence theorem, we have $\bar{E}L_T(\varphi_n) \uparrow \bar{E}L_T(\varphi)$. If $\bar{E}L_T(\varphi) = \infty$, there is nothing to prove. Assume therefore that

$$\bar{E}L_T(\varphi) < \infty. \tag{8.32}$$

Then $R(\bar{P} \circ (N^{\varphi_n})^{-1} \,\|\, \bar{P} \circ (N^1)^{-1}) \le \bar{E}L_T(\varphi_n) \le \bar{E}L_T(\varphi)$. We claim that N^{φ_n} converges in distribution to N^{φ}. If true, then using the uniform bound on relative entropies just noted, we can once again apply Lemma 2.5, pass to the limit on n, and thereby obtain (8.26).

Let $g \in \mathscr{C}_c(\mathscr{X}_T)$ and let n_0 be large enough that the support of g is contained in $[0,T] \times K_{n_0}$. Then for all $n \ge n_0$,

$$\bar{E}\,|\langle g, N^{\varphi_n}\rangle - \langle g, N^{\varphi}\rangle| \le \|g\|_\infty \, \bar{E} \int_{[0,T]\times K_{n_0}} \left(\frac{1}{n} + (\varphi(s,x) - n)^+ \right) \nu_T(ds \times dx).$$

Note that $\nu_T([0,T] \times K_{n_0}) < \infty$, $(\varphi(t,x) - n)^+ \to 0$ as $n \to \infty$, and $(\varphi(t,x) - n)^+ \le \ell(\varphi(t,x))$. These observations together with (8.32) show that the

right-hand side in the last display approaches 0 as $n \to \infty$. We can therefore apply Lemma A.10, and $N^{\varphi_n} \Rightarrow N^{\varphi}$ follows. □

8.2.4 *Proof of the Lower Bound in the Representation*

In this subsection we prove (again only for $\theta = 1$)

$$-\log E_1 \exp\{-G(N)\} = -\log \bar{E} \exp\{-G(N^1)\} \geq \inf_{\varphi \in \mathscr{A}_b} \bar{E}\left[L_T(\varphi) + G(N^{\varphi})\right], \tag{8.33}$$

which automatically gives the lower bound for the larger class $\mathscr{A}$ in (8.20). As in the Brownian motion case, the proof is in two steps.

Step 1. G of a particular form. Let $K, M \in \mathbb{N}$ be arbitrary, and consider any collection $0 = t_1 < t_2 < \cdots < T_K = T$. Let $h : \mathbb{N}_0^{KM} \to \mathbb{R}$ be a bounded map. Let $C_1, \ldots, C_M$ be precompact sets in $\mathscr{X}$ such that $C_i \cap C_j = \emptyset$ if $i \neq j$. Then G is of the form

$$G(N^1) = h(n(t_1), n(t_2) - n(t_1), \ldots, n(t_K) - n(t_{K-1})), \tag{8.34}$$

where for $0 \leq t \leq T$,

$$n(t) = (n_1(t), \ldots, n_M(t)) = (N^1((0,t] \times C_1), \ldots, N^1((0,t] \times C_M)). \tag{8.35}$$

For G of this particular form, we will construct $\varphi \in \mathscr{A}_b$ such that

$$-\log \bar{E} \exp\{-G(N^1)\} = \bar{E}\left[L_T(\varphi) + G(N^{\varphi})\right],$$

from which (8.33) is immediate. The underlying idea is the same as in the Brownian case. Using a conditioning argument, over each interval of the form $[t_i, t_{i+1}]$ we can interpret the logarithm of an exponential integral as the value function of a stochastic control problem. From the boundedness of h, the integral is smooth in the time variable, which means that the control problem has a classical-sense solution, and then an optimal control can be found from this solution and the corresponding dynamic programming equation. The controls over the various intervals are concatenated to produce $\varphi \in \mathscr{A}_b$, which actually achieves the infimum for the given G. The following lemma is analogous to Lemma 8.8 for the Brownian motion case and can be proved in a similar manner. When applied, the k in the statement of the lemma will be of the form jM, $j = 0, \ldots, K-1$.

Lemma 8.18 *Let $g : \mathbb{N}_0^k \times \mathbb{N}_0^M \to \mathbb{R}$ be uniformly bounded, and let $\{n(t)\}_{0 \leq t \leq T}$ be as in (8.35). Define $V : [0,T] \times \mathbb{N}_0^k \times \mathbb{N}_0^M \to \mathbb{R}$ by*

$$V(t,z,x) \doteq -\log \bar{E} e^{-g(z, x+n(T-t))}, \ (t,z,x) \in [0,T] \times \mathbb{N}_0^k \times \mathbb{N}_0^M.$$

For $(t, z, x) \in [0, T] \times \mathbb{N}_0^k \times \mathbb{N}_0^M$, *let*

$$\partial_{x_i} V(t, z, x) \doteq V(t, z, x + e_i) - V(t, z, x),$$

where $\{e_i\}_{i=1}^M$ *is the coordinate basis in* $\mathbb{R}^M$. *Then for each* z, x *and* i, $\partial_{x_i} V(t, z, x)$ *is continuous in* $t \in [0, T]$. *Let* $\{X(z, t)\}_{0 \le t \le T}$ *with* $X(z, t) = (X_1(z, t), \ldots, X_M(z, t))$ *be the unique solution of*

$$X_i(z, t) = \int_{(0,t] \times C_i \times [0,\infty)} 1_{[0, \exp\{-\partial_{x_i} V(s, z, X(z, s-))\}]}(r) \bar{N}(ds \times dx \times dr), \tag{8.36}$$

$i = 1, \ldots, M$. *Define*

$$\varphi(t, x) = \sum_{i=1}^M \varphi_i(t) 1_{C_i}(x) + 1_{C^c}(x), \ (t, x) \in [0, T] \times \mathscr{X},$$

where $C = \cup_{i=1}^M C_i$ *and* $\varphi_i(t) = \exp\{-\partial_{x_i} V(t, z, X(z, t-))\}$. *Then*

$$-\log \bar{E} \exp\{-g(z, n(T))\} = \bar{E}\left[L_T(\varphi) + g\left(z, n^\varphi(T)\right)\right],$$

where for $t \in [0, T]$,

$$n^\varphi(t) = (n_1^\varphi(t), \ldots, n_M^\varphi(t)) = (N^\varphi((0, t] \times C_1), \ldots, N^\varphi((0, t] \times C_M)). \tag{8.37}$$

Since $t \mapsto \partial_{x_i} V(t, z, x)$ is continuous and $\nu(C_i) < \infty$ for all (t, z, x) and $i = 1, \ldots, M$, the solution to (8.36) will jump a finite number of times a.s. over $[0, T]$. The equations can be solved recursively by updating the relevant component $X_i(z, \bar{t})$, $i = 1, \ldots, M$ if a jump occurs at time $\bar{t}$, using (8.36) to identify the next time that one of the components will jump, and repeating.

We next apply Lemma 8.18 recursively. For $j = 1, \ldots, K$, define $V_j : \mathbb{N}_0^{jM} \to \mathbb{R}$ as follows: $V_K = h$ and

$$V_j(z_j) = -\log \bar{E} e^{-V_{j+1}(z_j, n(t_{j+1}) - n(t_j))}, \ z_j \in \mathbb{N}_0^{jM}, \ j = 1, \ldots, K-1.$$

By successive conditioning it is easily checked that

$$V_0 \doteq -\log \bar{E} e^{-V_1(n(t_1) - n(t_0))} = -\log E e^{-G(N)}.$$

Note that V_j is a bounded map for each j. For $j = 1, \ldots, K$, let $Z_j = (n(t_1), n(t_2) - n(t_1), \ldots, n(t_j) - n(t_{j-1}))$ and note that Z_j is a $\mathbb{N}_0^{jM}$-valued random variable. For $z_j \in \mathbb{N}_0^{jM}$ and $j = 1, \ldots, K-1$, let $\{Y(z_j, t)\}_{t \in [t_j, t_{j+1}]}$ be the unique solution of

$$Y_i(z_j, t) = \int_{(t_j, t] \times C_i \times [0,\infty)} 1_{[0, \exp\{-\partial_{x_i} V_{j+1}(s, z_j, Y(z_j, s-))\}]}(r) \bar{N}(ds \times dx \times dr)$$

for $t \in [t_j, t_{j+1}]$, where $Y(z_j, t) = (Y_1(z_j, t), \ldots, Y_M(z_j, t))$. For $i = 1, \ldots, M$, define

$$\varphi_i(t) = \exp\{-\partial_{x_i} V_{j+1}(t, Z_j, Y(Z_j, t-))\}, \ t \in [t_j, t_{j+1}), \ j = 0, \ldots, K-1$$

and

$$\varphi(t, x) = \sum_{i=1}^{M} \varphi_i(t, x) 1_{C_i}(x) + 1_{C^c}(x), \ (t, x) \in [0, T] \times \mathscr{X}.$$

Then by a recursive argument using Lemma 8.18, we see that

$$\begin{aligned} -\log E e^{-G(N)} &= \bar{E}\left[L_T(\varphi) + h(n^\varphi(t_1), \ldots, n^\varphi(t_K) - n^\varphi(t_{K-1}))\right] \\ &= \bar{E}\left[L_T(\varphi) + G(N^\varphi)\right], \end{aligned}$$

where in the first line, n^φ is defined as in (8.37). Note that by construction, $\varphi \in \mathscr{A}_b$. Thus we have proved (8.33) for all G of the form (8.34).

Step 2. G that is bounded and measurable. Now suppose that G is simply bounded and measurable. We claim that there exist functions $\{G_n\}_{n\in\mathbb{N}}$ such that for each n, G_n is of the form assumed in Step 1, $\|G_n\|_\infty \le \|G\|_\infty$, and $G_n \to G$ a.s. with respect to P. This is shown, as in the Brownian case, by noting that each of the following classes admits an approximation of this form relative to elements of the preceding class, save of course the first:

- G bounded and measurable;
- G bounded and continuous;
- G bounded and continuous and depending on N^1 only through

$$\{N^1((0, t_k] \times \cdot)\}_{k=1,\ldots,K};$$

  where $K \in \mathbb{N}$ and $0 = t_1 < t_2 < \cdots < T_K = T$ are arbitrary;
- G bounded and depending on N^1 only through

$$\{N^1((0, t_k] \times C_i)\}_{k=1,\ldots,K, i=1,\ldots,M},$$

  where $K, M \in \mathbb{N}$, $0 = t_1 < t_2 < \cdots < T_K = T$ and $C_1, \ldots, C_M$ are precompact sets in $\mathscr{X}$ such that $C_i \cap C_j = \emptyset$ if $i \ne j$.

As before, these approximations follow by standard arguments based on Theorem E.4 and the martingale convergence theorem. We now complete the lower bound in exactly the same way as in the case of Brownian motion. With each $n \in \mathbb{N}$ we can associate $\varphi_n \in \mathscr{A}_b$ such that

$$-\log E \exp\{-G_n(N)\} = \bar{E}\left[L_T(\varphi_n) + G_n\left(N^{\varphi_n}\right)\right].$$

If Q_n is the distribution induced by N^{φ_n}, then [see (8.31)]

$$R(Q_n \,\|\, P) \le \bar{E}\,[L_T(\varphi_n)] \le 2\,\|G\|_\infty .$$

Thus $\{Q_n\}$ is tight, and from part (b) of Lemma 2.5,

$$\lim_{n\to\infty} \bar{E}\,|G_n(N^{\varphi_n}) - G(N^{\varphi_n})| = 0.$$

By the dominated convergence theorem,

$$\lim_{n\to\infty} |\log E\exp\{-G_n(N)\} - \log E\exp\{-G(N)\}| = 0.$$

Therefore, given $\varepsilon > 0$, we can find $n \in \mathbb{N}$ such that

$$\begin{aligned} -\log Ee^{-G(N)} &\ge -\log Ee^{-G_n(N)} - \varepsilon \\ &= \bar{E}\left[L_T(\varphi_n) + G_n\left(N^{\varphi_n}\right)\right] - \varepsilon \\ &\ge \bar{E}\left[L_T(\varphi_n) + G\left(N^{\varphi_n}\right)\right] - 2\varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary and $v_n \in \mathscr{A}_b$, we have (8.33), completing the proof. □

8.2.5 *Construction of Equivalent Controls*

In this section we give the proof of Lemma 8.17. We need to show that given $\varphi \in \mathscr{A}_s$, there is $\hat{\varphi} \in \mathscr{A}_s$ such that the distribution of $(L_T(\hat{\varphi}), N^1)$ under $\bar{\mathbb{Q}}^{\hat{\varphi}}$ is the same as that of $(L_T(\varphi), N^\varphi)$ under $\bar{\mathbb{P}}$. Let φ be as in (8.23):

$$\varphi(t,x,\bar{m}) = 1_{\{0\}}(t) + \sum_{i=1}^{\ell}\sum_{j=1}^{n_i} 1_{(t_{i-1},t_i]}(t)X_{ij}(\bar{m})1_{E_{ij}}(x) + 1_{K_n^c}(x)1_{(0,T]}(t).$$

We will need some notation to describe how measures on $[0,T]\times\mathscr{Y}$ are decomposed into parts on subintervals of the form $(t_{i-1},t_i]$, and also how after some manipulation such quantities can be recombined. For $i = 1,\dots,\ell$, let $\mathbb{I}_i \doteq (t_{i-1},t_i]$ and let $\mathscr{Y}_i \doteq \mathbb{I}_i \times \mathscr{Y}$. Denote by $\bar{\mathbb{M}}_i$ the space of nonnegative σ-finite integer-valued measures $\bar{m}_i$ on $(\mathscr{Y}_i, \mathscr{B}(\mathscr{Y}_i))$ that satisfy $\bar{m}_i(\mathbb{I}_i \times K) < \infty$ for all compact $K \subset \mathscr{Y}$. Endow $\bar{\mathbb{M}}_i$ with the weakest topology making the functions $m \mapsto \langle f, m\rangle$, $m \in \bar{\mathbb{M}}_i$ continuous, for every f in $\mathscr{C}_c(\mathscr{Y}_i)$ vanishing outside some compact subset of $\mathscr{Y}_i$. Denote by $\bar{\mathscr{M}}_i$ the corresponding Borel σ-field. Let $\bar{N}_i$ be the $\bar{\mathbb{M}}_i$-valued random variable on $(\bar{\mathbb{M}}, \mathscr{B}(\bar{\mathbb{M}}))$ defined by $\bar{N}_i(A) \doteq \bar{N}(A)$, $A \in \mathscr{B}(\mathscr{Y}_i)$. Also, define $\mathbb{J}_i \doteq [1/n, n]^{n_i}$, and the $\mathbb{J}_i$-valued random variable X_i by $X_i \doteq (X_{i1},\dots,X_{in_i})$. Let $\hat{\mathbb{M}} \doteq \bar{\mathbb{M}}_1 \times \cdots \times \bar{\mathbb{M}}_\ell$, and define $\varpi : \hat{\mathbb{M}} \to \bar{\mathbb{M}}$ by $\varpi(\hat{m}) = m$, where for $B \in \mathscr{B}(\mathscr{Y})$, $A \in \mathscr{B}[0,T]$, and with $\hat{m} = (m_1,\dots,m_\ell)$, $m_i \in \bar{\mathbb{M}}_i$, we have

$$m(A \times B) = \sum_{i=1}^{q} m_i((A \cap \mathbb{I}_i) \times B).$$

With these definitions, ϖ concatenates the measures back together, and in particular, $\varpi((\bar{N}_1, \ldots, \bar{N}_\ell)) = \bar{N}$.

From the predictability properties of φ it follows that for $i = 2, \ldots, \ell$, there are measurable maps $\xi_i : \bar{\mathbb{M}}_1 \times \cdots \times \bar{\mathbb{M}}_{i-1} \to \mathbb{J}_i$, which can be written in component form $\xi_i = (\xi_{i1}, \ldots, \xi_{in_i})$ such that

$$X_{ij}(\bar{m}) = \xi_{ij}(\bar{N}_1(\bar{m}), \ldots, \bar{N}_{i-1}(\bar{m})).$$

Also, for $i = 1$, we set $X_1 = \xi_1$ a.s.-$\bar{\mathbb{P}}$ for some fixed vector ξ_1 in $\mathbb{J}_1$. The construction of $\hat{\varphi}$, which takes the same form as φ, is recursive. For $s \in \mathbb{I}_1$ we set $\hat{\varphi}(s, x, \bar{m}) = \varphi(s, x, \bar{m})$. As we will see, if there were only one time interval, we would be done, in that N^φ under $\bar{\mathbb{P}}$ and N^1 under $\bar{\mathbb{Q}}^{\hat{\varphi}}$ would have the same distribution, and the costs $L_T(\varphi)$ and $L_T(\hat{\varphi})$ would obviously be the same. The definition on subsequent intervals will depend on maps $T_i : \bar{\mathbb{M}}_1 \times \cdots \times \bar{\mathbb{M}}_i \to \bar{\mathbb{M}}_i$ for $i = 1, \ldots, \ell$, which must also be defined recursively.

Observe that under $\bar{\mathbb{P}}$, $\bar{N}_1$ has intensity $ds \times \nu(dx) \times dr$. Under $\bar{\mathbb{Q}}^{\hat{\varphi}}$, regardless of the definition of $\hat{\varphi}$ on later intervals, $\bar{N}_1$ has intensity

$$ds \times \nu(dx) \times \left[\sum_{j=1}^{n_1} \xi_{1j} 1_{E_{1j}}(x) 1_{(0,1]}(r) + 1_{(1,\infty)}(r) \right] dr.$$

The task of T_1 is to "undo" the effect of the change of measure, so that under $\bar{\mathbb{Q}}^{\hat{\varphi}}$, $\hat{N}_1 = T_1\left[\bar{N}_1\right]$ has intensity $ds \times \nu(dx) \times dr$. For $\bar{m}_1 \in \bar{\mathbb{M}}_1$, $\hat{m}_1 = T_1[\bar{m}_1]$ is defined as follows: for all $j \in \{1, \ldots, n_1\}$ and Borel subsets $A \subset \mathbb{I}_1$, $B \subset E_{1j}$, $C_1 \subset [0, \xi_{1j}]$ and $C_2 \subset (\xi_{1j}, \infty)$,

$$\hat{m}_1\left(A \times B \times [C_1 \cup C_2]\right) = \bar{m}_1\left(A \times B \times \left[\frac{1}{\xi_{1j}} C_1 \cup \left(C_2 - \xi_{1j} + 1\right)\right]\right).$$

The mapping T_1 can thus be viewed as a transformation on the underlying space $\mathscr{Y}_1$ on which m_1 is defined. An equivalent characterization of $\hat{m}_1 = T_1(\bar{m}_1)$ that will be used below is that $\hat{m}_1$ is the unique measure that for all nonnegative $\psi \in M_b(\mathscr{Y}_1)$ satisfies

$$\int_{\mathscr{Y}_1} \psi(s, x, r) \hat{m}_1(ds \times dx \times dr) = \sum_{j=1}^{n_1} \int_{\mathscr{Y}_1} 1_{E_{1j}}(x) \left[\psi(s, x, \xi_{1j} r) 1_{(0,1]}(r) + \psi(s, x, r + \xi_{1j} - 1) 1_{(1,\infty)}(r)\right] \bar{m}_1(ds \times dx \times dr).$$

With T_1 in hand, the definition of $\hat{\varphi}(s,x,\bar{m})$ for $s \in \mathbb{I}_2$ is straightforward. Indeed, since $\hat{N}_1$ has the same distribution under $\bar{\mathbb{Q}}^{\hat{\varphi}}$ that $\bar{N}_1$ has under $\bar{\mathbb{P}}$, and since each ξ_{2j} is a function only of $\bar{N}_1$, with the definition $\hat{X}_{2j} = \xi_{2j}(T_1[\bar{N}_1]) = \xi_{2j}(\hat{N}_1)$, $\hat{X}_{2j}$ under $\bar{\mathbb{Q}}^{\hat{\varphi}}$ has the same distribution as X_{2j} under $\bar{\mathbb{P}}$. We now define $\hat{\varphi}$ on $\mathbb{I}_1 \cup \mathbb{I}_2$ as in (8.23) but with X_{2j} replaced by $\hat{X}_{2j}$. Then $\left\{\hat{\varphi}(s,x,\bar{m}), s \in \mathbb{I}_1 \cup \mathbb{I}_2, x \in \mathscr{X}\right\}$ under $\bar{\mathbb{Q}}^{\hat{\varphi}}$ has the same distribution as $\{\varphi(s,x,\bar{m}), s \in \mathbb{I}_1 \cup \mathbb{I}_2, x \in \mathscr{X}\}$ under $\bar{\mathbb{P}}$.

We now proceed recursively, and having defined $T_1, \ldots, T_{p-1}$ for some $1 < p \le \ell$, we define T_p by $T_p(\bar{m}_1, \ldots, \bar{m}_p) = \hat{m}_p$, where $\hat{m}_p$ is the unique measure satisfying, for all nonnegative $\psi \in M_b(\mathscr{Y}_p)$,

$$\int_{\mathscr{Y}_p} \psi(s,x,r)\hat{m}_p(ds \times dx \times dr) = \sum_{j=1}^{n_p} \int_{\mathscr{Y}_p} 1_{E_{pj}}(x) \Big[\psi(s,x,\hat{\xi}_{pj}r)1_{(0,1]}(r) + \psi(s,x,r+\hat{\xi}_{pj}-1)1_{(1,\infty)}(r)\Big]\bar{m}_p(ds \times dx \times dr),$$

where $\hat{\xi}_p = \xi_p(\hat{m}_1, \ldots, \hat{m}_{p-1})$ and $\hat{m}_i = T_i(\bar{m}_1, \ldots, \bar{m}_i)$. We define the transformation $T : \bar{\mathbb{M}} \to \bar{\mathbb{M}}$ by

$$T(\bar{m}) = \varpi\left(T_1(\bar{N}_1(\bar{m})), \ldots, T_\ell(\bar{N}_1(\bar{m}), \ldots, \bar{N}_\ell(\bar{m}))\right),$$

and define $\hat{\varphi} \in \mathscr{A}_s$ for all times s by replacing X_{ij} with $\hat{X}_{ij}$ in the right side of (8.23), where

$$\hat{X}_i(\bar{m}) = X_i(T(\bar{m})) = \xi_i(T_1(\bar{N}_1(\bar{m})), \ldots, T_i(\bar{N}_1(\bar{m}), \ldots, \bar{N}_i(\bar{m}))). \tag{8.38}$$

Denoting $T(\bar{N})$ by $\hat{N}$, we see that for ϑ in the class $\hat{A}_b$ defined above Theorem 8.15,

$$\int \vartheta(s,x,r)\hat{N}(ds \times dx \times dr) = \int \Big[\vartheta(s,x,\hat{\varphi}(s,x)r)1_{(0,1]}(r) + \vartheta(s,x,r+\hat{\varphi}(s,x)-1)1_{(1,\infty)}(r)\Big]\bar{N}(ds \times dx \times dr). \tag{8.39}$$

Also, let $h_\varphi : \bar{\mathbb{M}} \to \mathbb{M}$ be defined by

$$h_\varphi(\bar{m})(A \times B) \doteq \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \bar{m}((A \cap \mathbb{I}_i) \times (B \cap E_{ij}) \times [0, X_{ij}(\bar{m})])$$

for $A \times B \in \mathscr{B}(\mathscr{X}_T)$. Recall that L_T was defined in (8.17). In order to complete the proof of the lemma, we will prove the following:

(a) the distribution of $\hat{N} = T(\bar{N})$ under $\bar{\mathbb{Q}}^{\hat{\varphi}}$ is the same as that of $\bar{N}$ under $\bar{\mathbb{P}}$;

(b) $h_\varphi(\bar{N}) = N^\varphi$ and $h_\varphi(T(\bar{N})) = h_\varphi(\hat{N}) = N^1$;

(c) for some measurable map $\Theta : \bar{\mathbb{M}} \to [0,\infty)$, $L_T(\varphi) = \Theta(\bar{N})$ and $L_T(\hat{\varphi}) = \Theta(T(\bar{N}))$, a.s. $\bar{\mathbb{P}}$.

Item (c) is an immediate consequence of the definition of $\hat{\varphi}$ via (8.38). We next consider (b). Noting that $\bar{N}(\bar{m}) = \bar{m}$, suppressing $\bar{m}$ in notation, and recalling the form of φ in (8.23), we have for $A \times B \in \mathscr{B}(\mathscr{X}_T)$,

$$\begin{aligned} h_\varphi(\bar{N})(A \times B) &= \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \bar{N}((A \cap \mathbb{I}_i) \times (B \cap E_{ij}) \times [0, X_{ij}]) \\ &= \sum_{i=1}^{\ell} \int_{(t_{i-1},t_i] \times \mathscr{X} \times [0,\infty)} 1_{A \times B}(s,x) 1_{[0,\varphi(s,x)]}(r) \bar{N}(ds \times dx \times dr) \\ &= \int_{\mathscr{Y}_T} 1_{A \times B}(s,x) 1_{[0,\varphi(s,x)]}(r) \bar{N}(ds \times dx \times dr) \\ &= N^{\varphi}(A \times B). \end{aligned}$$

This proves the first statement in (b). Next, using (8.38), (8.39), and that $r > 1$ implies $r + \hat{\varphi}(s,x) - 1 > \hat{\varphi}(s,x)$, we obtain

$$\begin{aligned} h_\varphi(T(\bar{N}))(A \times B) &= \int 1_{A \times B}(s,x) 1_{[0,\hat{\varphi}(s,x)]}(r) \hat{N}(ds \times dx \times dr) \\ &= \int 1_{A \times B}(s,x) [1_{[0,\hat{\varphi}(s,x)]}(\hat{\varphi}(s,x) r) 1_{[0,1]}(r) \\ &\quad + 1_{[0,\hat{\varphi}(s,x)]}(r + \hat{\varphi}(s,x) - 1) 1_{(1,\infty)}(r)] \bar{N}(ds \times dx \times dr) \\ &= \int 1_{A \times B}(s,x) 1_{[0,1]}(r) \bar{N}(ds \times dx \times dr) \\ &= N^1(A \times B). \end{aligned}$$

This proves the second statement in (b). Lastly, we prove (a). It suffices to show that for every $\vartheta \in \hat{A}_b$,

$$\bar{E}^{\hat{\varphi}} \int \vartheta(s,x,r) \hat{N}(ds \times dx \times dr) = \bar{E}^{\hat{\varphi}} \int \vartheta(s,x,r) \bar{\nu}_T(ds \times dx \times dr).$$

Using (8.39) and the last part of Theorem 8.15 for the first equality and that the marginal of $\bar{\nu}_T(ds \times dx \times dr)$ in r is Lebesgue measure, we have

$$\begin{aligned} &\bar{E}^{\hat{\varphi}} \int \vartheta(s,x,r) \hat{N}(ds \times dx \times dr) \\ &= \bar{E}^{\hat{\varphi}} \int \big[\vartheta(s,x,\hat{\varphi}(s,x) r) \hat{\varphi}(s,x) 1_{(0,1]}(r) \\ &\quad + \vartheta(s,x,r + \hat{\varphi}(s,x) - 1) 1_{(1,\infty)}(r)\big] \bar{\nu}_T(ds \times dx \times dr) \\ &= \bar{E}^{\hat{\varphi}} \int \big[\vartheta(s,x,r) 1_{(0,\hat{\varphi}(s,x)]}(r) \\ &\quad + \vartheta(s,x,r) 1_{(\hat{\varphi}(s,x),\infty)}(r)\big] \bar{\nu}_T(ds \times dx \times dr) \end{aligned}$$

$$= \bar{E}^{\hat{\varphi}} \int \vartheta(s,x,r)\bar{\nu}_T(ds \times dx \times dr),$$

which proves (a), and completes the proof of the lemma. □

8.3 Representation for Functionals of PRM and Brownian Motion

In this section we state the representation for functionals of both a PRM and a Hilbert space valued Wiener process. In Chap. 11 we will show how representations for a Hilbert space valued Brownian motion can be translated into representations for related objects, such as a collection of infinitely many independent scalar Brownian motions and the Brownian sheet. The analogous conversions can also be done when one is considering a PRM along with a Hilbert space valued Brownian motion.

The independent processes we consider are thus a Λ-Wiener process as in Definition 8.1 and a PRM as in Definition 8.11. In the proof of such a result we would follow the same procedure as in the separate cases and consider first the canonical space and filtration and then generalize. However, in this instance we skip the proof because it is simply a combination of the arguments used for the two separate cases, and we instead present a result that holds for a general filtration on a general probability space.

Thus we let $(\Omega, \mathscr{F}, P)$ be a probability space with a filtration $\{\mathscr{F}_t\}_{0\le t\le T}$ satisfying the usual conditions, and assume that $(\Omega, \mathscr{F}, P)$ supports all the following processes. Let W be a Λ-Wiener process with respect to $\{\mathscr{F}_t\}$. Let ν be a σ-finite measure on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$ and let $\bar{N}$ be a PRM, with respect to $\{\mathscr{F}_t\}$, on $\mathscr{Y}_T \doteq [0,T] \times \mathscr{Y}$, where $\mathscr{Y} \doteq \mathscr{X} \times [0,\infty)$, with intensity measure $\bar{\nu}_T \doteq \lambda_T \times \nu \times \lambda_\infty$. We assume that for all $0 \le s \le t < \infty$, $(\bar{N}((s,t] \times \cdot), W(t) - W(s))$ is independent of $\mathscr{F}_s$. Let $\mathscr{P}\mathscr{F}$ be the predictable σ-field on $[0,T] \times \Omega$. Let $\bar{\mathscr{A}}^W$ and $\bar{\mathscr{A}}_b^W$ be the collections of controls for the Wiener process defined as $\bar{\mathscr{A}}$ was below Definition 8.2 and $\bar{\mathscr{A}}_b$ was below (8.2) respectively, and let $\bar{\mathscr{A}}^N$, $\bar{\mathscr{A}}_b^N$ be controls for the PRM defined as $\bar{\mathscr{A}}$ and $\bar{\mathscr{A}}_b$ were below (8.19). The classes $\bar{\mathscr{A}}_{b,M}^N$ and $\bar{\mathscr{A}}_{b,M}^W$, which give uniform (in ω) bounds, are defined as they were in (8.18) and (8.2). For each $\varphi \in \bar{\mathscr{A}}^N$, N^φ will be a counting process on $\mathscr{X}_T$ defined as in (8.16) with φ as its controlled intensity measure.

Let $\bar{\mathscr{A}}_{b,M} \doteq \bar{\mathscr{A}}_{b,M}^W \times \bar{\mathscr{A}}_{b,M}^N$, $\bar{\mathscr{A}}_b \doteq \bar{\mathscr{A}}_b^W \times \bar{\mathscr{A}}_b^N$, and $\bar{\mathscr{A}} \doteq \bar{\mathscr{A}}^W \times \bar{\mathscr{A}}^N$. For $\psi \in \bar{\mathscr{A}}^W$, define $L_T^W(\psi) \doteq \frac{1}{2}\int_0^T \|\psi(s)\|_0^2\, ds$, with the norm $\|\cdot\|_0$ as in Sect. 8.1.1. For $\varphi \in \bar{\mathscr{A}}^N$, define $L_T^N(\varphi) \doteq \int_{\mathscr{X}_T} \ell(\varphi(t,x))\, \nu_T(dt \times dx)$, and for $u = (\psi, \varphi) \in \bar{\mathscr{A}}$, set $\bar{L}_T(u) \doteq L_T^N(\varphi) + L_T^W(\psi)$. For $\psi \in \bar{\mathscr{A}}^W$, let W^ψ be defined by $W^\psi(t) = W(t) + \int_0^t \psi(s)ds$, $t \in [0,T]$. We recall the definition of the space of measures $\mathbb{M} = \Sigma(\mathscr{X}_T)$ from Sect. 8.2.1 and its associated topology. With these definitions, the following representation holds. The proof of the second part of the theorem is similar to the proofs of Theorems 8.4 and 8.13.

Theorem 8.19 *Let $G \in M_b(\mathscr{C}([0,T] : \mathscr{H}) \times \mathbb{M})$. Then for $\theta \in (0,\infty)$,*

$$-\log E\exp\{-G(W,N^\theta)\} = \inf_{u=(\psi,\varphi)\in\mathscr{R}} E\left[\theta\bar{L}_T(u) + G(W^{\sqrt{\theta}\psi}, N^{\theta\varphi})\right],$$

where $\mathscr{R}$ can be either $\bar{\mathscr{A}}_b$ or $\bar{\mathscr{A}}$. Furthermore, for every $\delta > 0$, there exists $M < \infty$ depending on $\|G\|_\infty$ and δ such that for all $\varepsilon \in (0,1)$,

$$\begin{aligned} &-\varepsilon\log E\exp\left\{-\frac{1}{\varepsilon}G(\sqrt{\varepsilon}W, \varepsilon N^{1/\varepsilon})\right\} \\ &\quad\geq \inf_{u=(\psi,\varphi)\in\bar{\mathscr{A}}_{b,M}} E\left[\bar{L}_T(u) + G(\sqrt{\varepsilon}W^{\psi/\sqrt{\varepsilon}}, \varepsilon N^{\varphi/\varepsilon})\right] - \delta. \end{aligned}$$

8.4 Notes

For basic results on Hilbert space valued Brownian motions see [69], and for Poisson random measures see [159, 161].

The representation for a finite dimensional Brownian motion first appeared in [32]. In Sect. 3.2, we saw how this representation allowed a straightforward large deviation analysis of small noise diffusions using weak convergence arguments. Other applications of the finite dimensional case include large deviation analysis of small noise diffusions with discontinuous statistics [33] and homogenization [111], and also the analysis of importance sampling for accelerating Monte Carlo in estimating rare events [112].

With respect to its application to large deviation analysis, the representation is convenient because it eliminates the need for superexponentially close approximation and exponential tightness results used by other methods. A special case of the representation, rediscovered by Borell in [31], has found use in proving various functional inequalities, as in [184].

While convenient in the finite dimensional setting, the representation for functionals of Brownian motion and associated weak convergence methods are even more important for processes with an infinite dimensional state, where the proof of approximation and tightness results can be very demanding, and which often require assumptions beyond those needed for the large deviation result itself to be valid. Representations for infinite dimensional problems first appeared in [39] for the case of infinite dimensional Brownian motion, and in [45] for the case of Poisson random measures. The proof given here differs substantially from those of [39, 45], in particular in that they use classical-sense solutions to dynamic programming equations to establish the first step in the proof of the lower bound. As noted in the overview of Part III of the book, other authors have made numerous and varied applications of these representations and the associated abstract large deviation theorems that can be based on them. These abstract large deviation theorems are the topic of the next chapter.

A generalization of the representation that is sometimes useful (see, e.g., [11] for its use in a problem studying large deviations from a hydrodynamic limit) is that the infimum in the representation can be restricted to simple adapted processes [39]. Another extension is to the case in which the functional, in addition to depending on a BM and PRM, depends also on a $\mathscr{F}_0$-valued random variable, such as an initial condition (see [11, 46]).

A somewhat different variational representation for functionals of a PRM is presented in [269]. This representation is given in terms of some predictable transformations on the canonical Poisson space whose existence relies on solvability of certain nonlinear partial differential equations from the theory of mass transportation. This imposes restrictive conditions on the intensity measure (e.g., absolute continuity with respect to Lebesgue measure) of the PRM, and in particular, a standard Poisson process is not covered. The use of such a representation for proving large deviation results for general continuous time models with jumps appears to be unclear.

Chapter 9
Abstract Sufficient Conditions for Large and Moderate Deviations in the Small Noise Limit

In this chapter we use the representations derived in Chap. 8 to study large and moderate deviations for stochastic systems driven by Brownian and/or Poisson noise, and consider a "small noise" limit, as in Sects. 3.2 and 3.3. We will prove general abstract large deviation principles, and in later chapters apply these to models in which the noise enters the system in an additive and independent manner.[1] For these systems, one can view the mapping that takes the noise into the state of the system as "nearly" continuous, and it is this property that allows a unified and relatively straightforward treatment. In contrast, for the corresponding discrete time processes of Chap. 4, the noise entered in a possibly nonadditive way, and a more involved analysis was required. If we had restricted our attention in Chap. 4 to recursive models of the form

$$X_{i+1}^n = X_i^n + \frac{1}{n} b(X_i^n) + \frac{1}{n} \sigma(X_i^n)\theta_i, \quad X_0^n = x_0,$$

with $\{\theta_i\}_{i\in\mathbb{N}}$ an iid sequence (the discrete time analogues of the models in this chapter), then the analysis of Chap. 4 would have been much simpler. If one were to generalize within the continuous time framework to systems in which the noise enters in a more complicated manner, as in for example processes with multiple time or space scales (e.g., [111]), then the mapping from noise to state becomes more complex, as do the formulation of large deviation results and the methods of proof.

The main results of this chapter are Theorems 9.2 and 9.9 on uniform Laplace principles for a sequence of measurable functions of a Brownian motion and a PRM. Theorem 9.2 is well suited for proving large deviation results for small noise systems, whereas Theorem 9.9 is motivated by applications to moderate deviations. The proof of Theorem 9.2 is given in Sect. 9.3, and that of Theorem 9.9 is in Sect. 9.4. Theorems 9.2 and 9.9 are applied in Chap. 10 to develop large and moderate deviation approximations for certain finite dimensional systems. Infinite dimensional systems

[1] In our terminology, this includes systems with multiplicative noise, namely settings in which the noise term is multiplied by a state-dependent coefficient.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_9

are considered in Chap. 11, with the case of reaction–diffusion equations being developed in some detail, and in Chap. 12, where stochastic flows of diffeomorphisms are considered.

9.1 Definitions and Notation

The noise processes that drive the stochastic dynamical systems of this chapter were introduced in Chap. 8, and we adopt the notation used there. All stochastic processes are on the time horizon $[0, T]$, for some $T \in (0, \infty)$. We recall that Λ is a symmetric strictly positive trace class operator on the real separable Hilbert space $(\mathscr{H}, \langle \cdot, \cdot \rangle)$, $\mathscr{H}_0 \doteq \Lambda^{1/2}\mathscr{H}$, and $\mathbb{W} \doteq \mathscr{C}([0, T] : \mathscr{H}_0)$. Recall also that for a locally compact Polish space $\mathscr{S}$, $\Sigma(\mathscr{S})$ is the space of all measures ν on $(\mathscr{S}, \mathscr{B}(\mathscr{S}))$ satisfying $\nu(K) < \infty$ for every compact $K \subset \mathscr{S}$.

Throughout this chapter we deal simultaneously with Brownian and Poisson noise models. Because of this, we slightly modify the notation from Chap. 8 to make associations clear. Given a locally compact Polish space $\mathscr{X}$ that models the different types of jumps under the PRM, define $\mathscr{X}_T \doteq [0, T] \times \mathscr{X}$, the augmented space $\mathscr{Y} \doteq \mathscr{X} \times \mathbb{R}_+$ and the time-dependent version $\mathscr{Y}_T \doteq [0, T] \times \mathscr{Y}$, and canonical spaces $\mathbb{M} \doteq \Sigma(\mathscr{X}_T)$, $\bar{\mathbb{M}} \doteq \Sigma(\mathscr{Y}_T)$, $\mathbb{V} \doteq \mathbb{W} \times \mathbb{M}$, and $\bar{\mathbb{V}} \doteq \mathbb{W} \times \bar{\mathbb{M}}$. Let $\bar{N}$ and W be the maps from $\bar{\mathbb{V}}$ to $\bar{\mathbb{M}}$ and $\bar{\mathbb{V}}$ to $\mathbb{W}$ such that

$$\bar{N}(w, m) = m, \quad W(w, m) = w, \text{ for } (w, m) \in \bar{\mathbb{V}}.$$

Define

$$\mathscr{F}_t^0 \doteq \sigma\left\{\bar{N}((0, s] \times A), W(s) : 0 \le s \le t, A \in \mathscr{B}(\mathscr{Y})\right\}.$$

Assume $\nu \in \Sigma(\mathscr{X})$, and define $\bar{\nu} \doteq \nu \times \lambda_\infty$ and $\bar{\nu}_T \doteq \lambda_T \times \bar{\nu}$, where λ_T and λ_∞ are Lebesgue measure on $[0, T]$ and $[0, \infty)$, respectively. Let P denote the unique probability measure on $(\bar{\mathbb{V}}, \mathscr{B}(\bar{\mathbb{V}}))$ such that under P:

(a) W is a Λ-Wiener process with respect to $\mathscr{F}_t^0$;
(b) $\bar{N}$ is an $\mathscr{F}_t^0$-PRM with intensity measure $\bar{\nu}_T$;
(c) for all $0 \le s \le t < \infty$, $(\bar{N}((s, t] \times \cdot), W(t) - W(s))$ is independent of $\mathscr{F}_s^0$.

It follows that W and $\bar{N}$ are independent under P [167, Lemma 13.6]. Throughout this chapter we use $\{\mathscr{F}_t\}$, the augmentation of the filtration $\{\mathscr{F}_t^0\}$ with all P-null sets in $\mathscr{B}(\bar{\mathbb{V}})$. Recall the collections of controls $\bar{\mathscr{A}}^W$, $\bar{\mathscr{A}}_b^W$, $\bar{\mathscr{A}}_{b,n}^W$, $\bar{\mathscr{A}}^N$, $\bar{\mathscr{A}}_b^N$, and $\bar{\mathscr{A}}_{b,n}^N$ introduced in Sect. 8.3, where a subscript b, n means that costs are w.p.1 bounded by n, a b denotes the union over finite n of such controls, and the overbar indicates that the filtration used in defining these spaces is $\{\mathscr{F}_t\}$. We also have the definitions $\bar{\mathscr{A}}_{b,n} \doteq \bar{\mathscr{A}}_{b,n}^W \times \bar{\mathscr{A}}_{b,n}^N$, and $\bar{\mathscr{A}}_b \doteq \cup_{n \in \mathbb{N}} \bar{\mathscr{A}}_{b,n}$. For $u = (\psi, \varphi) \in \bar{\mathscr{A}}_b$, define the costs

$$L_T^W(\psi) \doteq \frac{1}{2} \int_0^T \|\psi(s)\|_0^2 \, ds \text{ and } L_T^N(\varphi) \doteq \int_{\mathscr{X}_T} \ell(\varphi(t, x)) \, \nu_T(dt \times dx) \tag{9.1}$$

as in Sect. 8.3, and let $\bar{L}_T(u) \doteq L_T^W(\psi) + L_T^N(\varphi)$. The controlled PRM N^φ is also defined as in that section.

Let S_n^W denote the subset of $\mathscr{L}^2([0,T] : \mathscr{H}_0)$ defined as in (8.1), and recall that this is a compact space with the weak topology on $\mathscr{L}^2([0,T] : \mathscr{H}_0)$. For $n \in \mathbb{N}$, define the analogous space

$$S_n^N \doteq \left\{ g : \mathscr{X}_T \to [0,\infty) : L_T^N(g) \le n \right\}.$$

A function $g \in S_n^N$ can be identified with a measure $\nu_T^g \in \mathbb{M}$ according to $\nu_T^g(A) = \int_A g(s,x)\, \nu_T(ds \times dx)$, $A \in \mathscr{B}(\mathscr{X}_T)$. Since convergence in $\mathbb{M}$ is essentially equivalent to weak convergence on compact subsets, the superlinear growth of ℓ implies that $\left\{ \nu_T^g : g \in S_n^N \right\}$ is a compact subset of $\mathbb{M}$. The proof of this fact is given in Appendix A.4.3. We equip S_n^N with the topology obtained through this identification, which makes S_n^N a compact space. We then let $S_n \doteq S_n^W \times S_n^N$ with the usual product topology, with respect to which it is also a compact space. An element $u \in \mathscr{A}_{b,n}$ is regarded as a random variable with values in the compact space S_n. Finally, let $S \doteq \cup_{n\in\mathbb{N}} S_n$.

9.2 Abstract Sufficient Conditions for LDP and MDP

Recall from Chap. 1 that various normalizations or scaling sequences are possible when one is formulating an LDP. In this section we formulate sufficient conditions for a Laplace principle to hold for general measurable functions of $(\sqrt{\varepsilon}W, \varepsilon N^{1/\varepsilon})$ and with two different scaling sequences. The first sufficient condition will be used in Chaps. 10, 11 and 12 to study large deviation principles for small noise stochastic dynamical systems. The second sufficient condition is for a moderate deviation principle. The condition is applied to finite dimensional models in Chap. 10, and for an example of its use in an infinite dimensional setting we refer to [41]. The results that we prove in fact give more, namely uniform Laplace principles in the sense of Definition 1.11. The uniformity is with respect to a parameter z (typically an initial condition), which takes values in some compact subset of a Polish space $\mathscr{Z}$.

The definition of a uniform Laplace principle was given in Chap. 1. The statement there considered the scale sequence $\varepsilon = 1/n$, and the analogous definition for a general scale function $\varkappa(\varepsilon)$ is as follows. Let $\{I_z, z \in \mathscr{Z}\}$ be a family of rate functions on $\mathscr{X}$ parametrized by z in a Polish space $\mathscr{Z}$ and assume that this family has compact level sets on compacts, namely, for each compact subset K of $\mathscr{Z}$ and each $M < \infty$, $\cup_{z\in K}\{x \in \mathscr{X} : I_z(x) \le M\}$ is a compact subset of $\mathscr{X}$. Let $\{X^\varepsilon\}$ be a collection of $\mathscr{X}$-valued random variables with distributions that depend on $z \in \mathscr{Z}$ and denote the corresponding expectation operator by E_z. The collection $\{X^\varepsilon\}$ is said to satisfy the Laplace principle on $\mathscr{X}$ with scale function $\varkappa(\varepsilon)$ and rate function I_z, uniformly on compacts, if for all compact subsets K of $\mathscr{Z}$ and all bounded continuous functions h mapping $\mathscr{X}$ into $\mathbb{R}$,

$$\lim_{\varepsilon\to 0}\sup_{z\in K}\left|\varkappa(\varepsilon)\log E_z\exp\{-\varkappa(\varepsilon)^{-1}h(X^\varepsilon)\}-F(z,h)\right|=0,$$

where $F(z,h)\doteq-\inf_{x\in\mathscr{X}}[h(x)+I_z(x)]$.

In this chapter it will be convenient to work with a common probability measure (instead of a collection parametrized by $z\in\mathscr{Z}$) and instead note the dependence on z in the collection of random variables, i.e., we write X_z^ε instead of X^ε.

9.2.1 An Abstract Large Deviation Result

In this section we present a sufficient condition for a uniform Laplace principle with scale function $\varkappa(\varepsilon)=\varepsilon$ to hold for measurable functions of $(\sqrt{\varepsilon}W,\varepsilon N^{1/\varepsilon})$. Recall that $N^{1/\varepsilon}$ is the PRM defined through (8.16) with $\varphi\equiv 1/\varepsilon$. It is defined on $\bar{\mathbb{V}}$, takes values in $\mathbb{M}$, and has an intensity measure that is scaled by $1/\varepsilon$. Let $\{\mathscr{G}^\varepsilon\}_{\varepsilon>0}$, be a family of measurable maps from $\mathscr{Z}\times\mathbb{V}$ to $\mathbb{U}$, where $\mathscr{Z}$ and $\mathbb{U}$ are some Polish spaces. Let $\left\{Z_z^\varepsilon\right\}_{\varepsilon>0,z\in\mathscr{Z}}$ be the collection of $\mathbb{U}$-valued random variables on $(\bar{\mathbb{V}},\mathscr{B}(\bar{\mathbb{V}}),P)$ defined by

$$Z_z^\varepsilon\doteq\mathscr{G}^\varepsilon(z,\sqrt{\varepsilon}W,\varepsilon N^{1/\varepsilon}). \tag{9.2}$$

We are interested in a uniform large deviation principle for the family $\left\{Z_z^\varepsilon\right\}$ as $\varepsilon\to 0$. We recall from Proposition 1.14 that a uniform large deviation principle is implied by a uniform Laplace principle.

A control $u=(\psi,\varphi)\in\bar{\mathscr{A}}_{b,n}$ will be regarded as a random variable with values in the compact metric space S_n. The following is a sufficient condition for a large deviation property. As noted in Sect. 9.1, $\bar{L}_T(u)=L_T^W(\psi)+L_T^N(\varphi)$. Recall also the notation $W^\psi(\cdot)=W(\cdot)+\int_0^\cdot\psi(s)ds$.

Condition 9.1 *There exists a measurable map $\mathscr{G}^0:\mathscr{Z}\times\mathbb{V}\to\mathbb{U}$ such that the following hold.*

(a) *For $n\in\mathbb{N}$ and compact $K\subset\mathscr{Z}$, the set*

$$\Gamma_{n,K}\doteq\left\{\mathscr{G}^0\left(z,\int_0^\cdot f(s)ds,\nu_T^g\right):q=(f,g)\in S,\bar{L}_T(q)\le n,z\in K\right\} \tag{9.3}$$

is a compact subset of $\mathbb{U}$.

(b) *For $n\in\mathbb{N}$, let $u^\varepsilon=(\psi^\varepsilon,\varphi^\varepsilon)\in\bar{\mathscr{A}}_{b,n}$, $u=(\psi,\varphi)\in\bar{\mathscr{A}}_{b,n}$ be such that u^ε converges in distribution to u as $\varepsilon\to 0$. Also, let $\{z^\varepsilon\}\subset\mathscr{Z}$ be such that $z^\varepsilon\to z$ as $\varepsilon\to 0$. Then*

$$\mathscr{G}^\varepsilon\left(z^\varepsilon,\sqrt{\varepsilon}W^{\psi^\varepsilon/\sqrt{\varepsilon}},\varepsilon N^{\varphi^\varepsilon/\varepsilon}\right)\Rightarrow\mathscr{G}^0\left(z,\int_0^\cdot\psi(s)ds,\nu_T^\varphi\right).$$

For $\phi\in\mathbb{U}$ and $z\in\mathscr{Z}$, define $S_{z,\phi}^{\mathscr{G}}\doteq\left\{(f,g)\in S:\phi=\mathscr{G}^0(z,\int_0^\cdot f(s)ds,\nu_T^g)\right\}$. These are the controls that produce the output ϕ. For $z\in\mathscr{Z}$, let $I_z:\mathbb{U}\to[0,\infty]$ be

defined by

$$I_z(\phi) \doteq \inf_{q=(f,g)\in S^{\mathscr{G}}_{z,\phi}} \bar{L}_T(q). \tag{9.4}$$

Theorem 9.2 *Suppose that $\mathscr{G}^\varepsilon$ and $\mathscr{G}^0$ satisfy Condition 9.1. Suppose also that for all $\phi \in \mathbb{U}$, $z \mapsto I_z(\phi)$ is a lower semicontinuous mapping from $\mathscr{Z}$ to $[0,\infty]$. Then for all $z \in \mathscr{Z}$, I_z defined in (9.4) is a rate function on $\mathbb{U}$, the family $\{I_z, z \in \mathscr{Z}\}$ of rate functions has compact level sets on compacts, and $\left\{Z^\varepsilon_z\right\}$ satisfies a Laplace principle with scale function ε and rate function I_z, uniformly on compact subsets of $\mathscr{Z}$.*

Remark 9.3 Note that the lower semicontinuity of $z \mapsto I_z(\phi)$ is typically automatic in the situation in which z is an initial condition for a stochastic process defined by the mapping $\mathscr{G}^\varepsilon$, since in this case $I_z(\phi) < \infty$ only when $z = \phi(0)$.

The proof of Theorem 9.2 is given in Sect. 9.3. Two examples were given in Chap. 3 to illustrate the role of $\mathscr{G}^\varepsilon$ and the scaling. For convenience, we recall the diffusion example, and hence take $\mathbb{V} = \mathscr{C}([0,1] : \mathbb{R}^k)$, $\mathbb{U} = \mathscr{C}([0,1] : \mathbb{R}^d)$, and $\mathscr{Z} = \mathbb{R}^d$.

Example 9.4 Suppose $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times k}$ satisfy

$$\|b(x) - b(y)\| + \|\sigma(x) - \sigma(y)\| \le C\,\|x - y\|$$

for all $x, y \in \mathbb{R}^d$, with $C \in (0,\infty)$. Let W be a standard k-dimensional Brownian motion, and for fixed $z \in \mathbb{R}^d$ and $\varepsilon > 0$, let $X^\varepsilon_z = \left\{X^\varepsilon_z(t)\right\}_{0\le t\le 1}$ be the strong solution of the SDE

$$dX^\varepsilon_z(t) = b(X^\varepsilon_z(t))dt + \sqrt{\varepsilon}\sigma(X^\varepsilon_z(t))dW(t),\ \ X^\varepsilon_z(0) = z. \tag{9.5}$$

From the unique pathwise solvability of this SDE (see [172, Definition 5.3.2 and Corollary 5.3.23]), it follows that for each $\varepsilon > 0$, there is a measurable map $\mathscr{G}^\varepsilon : \mathbb{R}^d \times \mathscr{C}([0,1] : \mathbb{R}^k) \to \mathscr{C}([0,1] : \mathbb{R}^d)$ such that $X^\varepsilon_z = \mathscr{G}^\varepsilon(z, \sqrt{\varepsilon}W)$ is the solution to (9.5). The corresponding map $\mathscr{G}^0$ can be defined by $\mathscr{G}^0\left(z, \int_0^\cdot f(s)ds\right) = \varphi$ if for $z \in \mathbb{R}^d$ and $f \in \mathscr{L}^2([0,1] : \mathbb{R}^k)$,

$$\varphi(t) = z + \int_0^t b(\varphi(s))ds + \int_0^t \sigma(\varphi(s))f(s)ds, \quad t \in [0,1],$$

and $\mathscr{G}^0(z,\gamma) \equiv 0$ for all other $(z,\gamma) \in \mathbb{R}^d \times \mathscr{C}([0,1] : \mathbb{R}^d)$. Along the lines of the discussion in Chap. 3, it is easily checked that Condition 9.1 is valid, and in particular, part (b) is simply a restatement of the LLN limit $\bar{X}^\varepsilon_{z^\varepsilon} \Rightarrow \bar{X}_z$, where $\bar{X}^\varepsilon_{z^\varepsilon}$ and $\bar{X}_z$ are the solutions to

$$d\bar{X}^\varepsilon_{z^\varepsilon}(t) = b(\bar{X}^\varepsilon_{z^\varepsilon}(t))dt + \sigma(\bar{X}^\varepsilon_{z^\varepsilon}(t))\psi^\varepsilon(t)dt + \sqrt{\varepsilon}\sigma(\bar{X}^\varepsilon_{z^\varepsilon}(t))dW(t),$$

$\bar{X}^{\varepsilon}_{z^\varepsilon}(0) = z_\varepsilon$, and

$$d\bar{X}_z(t) = b(\bar{X}_z(t))dt + \sigma(\bar{X}_z(t))\psi(t)dt, \;\; \bar{X}_z(0) = z.$$

In Chaps. 10, 11, and 12 we consider other applications of Theorem 9.2.

Remark 9.5 The discussion of Example 9.4 shows that the main additional work needed to prove a uniform LDP instead of the ordinary LDP is to prove, instead of the convergence $\bar{X}^{\varepsilon}_z \Rightarrow \bar{X}_z$, the stronger convergence property $\bar{X}^{\varepsilon}_{z^\varepsilon} \Rightarrow \bar{X}_z$ whenever $z^\varepsilon \to z$. However, the proof of this stronger convergence property, at least in most situations of interest, requires the same analysis as that used for the convergence with a fixed initial condition. Thus in some uses later of Theorems 9.2 and 9.9 we present the argument for a fixed initial condition, and leave it to the reader to check that the same arguments could be used for converging initial conditions and thereby yield the uniform Laplace principle. Two exceptions are the reaction–diffusion example of Chap. 11 and the serve-the-longest queueing example of Chap. 13. For the latter example, as with many models in queueing, the discrete nature of the state space for the prelimit models requires initial conditions that depend on the scaling parameter.

9.2.2 An Abstract Moderate Deviation Result

Let $\{\mathscr{K}^\varepsilon\}_{\varepsilon>0}$ be a family of measurable maps from $\mathscr{Z} \times \mathbb{V}$ to $\mathbb{U}$. Let $a : (0, \infty) \to (0, \infty)$ be such that as $\varepsilon \to 0$,

$$a(\varepsilon) \to 0 \text{ and } \varkappa(\varepsilon) \doteq \frac{\varepsilon}{a^2(\varepsilon)} \to 0. \tag{9.6}$$

For $\varepsilon > 0$ and $z \in \mathscr{Z}$, let $Y^\varepsilon_z \doteq \mathscr{K}^\varepsilon(z, \sqrt{\varepsilon}W, \varepsilon N^{1/\varepsilon})$. In this section we formulate a sufficient condition for the collection $\left\{Y^\varepsilon_z\right\}_{\varepsilon>0}$ to satisfy a uniform Laplace principle with scale function $\varkappa(\varepsilon)$ and a rate function that is given through a suitable quadratic form.

While the large and moderate deviation assumptions and arguments are very similar when one is considering a diffusion model, a significant difference occurs when a PRM driving noise is included. This is very similar to the situation encountered in the discrete time analogue presented in Chap. 5. In particular, the Poisson cost must be replaced by an appropriate quadratic functional in the limit $\varepsilon \to 0$, and the dynamics are also adjusted to make analysis of the LLN limits easier. This is done by centering the controls on $\varphi \equiv 1$ and rescaling. The following inequalities will be used to translate bounds on controls φ^ε into bounds on this quadratic approximation. Recall the function $\ell(r) \doteq r \log r - r + 1$. The following properties can be easily shown. Part (a) has been used many times already, but is included here for convenience.

Lemma 9.6 (a) *For $a, b \in (0,\infty)$ and $\sigma \in [1,\infty)$, $ab \le e^{\sigma a} + \frac{1}{\sigma}\ell(b)$.*

(b) *For every $\beta > 0$, there exist $\kappa_1(\beta), \bar{\kappa}_1(\beta) \in (0,\infty)$ such that $\kappa_1(\beta)$ and $\bar{\kappa}_1(\beta)$ converge to 0 as $\beta \to \infty$, and for $r \ge 0$,*

$$|r-1| \le \kappa_1(\beta)\ell(r) \text{ if } |r-1| \ge \beta, \text{ and } r \le \bar{\kappa}_1(\beta)\ell(r) \text{ if } r \ge \beta > 1.$$

(c) *There is a nondecreasing function $\kappa_2 : (0,\infty) \to (0,\infty)$ such that for each $\beta > 0$,*

$$|r-1|^2 \le \kappa_2(\beta)\ell(r) \quad \text{for } |r-1| \le \beta, r \ge 0.$$

(d) *There exists $\kappa_3 \in (0,\infty)$ such that*

$$\ell(r) \le \kappa_3|r-1|^2, \quad |\ell(r) - (r-1)^2/2| \le \kappa_3|r-1|^3 \text{ for all } r \ge 0.$$

Recall $L_T^N(g) \doteq \int_{\mathscr{X}_T} \ell(g(s,y))\nu_T(ds \times dy)$ and $L_T^W(\psi) \doteq \frac{1}{2}\int_0^T \|\psi(s)\|_0^2\, ds$. For $\varepsilon > 0$ and $n \in \mathbb{N}$, define the spaces

$$\begin{aligned} S_{n,+}^{N,\varepsilon} &\doteq \{g : \mathscr{X}_T \to \mathbb{R}_+ \text{ such that } L_T^N(g) \le na^2(\varepsilon)\} \\ S_n^{N,\varepsilon} &\doteq \{f : \mathscr{X}_T \to \mathbb{R} \text{ such that } f = (g-1)/a(\varepsilon), \text{ with } g \in S_{n,+}^{N,\varepsilon}\}. \end{aligned} \tag{9.7}$$

Thus $S_n^{N,\varepsilon}$ are the centered and rescaled versions of the nonnegative functions appearing in $S_{n,+}^{N,\varepsilon}$. The following result is immediate from Lemma 9.6.

Lemma 9.7 *Suppose $g \in S_{n,+}^{N,\varepsilon}$ for some $n < \infty$ and let $f = (g-1)/a(\varepsilon)$. Then:*

(a) $\displaystyle\int_{\mathscr{X}_T} |f(s,y)|1_{\{|f(s,y)| \ge \beta/a(\varepsilon)\}}\nu_T(ds \times dy) \le na(\varepsilon)\kappa_1(\beta)$ *for all $\beta > 0$;*

(b) $\displaystyle\int_{\mathscr{X}_T} g(s,y)1_{\{g(s,y) \ge \beta\}}\nu_T(ds \times dy) \le na^2(\varepsilon)\bar{\kappa}_1(\beta)$ *for all $\beta > 1$;*

(c) $\displaystyle\int_{\mathscr{X}_T} |f(s,y)|^2 1_{\{|f(s,y)| \le \beta/a(\varepsilon)\}}\nu_T(ds \times dy) \le n\kappa_2(\beta)$ *for all $\beta > 0$,*

where $\kappa_1, \bar{\kappa}_1$ and κ_2 are as in Lemma 9.6.

Let

$$\mathscr{U}_{n,+}^{\varepsilon} \doteq \left\{(u_W, u_N) \in \bar{\mathscr{A}}_b : u_W(\cdot,\omega) \in S_{na(\varepsilon)^2}^W,\, u_N(\cdot,\cdot,\omega) \in S_{n,+}^{N,\varepsilon},\ \bar{P}\text{-a.s.}\right\}. \tag{9.8}$$

Thus by (9.7), $\mathscr{U}_{n,+}^{\varepsilon}$ is the class of controls for both types of noise for which the cost scales proportionally with $a(\varepsilon)^2$. Owing to the moderate deviation scaling, one can assume without loss that the control appearing in the representation can be restricted to a class of this form, with n depending on the function F. However, as $\varepsilon \to 0$ we will need to use the centered and rescaled analogues, which are related to a diffusion approximation to the original process. This requires additional notation and definitions.

The norm in the Hilbert space $\mathscr{L}^2(\nu_T)$ is denoted by $\|\cdot\|_{N,2}$, and the norm in $\mathscr{L}^2([0,T]:\mathscr{H}_0)$ by $\|\cdot\|_{W,2}$. Let $\mathscr{L}^2 \doteq \mathscr{L}^2([0,T]:\mathscr{H}_0)\times\mathscr{L}^2(\nu_T)$ and recall that $\mathscr{P}\mathscr{F}$ is the predictable σ-field on $[0,T]\times\bar{\mathbb{V}}$ with the filtration $\{\mathscr{F}_t\}$ on $(\bar{\mathbb{V}},\mathscr{B}(\bar{\mathbb{V}}))$.

Given a map $\mathscr{K}^0 : \mathscr{Z}\times\mathscr{L}^2\to\mathbb{U}$, $z\in\mathscr{Z}$, and $\eta\in\mathbb{U}$, let

$$S^{\mathscr{K}}_{z,\eta} \doteq \{q=(f_1,f_2)\in\mathscr{L}^2 : \eta = \mathscr{K}^0(z,q)\},$$

and define I_z for $z\in\mathscr{Z}$ by

$$I_z(\eta) \doteq \inf_{q=(f_1,f_2)\in S^{\mathscr{K}}_{z,\eta}} \left[\frac{1}{2}\left(\|f_1\|^2_{W,2}+\|f_2\|^2_{N,2}\right)\right]. \tag{9.9}$$

Here $S^{\mathscr{K}}_{z,\eta}$ identifies the $\mathscr{L}^2$ spaces that lead to the outcome q under the map $\mathscr{K}^0$, which can be associated with the map $\mathscr{K}^\varepsilon$ linearized about the LLN limit. As always, we follow the convention that the infimum over an empty set is $+\infty$.

We now introduce a sufficient condition that ensures that I_z is a rate function for every $z\in\mathscr{Z}$, the collection $\{I_z\}_{z\in\mathscr{Z}}$ has compact level sets on compacts, and the collection $\{Y^\varepsilon_z\}$ satisfies a Laplace principle with scale function $\varkappa(\varepsilon)$ and rate function I_z as $\varepsilon\to 0$. Let

$$\hat{S}_n \doteq \{(f_1,f_2)\in\mathscr{L}^2 : \|f_1\|^2_{W,2}+\|f_2\|^2_{N,2}\le n\}. \tag{9.10}$$

Condition 9.8 *For some measurable map $\mathscr{K}^0 : \mathscr{Z}\times\mathscr{L}^2\to\mathbb{U}$, the following two conditions hold.*

(a) *For every $n\in\mathbb{N}$ and compact $K\subset\mathscr{Z}$, the set*

$$\Gamma_{n,K} \doteq \left\{\mathscr{K}^0(z,q) : z\in K, q\in\hat{S}_n\right\}$$

is a compact subset of $\mathbb{U}$.

(b) *Given $n\in\mathbb{N}$ and $\varepsilon>0$, let $(\psi^\varepsilon,\varphi^\varepsilon)\in\mathscr{U}^\varepsilon_{n,+}$ [defined in (9.8)]. Let $\theta^\varepsilon = \psi^\varepsilon/a(\varepsilon)$ and $\zeta^\varepsilon = (\varphi^\varepsilon-1)/a(\varepsilon)$. Suppose that for some $\beta\in(0,1]$, there is $m\in\mathbb{N}$ such that $(\theta^\varepsilon,\zeta^\varepsilon 1_{\{|\zeta^\varepsilon|\le\beta/a(\varepsilon)\}})\Rightarrow(\theta,\zeta)$ in $\hat{S}_m$. Also, let $\{z^\varepsilon\}\subset\mathscr{Z}$ be such that $z^\varepsilon\to z$ as $\varepsilon\to 0$. Then*

$$\mathscr{K}^\varepsilon\left(z^\varepsilon,\sqrt{\varepsilon}W^{\psi^\varepsilon/\sqrt{\varepsilon}},\ \varepsilon N^{\varphi^\varepsilon/\varepsilon}\right)\Rightarrow\mathscr{K}^0(z,\theta,\zeta).$$

Note that from Lemma 9.7, $(\theta^\varepsilon,\zeta^\varepsilon 1_{\{|\zeta^\varepsilon|\le\beta/a(\varepsilon)\}})$, as in part (b) of Condition 9.8, takes values in $\hat{S}_m$ with $m=n(1+\kappa_2(\beta))$. Thus if the condition holds, one can take $m=n(1+\kappa_2(\beta))$ without loss of generality. The following is the analogue of Theorem 9.2 for the moderate deviation scaling.

Theorem 9.9 *Suppose that $\mathscr{K}^\varepsilon$ and $\mathscr{K}^0$ satisfy Condition 9.8. Suppose also that for all $\phi\in\mathbb{U}$, $z\mapsto I_z(\phi)$ is a lower semicontinuous mapping from $\mathscr{Z}$ to $[0,\infty]$.*

Then for $z \in \mathscr{Z}$, I_z defined in (9.9) is a rate function on $\mathbb{U}$, the family $\{I_z, z \in \mathscr{Z}\}$ of rate functions has compact level sets on compacts, and $\{Y_z^\varepsilon\}$ satisfies a uniform Laplace principle with scale function $\varkappa(\varepsilon)$ and rate function I_z as $\varepsilon \to 0$.

The proof of Theorem 9.9 is in Sect. 9.4. As with Theorem 9.2, the assumption of lower semicontinuity of $z \mapsto I_z(\phi)$ is often vacuous when z plays the role of an initial condition (see Remark 9.3).

Example 9.10 Let X_z^ε be as in Example 9.4. In addition to the Lipschitz condition on the coefficients b, σ, assume that b is continuously differentiable. Let X_z^0 be the solution of the ODE $\dot{X}_z^0(t) = b(X_z^0(t))$, $X_z^0(0) = z$, and let $Y_z^\varepsilon = (X_z^\varepsilon - X_z^0)/a(\varepsilon)$. Then for each $\varepsilon > 0$, there is a measurable map $\mathscr{K}^\varepsilon$ from $\mathbb{R}^d \times \mathscr{C}([0,1] : \mathbb{R}^k)$ to $\mathscr{C}([0,1] : \mathbb{R}^d)$ such that $Y_z^\varepsilon = \mathscr{K}^\varepsilon(z, \sqrt{\varepsilon} W)$. The corresponding map

$$\mathscr{K}^0 : \mathbb{R}^d \times \mathscr{L}^2([0,1] : \mathbb{R}^k) \to \mathscr{C}([0,1] : \mathbb{R}^d)$$

is defined by $\mathscr{K}^0(z, f) = \eta_{z,f}$, where $\eta_{z,f}$ is the unique solution of the equation

$$\eta_{z,f}(t) = \int_0^t [Db(X_z^0(s))](\eta_{z,f}(s))ds + \int_0^t \sigma(X_z^0(s)) f(s) ds,$$

where $Db(x)$ is the matrix $(\partial b_i(x)/\partial x_j)_{ij}$. In Chap. 10, under the assumption that Db is Lipschitz continuous, it will be shown using Theorem 9.9 that for each fixed z, Y_z^ε satisfies a Laplace principle with scale function $\varkappa(\varepsilon)$. The proof of the uniform Laplace principle can be given similarly by considering arbitrary $z_\varepsilon \to z$ as $\varepsilon \to 0$. This result can be viewed as a moderate deviation principle for the diffusion process X_z^ε. Chapter 10 will treat the more general setting of d-dimensional jump-diffusions, and we refer the reader to [41] for analogous results in an infinite dimensional setting.

9.3 Proof of the Large Deviation Principle

In this section we prove Theorem 9.2. We first argue that for all compact $K \subset \mathscr{Z}$ and each $M < \infty$,

$$\Lambda_{M,K} \doteq \cup_{z \in K} \{\phi \in \mathbb{U} : I_z(\phi) \le M\} \tag{9.11}$$

is a compact subset of $\mathbb{U}$. Note that this will show that for each $z \in \mathscr{Z}$, I_z is a rate function and the collection $\{I_z, z \in \mathscr{Z}\}$ has compact level sets on compacts. To establish this, we will show that $\Lambda_{M,K}$ equals $\cap_{\delta \in (0,1)} \Gamma_{M+\delta,K}$, where $\Gamma_{M,K}$ is as in (9.3). In view of part (a) of Condition 9.1, the compactness of $\Lambda_{M,K}$ will then follow. Let $\phi \in \Lambda_{M,K}$. Then there exists $z \in K$ such that $I_z(\phi) \le M$. We can now find, for each $\delta \in (0,1)$, $q_\delta = (f_\delta, g_\delta) \in S_{z,\phi}^{\mathscr{G}}$, i.e., $\phi = \mathscr{G}^0(z, \int_0^\cdot f_\delta(s)ds, \nu_T^{g_\delta})$, such that $\bar{L}_T(q_\delta) \le M + \delta$. In particular, $\phi \in \Gamma_{M+\delta,K}$. Since $\delta \in (0,1)$ is arbitrary, we have $\Lambda_{M,K} \subset \cap_{\delta \in (0,1)} \Gamma_{M+\delta,K}$. Conversely, suppose $\phi \in \Gamma_{M+\delta,K}$ for all $\delta \in (0,1)$.

Then for each $\delta \in (0,1)$, there exists $z_\delta \in K$, $q_\delta = (f_\delta, g_\delta) \in S$, $\bar{L}_T(q_\delta) \le M + \delta$, such that $\phi = \mathscr{G}^0(z_\delta, \int_0^\cdot f_\delta(s)ds, \nu_T^{g_\delta})$. In particular, we have

$$\inf_{z\in K} I_z(\phi) \le I_{z_\delta}(\phi) \le \bar{L}_T(q_\delta) \le M + \delta.$$

Sending $\delta \to 0$ gives $\inf_{z\in K} I_z(\phi) \le M$. Since the map $z \mapsto I_z(\phi)$ is lower semicontinuous, $\phi \in \Lambda_{M,K}$, and the inclusion $\cap_{\delta\in(0,1)}\Gamma_{M+\delta,K} \subset \Lambda_{M,K}$ follows. This proves the compactness of $\Lambda_{M,K}$ and finishes the first part of the theorem.

We next prove the second statement in the theorem. Fix $z \in \mathscr{Z}$ and let $\{z^\varepsilon\}_{\varepsilon>0} \subset \mathscr{Z}$ be such that $z^\varepsilon \to z$ as $\varepsilon \to 0$. Fix $F \in \mathscr{C}_b(\mathbb{U})$. In view of Proposition 1.12, it suffices to prove the Laplace upper bound

$$\limsup_{\varepsilon\to 0} \varepsilon \log E \exp\left\{-\frac{1}{\varepsilon}F(Z^\varepsilon_{z^\varepsilon})\right\} \le -\inf_{\phi\in\mathbb{U}}[I_z(\phi) + F(\phi)] \tag{9.12}$$

and lower bound

$$\liminf_{\varepsilon\to 0} \varepsilon \log E \exp\left\{-\frac{1}{\varepsilon}F(Z^\varepsilon_{z^\varepsilon})\right\} \ge -\inf_{\phi\in\mathbb{U}}[I_z(\phi) + F(\phi)]. \tag{9.13}$$

Proof of the Laplace upper bound. Fix $\delta > 0$ and recall from Sect. 8.3 that W^ψ denotes $W + \int_0^\cdot \psi(s)ds$. Using Theorem 8.19 and the definition (9.2) of $Z^\varepsilon_{z^\varepsilon}$, there exists $M < \infty$ such that for each $\varepsilon \in (0,1)$, one can find $u^\varepsilon = (\psi^\varepsilon, \varphi^\varepsilon) \in \mathscr{A}_{b,M}$ with

$$\begin{aligned}
&-\varepsilon \log E \exp\left\{-\frac{1}{\varepsilon}F(Z^\varepsilon_{z^\varepsilon})\right\} \\
&= -\varepsilon \log E \exp\left\{-\frac{1}{\varepsilon}F \circ \mathscr{G}^\varepsilon(z^\varepsilon, \sqrt{\varepsilon}W, \varepsilon N^{1/\varepsilon})\right\} \\
&\ge E\left[\bar{L}_T(u^\varepsilon) + F \circ \mathscr{G}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}W^{\psi^\varepsilon/\sqrt{\varepsilon}}, \varepsilon N^{\varphi^\varepsilon/\varepsilon}\right)\right] - \delta.
\end{aligned}$$

Using the compactness of S_M, we can find a subsequence $\{\varepsilon_k\}$ along which u^{ε_k} converges in distribution to some $u{=}(\psi, \varphi)$ that takes values in S_M a.s. By a standard subsequential argument, it is enough to demonstrate the lower bound for this subsequence, which for simplicity we label as ε. From part (b) of Condition 9.1, it follows that

$$\begin{aligned}
&\liminf_{\varepsilon\to 0} E\left[\bar{L}_T(u^\varepsilon) + F \circ \mathscr{G}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}W^{\psi^\varepsilon/\sqrt{\varepsilon}}, \varepsilon N^{\varphi^\varepsilon/\varepsilon}\right)\right] \\
&\ge E\left[\bar{L}_T(u) + F \circ \mathscr{G}^0\left(z, \int_0^\cdot \psi(s)ds, \nu_T^\varphi\right)\right] \\
&\ge \inf_{\{(\phi,q)\in\mathbb{U}\times S_M : q\in S^{\mathscr{G}}_{z,\phi}\}} \left[\bar{L}_T(q) + F(\phi)\right] \\
&= \inf_{\phi\in\mathbb{U}}[I_z(\phi) + F(\phi)],
\end{aligned}$$

where the first inequality is a consequence of Fatou's lemma and the lower semicontinuity of $q \mapsto \bar{L}_T(q)$ on S_M, and the second inequality follows from the definition of $S^{\mathscr{G}}_{z,\phi}$, and the equality is due to the definition of $I_z(\phi)$ in (9.4). Since $\delta > 0$ is arbitrary, this completes the proof of the upper bound (9.12). □

Proof of the Laplace lower bound. We need to prove the inequality in (9.13). Without loss of generality we can assume that $\inf_{\phi\in\mathbb{U}}[I_z(\phi) + F(\phi)] < \infty$. Let $\delta > 0$ be arbitrary, and let $\phi_0 \in \mathbb{U}$ be such that

$$I_z(\phi_0) + F(\phi_0) \le \inf_{\phi\in\mathbb{U}}[I_z(\phi) + F(\phi)] + \frac{\delta}{2}. \tag{9.14}$$

Choose $q_0 = (f_0, g_0) \in S^{\mathscr{G}}_{z,\phi_0}$ such that

$$\bar{L}_T(q_0) \le I_z(\phi_0) + \frac{\delta}{2}. \tag{9.15}$$

Note that $\phi_0 = \mathscr{G}^0\left(z, \int_0^\cdot f_0(s)ds, \nu_T^{g_0}\right)$. Using the representation in Theorem 8.19, we obtain

$$\begin{aligned}
&\limsup_{\varepsilon\to 0} -\varepsilon \log E \exp\left\{-\frac{1}{\varepsilon}F(Z^\varepsilon_{z^\varepsilon})\right\} \\
&= \limsup_{\varepsilon\to 0} \inf_{u=(\psi,\varphi)\in\mathscr{A}} E\left[\bar{L}_T(u) + F\circ\mathscr{G}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}W^{\psi/\sqrt{\varepsilon}}, \varepsilon N^{\varphi/\varepsilon}\right)\right] \\
&\le \limsup_{\varepsilon\to 0} E\left[\bar{L}_T(q_0) + F\circ\mathscr{G}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}W^{f_0/\sqrt{\varepsilon}}, \varepsilon N^{g_0/\varepsilon}\right)\right] \\
&= \bar{L}_T(q_0) + \limsup_{\varepsilon\to 0} E\left[F\circ\mathscr{G}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}W^{f_0/\sqrt{\varepsilon}}, \varepsilon N^{g_0/\varepsilon}\right)\right].
\end{aligned}$$

By part (b) of Condition 9.1, we have

$$\begin{aligned}
\lim_{\varepsilon\to 0} E\left[F\circ\mathscr{G}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}W^{f_0/\sqrt{\varepsilon}}, \varepsilon N^{g_0/\varepsilon}\right)\right] &= F\circ\mathscr{G}^0\left(z, \int_0^\cdot f_0(s)ds, \nu_T^{g_0}\right) \\
&= F(\phi_0).
\end{aligned}$$

In view of (9.14) and (9.15), the left side of (9.13) can be at most $\inf_{\phi\in\mathbb{U}}[I_z(\phi) + F(\phi)] + \delta$. Since δ is arbitrary, the proof of the Laplace lower bound is complete. □

9.4 Proof of the Moderate Deviation Principle

In this section we prove Theorem 9.9. In order to show that I_z defined in (9.9) is a rate function on $\mathbb{U}$ and the family $\{I_z, z \in \mathscr{Z}\}$ of rate functions has compact level sets on compacts, we need to show that for every compact $K \subset \mathscr{Z}$ and $M < \infty$,

the set $\Lambda_{M,K}$ defined in (9.11), but with I_z as in (9.9), is compact. The proof of this property is exactly the same as that in the proof of Theorem 9.2, except we make use of part (a) of Condition 9.8 instead of the corresponding part of Condition 9.1. We omit the details.

We next prove the second statement in the theorem. Fix $z \in \mathscr{Z}$ and let $\{z^\varepsilon\}_{\varepsilon>0} \subset \mathscr{Z}$ be such that $z^\varepsilon \to z$ as $\varepsilon \to 0$. Fix $F \in \mathscr{C}_b(\mathbb{U})$. It suffices to prove the Laplace upper and lower bounds:

$$\limsup_{\varepsilon\to 0} \varkappa(\varepsilon) \log E \exp\left\{-\frac{1}{\varkappa(\varepsilon)} F(Y^\varepsilon_{z^\varepsilon})\right\} \le -\inf_{\phi\in\mathbb{U}} [I_z(\phi) + F(\phi)], \tag{9.16}$$

$$\liminf_{\varepsilon\to 0} \varkappa(\varepsilon) \log E \exp\left\{-\frac{1}{\varkappa(\varepsilon)} F(Y^\varepsilon_{z^\varepsilon})\right\} \ge -\inf_{\phi\in\mathbb{U}} [I_z(\phi) + F(\phi)]. \tag{9.17}$$

Proof of the Laplace upper bound. Since $Y^{\varepsilon,z^\varepsilon} \doteq \mathscr{K}^\varepsilon(z^\varepsilon, \sqrt{\varepsilon}W, \varepsilon N^{\varepsilon^{-1}})$, Theorem 8.19 implies

$$\begin{aligned} &-\varkappa(\varepsilon) \log E \exp\left\{-\frac{1}{\varkappa(\varepsilon)} F(Y^\varepsilon_{z^\varepsilon})\right\} \\ &= \inf_{u=(\psi,\varphi)\in\bar{\mathscr{A}}_b} E\left[\frac{1}{a^2(\varepsilon)} \bar{L}_T(u) + F \circ \mathscr{K}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}W^{\psi/\sqrt{\varepsilon}}, \varepsilon N^{\varphi/\varepsilon}\right)\right]. \end{aligned} \tag{9.18}$$

For later use, recall that by Theorem 8.3, $\bar{\mathscr{A}}_b$ in the representation can be replaced by $\bar{\mathscr{A}}$. Choose $\tilde{u}^\varepsilon = (\tilde{\psi}^\varepsilon, \tilde{\varphi}^\varepsilon) \in \bar{\mathscr{A}}_b$ such that

$$\begin{aligned} &-\varkappa(\varepsilon) \log E \exp\left\{-\frac{1}{\varkappa(\varepsilon)} F(Y^\varepsilon_{z^\varepsilon})\right\} + \varepsilon \\ &\ge E\left[\frac{1}{a^2(\varepsilon)}\left[L^W_T(\tilde{\psi}^\varepsilon) + L^N_T(\tilde{\varphi}^\varepsilon)\right] + F \circ \mathscr{K}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}W^{\tilde{\psi}^\varepsilon/\sqrt{\varepsilon}}, \varepsilon N^{\tilde{\varphi}^\varepsilon/\varepsilon}\right)\right]. \end{aligned} \tag{9.19}$$

Note that for $\varepsilon \in (0, 1)$,

$$\frac{1}{a^2(\varepsilon)} E\left[L^W_T(\tilde{\psi}^\varepsilon) + L^N_T(\tilde{\varphi}^\varepsilon)\right] \le \tilde{M} \doteq (2\|F\|_\infty + 1).$$

We would like to argue as in Theorem 8.4 that one can consider controls that are in a certain sense bounded, but in this case the bound should depend on ε. Fix $\delta > 0$ and define

$$\tau^\varepsilon \doteq \inf\left\{t \in [0, T] : L^W_t(\tilde{\psi}^\varepsilon) \ge a^2(\varepsilon)2M \text{ or } L^N_t(\tilde{\varphi}^\varepsilon) \ge a^2(\varepsilon)2M\right\} \wedge T,$$

where $M \doteq \tilde{M}\|F\|_\infty/\delta$. Let

$$\varphi^\varepsilon(s, y) \doteq \tilde{\varphi}^\varepsilon(y, s)1_{\{s\le\tau^\varepsilon\}} + 1_{\{s>\tau^\varepsilon\}}, \quad \psi^\varepsilon(s) \doteq \tilde{\psi}^\varepsilon(s)1_{\{s\le\tau^\varepsilon\}}$$

for $(s, y) \in \mathscr{X}_T$. Then $u^\varepsilon \doteq (\psi^\varepsilon, \varphi^\varepsilon) \in \bar{\mathscr{A}}_b$,

$$L_T^N(\varphi^\varepsilon) \le a^2(\varepsilon)2M, \quad L_T^W(\psi^\varepsilon) \le a^2(\varepsilon)2M,$$

and

$$P\{\varphi^\varepsilon \neq \tilde{\varphi}^\varepsilon \text{ or } \psi^\varepsilon \neq \tilde{\psi}^\varepsilon\} \le \frac{1}{a^2(\varepsilon)2M} E\left[L_T^W(\tilde{\psi}^\varepsilon) + L_T^N(\tilde{\varphi}^\varepsilon)\right] \le \frac{\delta}{2\|F\|_\infty}.$$

For $(s, y) \in \mathscr{X}_T$, define the rescaled and (for φ^ε) centered controls

$$\zeta^\varepsilon(s, y) \doteq \frac{\varphi^\varepsilon(s, y) - 1}{a(\varepsilon)}, \quad \theta^\varepsilon(s) \doteq \frac{\psi^\varepsilon(s)}{a(\varepsilon)}.$$

Fix any $\beta \in (0, 1]$. Applying part (d) of Lemma 9.6 yields

$$\begin{aligned} E\left[\frac{1}{a^2(\varepsilon)} \int_{\mathscr{X}_T} \ell(\tilde{\varphi}^\varepsilon) d\nu_T\right] &\ge E\left[\frac{1}{a^2(\varepsilon)} \int_{\mathscr{X}_T} \ell(\varphi^\varepsilon) 1_{\{|\zeta^\varepsilon| \le \beta/a(\varepsilon)\}} d\nu_T\right] \\ &\ge E\left[\int_{\mathscr{X}_T} \left(\frac{1}{2}(\zeta^\varepsilon)^2 - \kappa_3 a(\varepsilon)|\zeta^\varepsilon|^3\right) 1_{\{|\zeta^\varepsilon| \le \beta/a(\varepsilon)\}} d\nu_T\right] \\ &\ge \left(\frac{1}{2} - \kappa_3\beta\right) E\left[\int_{\mathscr{X}_T} (\zeta^\varepsilon)^2 1_{\{|\zeta^\varepsilon| \le \beta/a(\varepsilon)\}} d\nu_T\right]. \end{aligned} \tag{9.20}$$

Also, from the definition of τ^ε, it follows that

$$\begin{aligned} &E\left[F \circ \mathscr{K}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon} W^{\tilde{\psi}^\varepsilon/\sqrt{\varepsilon}}, \varepsilon N^{\tilde{\varphi}^\varepsilon/\varepsilon}\right) - F \circ \mathscr{K}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon} W^{\psi^\varepsilon/\sqrt{\varepsilon}}, \varepsilon N^{\varphi^\varepsilon/\varepsilon}\right)\right] \\ &\le 2\|F\|_\infty P\{\varphi^\varepsilon \neq \tilde{\varphi}^\varepsilon \text{ or } \psi^\varepsilon \neq \tilde{\psi}^\varepsilon\} \\ &\le \delta. \end{aligned}$$

The definition of τ^ε implies $\varphi^\varepsilon \in S^{N,\varepsilon}_{2M,+}$, which was defined in (9.7), and thus part (c) of Lemma 9.7 implies an upper bound of $2M\kappa_2(\beta)$ on the expected value in (9.20). Using the last two displays, (9.19), and $\kappa_2(1) \ge \kappa_2(\beta)$, we have

$$\begin{aligned} &-\varkappa(\varepsilon) \log E \exp\left\{-\frac{1}{\varkappa(\varepsilon)} F(Y^\varepsilon_{z^\varepsilon})\right\} \\ &\ge E\left[\frac{1}{2} \int_{\mathscr{X}_T} (\zeta^\varepsilon)^2 1_{\{|\zeta^\varepsilon| \le \beta/a(\varepsilon)\}} d\nu_T + L_T^W(\theta^\varepsilon)\right] \\ &\quad + E\left[F \circ \mathscr{K}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon} W^{\psi^\varepsilon/\sqrt{\varepsilon}}, \varepsilon N^{\varphi^\varepsilon/\varepsilon}\right)\right] - \delta - \varepsilon - 2\beta\kappa_3 M\kappa_2(1). \end{aligned} \tag{9.21}$$

Recall $\hat{S}_n$ defined in (9.10), and note that $\{\theta^\varepsilon, \zeta^\varepsilon 1_{\{|\zeta^\varepsilon| \le \beta/a(\varepsilon)\}}\}$ is a sequence in the compact set $\hat{S}_K$ for sufficiently large but finite K, and is therefore automatically tight. Let (θ, ζ) be a limit point along a subsequence that we index once more by

ε. By a standard argument by contradiction, it suffices to prove (9.16) along this subsequence. Using part (b) of Condition 9.8, we have that along this subsequence,

$$\mathscr{K}^{\varepsilon}\left(z^{\varepsilon}, \sqrt{\varepsilon}W^{\psi^{\varepsilon}/\sqrt{\varepsilon}}, \varepsilon N^{\varphi^{\varepsilon}/\varepsilon}\right) \Rightarrow \mathscr{K}^{0}(z,\theta,\zeta) \doteq \eta.$$

Hence taking limits in (9.21) along this subsequence yields

$$\begin{aligned}
&\liminf_{\varepsilon\to 0} -\varkappa(\varepsilon)\log E\exp\left\{-\frac{1}{\varkappa(\varepsilon)}F(Y^{\varepsilon}_{z^{\varepsilon}})\right\}\\
&\geq E\left[\frac{1}{2}(\|\theta\|^2_{W,2}+\|\zeta\|^2_{N,2})+F(\eta)\right]-\delta-\beta\kappa_3 M\kappa_2(1)\\
&\geq E\left[I_z(\eta)+F(\eta)\right]-\delta-\beta\kappa_3 M\kappa_2(1)\\
&\geq \inf_{\eta\in\mathbb{U}}\left[I_z(\eta)+F(\eta)\right]-\delta-\beta\kappa_3\kappa_2(1)\frac{\tilde{M}\|F\|_{\infty}}{\delta},
\end{aligned}$$

where the first line is from Fatou's lemma, and the second uses the definition of I_z in (9.9). Sending first β to 0 and then δ to 0 gives (9.16). □

Proof of the Laplace lower bound. For $\delta > 0$, there exists $\eta \in \mathbb{U}$ such that

$$I_z(\eta)+F(\eta) \leq \inf_{\eta\in\mathbb{U}}[I_z(\eta)+F(\eta)]+\delta/2. \tag{9.22}$$

Choose $(\theta,\zeta) \in S^{\mathscr{K}}_{z,\eta}$ such that

$$\frac{1}{2}(\|\theta\|^2_{W,2}+\|\zeta\|^2_{N,2}) \leq I_z(\eta)+\delta/2. \tag{9.23}$$

For $\beta \in (0,1]$, define

$$\zeta^{\varepsilon} \doteq \zeta 1_{\{|\zeta|\leq\beta/a(\varepsilon)\}}, \quad \varphi^{\varepsilon} \doteq 1+a(\varepsilon)\zeta^{\varepsilon}, \quad \psi^{\varepsilon} \doteq a(\varepsilon)\theta.$$

For every $\varepsilon > 0$, using $\zeta^{\varepsilon} = (\varphi^{\varepsilon}-1)/a(\varepsilon)$ and part (d) of Lemma 9.6, we have

$$\int_{\mathscr{X}_T}\ell(\varphi^{\varepsilon})d\nu_T \leq \kappa_3\int_{\mathscr{X}_T}(\varphi^{\varepsilon}-1)^2 d\nu_T = a^2(\varepsilon)\kappa_3\int_{\mathscr{X}_T}|\zeta^{\varepsilon}|^2 d\nu_T \leq a^2(\varepsilon)M,$$

where $M \doteq \kappa_3\int_{\mathscr{X}_T}|\zeta|^2 d\nu_T$. Thus $\varphi^{\varepsilon} \in \mathscr{U}^{\varepsilon}_{M,+}$, with this space defined in (9.8), for all $\varepsilon > 0$. Also

$$\zeta^{\varepsilon}1_{\{|\zeta^{\varepsilon}|\leq\beta/a(\varepsilon)\}} = \zeta 1_{\{|\zeta|\leq\beta/a(\varepsilon)\}},$$

which converges to ζ in $L^2(\nu_T)$ as $\varepsilon \to 0$. Thus by part (b) of Condition 9.8,

$$\mathscr{K}^{\varepsilon}\left(z^{\varepsilon}, \sqrt{\varepsilon}W^{\psi^{\varepsilon}/\sqrt{\varepsilon}}, \varepsilon N^{\varphi^{\varepsilon}/\varepsilon}\right) \Rightarrow \mathscr{K}^{0}(z,\theta,\zeta) = \eta. \tag{9.24}$$

Using part (d) of Lemma 9.6 and $\varkappa(\varepsilon)\varepsilon^{-1} = 1/a(\varepsilon)^2$, we obtain

$$\begin{aligned}\varkappa(\varepsilon)\varepsilon^{-1} L_T^N(\varphi^\varepsilon) &\le \frac{1}{2}\int_{\mathscr{X}_T} |\zeta^\varepsilon|^2 \, d\nu_T + \kappa_3 \int_{\mathscr{X}_T} a(\varepsilon)|\zeta^\varepsilon|^3 \, d\nu_T \\ &\le \frac{1}{2}(1 + 2\kappa_3\beta)\int_{\mathscr{X}_T} |\zeta|^2 \, d\nu_T.\end{aligned}$$

For φ^ε as defined in terms of ζ, there is no guarantee that $\varphi^\varepsilon \in \bar{\mathscr{A}}_b^N$. However, as noted previously, the variational representation (9.18) holds with $\bar{\mathscr{A}}_b$ replaced by $\bar{\mathscr{A}}$. Hence by the last display,

$$\begin{aligned}&-\varkappa(\varepsilon)\log E\exp\left\{-\frac{1}{\varkappa(\varepsilon)}F(Y_{z^\varepsilon}^\varepsilon)\right\} \\ &\le \frac{1}{a^2(\varepsilon)}\left[L_T^W(\psi^\varepsilon) + L_T^N(\varphi^\varepsilon)\right] + E\left[F\circ\mathscr{K}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}W^{\psi^\varepsilon/\sqrt{\varepsilon}}, \varepsilon N^{\varphi^\varepsilon/\varepsilon}\right)\right] \\ &\le \frac{1}{2}(\|\theta\|_{W,2}^2 + \|\zeta\|_{N,2}^2) + E\left[F\circ\mathscr{K}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}W^{\psi^\varepsilon/\sqrt{\varepsilon}}, \varepsilon N^{\varphi^\varepsilon/\varepsilon}\right)\right] \\ &\quad + \kappa_3\beta\int_{\mathscr{X}_T}|\zeta|^2\,d\nu_T.\end{aligned}$$

Taking the limit as $\varepsilon \to 0$ and using (9.24) yields

$$\begin{aligned}&\limsup_{\varepsilon\to 0} -\varkappa(\varepsilon)\log E\exp\left\{-\frac{1}{\varkappa(\varepsilon)}F(Y_{z^\varepsilon}^\varepsilon)\right\} \\ &\le \frac{1}{2}(\|\theta\|_{W,2}^2 + \|\zeta\|_{N,2}^2) + F(\eta) + \kappa_3\beta\int|\zeta|^2 d\nu_T.\end{aligned}$$

Finally, sending $\beta \to 0$ gives

$$\begin{aligned}\limsup_{\varepsilon\to 0} -\varkappa(\varepsilon)\log E\exp\left\{-\frac{1}{\varkappa(\varepsilon)}F(Y_{z^\varepsilon}^\varepsilon)\right\} &\le \frac{1}{2}(\|\theta\|_{W,2}^2 + \|\zeta\|_{N,2}^2) + F(\eta) \\ &\le I_z(\eta) + F(\eta) + \delta/2 \\ &\le \inf_{\eta\in\mathbb{U}}[I_z(\eta) + F(\eta)] + \delta,\end{aligned}$$

where the second inequality is from (9.23) and the last inequality follows from (9.22). Since $\delta > 0$ is arbitrary, this completes the proof of (9.17) and consequently the proof of Theorem 9.9. □

9.5 Notes

The sufficient condition for a Laplace principle given in Theorem 9.2, in the case that there is no Poisson noise, was established in [39], and the general case was treated in

[45], where its application to the study of a small noise LDP for finite dimensional jump-diffusions was studied as well. The sufficient condition given in Theorem 9.9 for the case in which the driving noise does not have a Gaussian component was given in [41]. This work also gave applications of Theorem 9.9 to the study of moderate deviation principles for finite and infinite dimensional stochastic dynamical systems driven by Poisson random measures.

The sufficient conditions given in this chapter have found applications in many different problems. Some of the works that have used the sufficient condition for Brownian motion functional given in [39] include [20–22, 37, 43, 44, 63, 64, 71, 91, 142, 143, 156, 191, 192, 196, 203, 207, 212–216, 218, 236, 244, 250, 254–256, 263, 268, 270]. Sufficient conditions given in this chapter for functionals of PRM and BM have been used in [12, 38, 47, 55, 70, 75, 251, 253, 258, 262, 264, 267, 272]. MDP sufficient conditions have found applications in [48, 57, 186–188, 194, 257, 265, 273].

Chapter 10
Large and Moderate Deviations for Finite Dimensional Systems

In this chapter we use the abstract sufficient conditions from Chap. 9 to prove large and moderate deviation principles for small noise finite dimensional jump-diffusions. We will consider only Laplace principles rather than uniform Laplace principles, since, as was noted in Chap. 9, the extension from the nonuniform to the uniform case is straightforward. The first general results on large deviation principles for jump-diffusions of the form considered in this chapter are due to Wentzell [245–248] and Freidlin and Wentzell [140]. The conditions for an LDP identified in the current chapter relax some of the assumptions made in these works. Results on moderate deviation principles in this chapter are based on the recent work [41]. We do not aim for maximal generality, and from the proofs it is clear that many other models (e.g., time inhomogeneous jump diffusions, SDEs with delay) can be treated in an analogous fashion.

The perspective here is different from that of the discrete time model of Chap. 4. In particular, the emphasis is on viewing the stochastic model as a relatively well behaved mapping on a fixed noise space consisting of a Brownian motion and one or more Poisson random measures. This has important consequences for models with degeneracy, i.e., systems for which the noise does not push the state in all directions. Processes of this sort in the general discrete time setting required complicated assumptions such as Condition 4.8 and a delicate mollification argument as in Sect. 4.8. In contrast, the degeneracy is essentially irrelevant when the process of interest can be viewed as a nice mapping (at least asymptotically) on a fixed noise space. This distinction becomes even more significant for the infinite dimensional models of Chap. 11, where the analogous degeneracy is ubiquitous. In this chapter we use Lipschitz continuity assumptions on the coefficients to guarantee that the mapping is well behaved. However, this is not necessary, especially with regard to Poisson noise, and for one such weakening we refer to Sect. 13.3.

The chapter is organized as follows. Section 10.1 introduces the basic stochastic process model that will be considered. Conditions under which the stochastic equation and its deterministic analogue have unique solutions are given. In Sect. 10.2 we use Theorem 9.2 to establish an LDP for the solution under additional integrability

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_10

conditions. Finally, in Sect. 10.3 we apply Theorem 9.9 to prove a moderate deviations result. For this result the integrability conditions we require are somewhat weaker, though additional smoothness conditions on the coefficients are assumed so that one can easily expand around the LLN limit, and the proof of tightness, as with the discrete time model of Chap. 5, is more involved than for the large deviation counterpart. We use the notation from Chap. 9, except that in this chapter, W is a finite dimensional standard Brownian motion, i.e., $\mathscr{H}_0 = \mathscr{H} = \mathbb{R}^d$, $\mathbb{W} = \mathscr{C}([0,T] : \mathbb{R}^d)$, and Λ is the identity operator.

10.1 Small Noise Jump-Diffusion

We consider small noise stochastic differential equations (SDEs) of the form

$$X^\varepsilon(t) = x_0 + \int_0^t b(X^\varepsilon(s))ds + \sqrt{\varepsilon}\int_0^t \sigma(X^\varepsilon(s))dW(s) + \varepsilon\int_{\mathscr{X}_t} G(X^\varepsilon(s-), y)N^{1/\varepsilon}(ds \times dy), \tag{10.1}$$

where W is a standard d-dimensional Wiener process, $N^{1/\varepsilon}$ is a PRM with intensity measure $\lambda_T \times \nu$ (see Definition 8.11) constructed from $\bar{N}$ as in (8.16) with $\varphi = 1/\varepsilon$, and $W, \bar{N}$ satisfy (a)–(c) in Sect. 9.1. The coefficients are assumed to satisfy the following condition.

Condition 10.1 *The functions $b : \mathbb{R}^d \to \mathbb{R}^d$, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, and $G : \mathbb{R}^d \times \mathscr{X} \to \mathbb{R}^d$ satisfy*

(a) *for some $L_b \in (0,\infty)$,*

$$\|b(x) - b(\bar{x})\| \le L_b\|x - \bar{x}\|,\ x, \bar{x} \in \mathbb{R}^d;$$

(b) *for some $L_\sigma \in (0,\infty)$,*

$$\|\sigma(x) - \sigma(\bar{x})\| \le L_\sigma\|x - \bar{x}\|,\ x, \bar{x} \in \mathbb{R}^d;$$

(c) *for some $L_G \in \mathscr{L}^1(\nu)$,*

$$\|G(x,y) - G(\bar{x},y)\| \le L_G(y)\|x - \bar{x}\|,\ x, \bar{x} \in \mathbb{R}^d,\ y \in \mathscr{X};$$

(d) *for some $M_G \in \mathscr{L}^1(\nu)$,*

$$\|G(x,y)\| \le M_G(y)(1 + \|x\|),\ x \in \mathbb{R}^d,\ y \in \mathscr{X}.$$

The following result follows by standard arguments (see Theorem IV.9.1 of [159]). It says that under Condition 10.1, Eq. (10.1) has a unique pathwise solution. In

applying results from Chap. 9, we take $\mathbb{U} = \mathscr{D}([0,T]:\mathbb{R}^d)$, i.e., the space of $\mathbb{R}^d$-valued right-continuous functions with left limits and the usual Skorokhod topology [24, Chap. 3, Sect. 12].

Theorem 10.2 *Fix $x_0 \in \mathbb{R}^d$, and assume Condition 10.1. Then for each $\varepsilon > 0$, there is a measurable map $\mathscr{G}^\varepsilon : \mathbb{V} \to \mathscr{D}([0,T]:\mathbb{R}^d)$ such that for every probability space $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{P})$ on which are given a d-dimensional Brownian motion $\tilde{W}$ and an independent Poisson random measure $\tilde{N}_\varepsilon$ on $\mathscr{X}_T$ with intensity measure $\varepsilon^{-1}\nu_T$, $\tilde{X}^\varepsilon \doteq \mathscr{G}^\varepsilon(\sqrt{\varepsilon}\tilde{W}, \varepsilon\tilde{N}_\varepsilon)$ is an $\tilde{\mathscr{F}}_t \doteq \sigma\{\tilde{W}(s), \tilde{N}_\varepsilon(B\times[0,s]), s \le t, B \in \mathscr{B}(\mathscr{X}), \nu(B) < \infty\}$ adapted process that is the unique solution of the stochastic integral equation*

$$\tilde{X}^\varepsilon(t) = x_0 + \int_0^t b(\tilde{X}^\varepsilon(s))ds + \sqrt{\varepsilon}\int_0^t \sigma(\tilde{X}^\varepsilon(s))d\tilde{W}(s) + \varepsilon\int_{\mathscr{X}_t} G(\tilde{X}^\varepsilon(s-), y)\tilde{N}_\varepsilon(ds\times dy), \tag{10.2}$$

for $t \in [0,T]$. In particular, $X^\varepsilon = \mathscr{G}^\varepsilon(\sqrt{\varepsilon}W, \varepsilon N^{1/\varepsilon})$ is the unique solution of (10.1).

10.2 An LDP for Small Noise Jump-Diffusions

The solution X^ε of (10.1) is a $\mathscr{D}([0,T]:\mathbb{R}^d)$-valued random variable. To prove a large deviation principle for $\{X^\varepsilon\}_{\varepsilon>0}$ as $\varepsilon \to 0$, we will assume the following additional condition on the coefficient function G. For $\rho \in (0,\infty)$, let $\mathscr{L}^\rho_{\exp}$ be the collection of all measurable $\theta : \mathscr{X} \to \mathbb{R}_+$ such that whenever $A \in \mathscr{B}(\mathscr{X})$ satisfies $\nu(A) < \infty$,

$$\int_A e^{\rho\theta(y)}\nu(dy) < \infty. \tag{10.3}$$

Let $\mathscr{L}_{\exp} \doteq \cap_{\rho\in(0,\infty)}\mathscr{L}^\rho_{\exp}$.

Condition 10.3 *$M_G \in \mathscr{L}_{exp}$ and $L_G \in \mathscr{L}^\rho_{exp}$ for some $\rho > 0$.*

Remark 10.4 This exponential integrability condition on jump distributions is a natural requirement for the model; it should be compared with Condition 4.3, assumed in the study of small noise discrete time Markov recursive systems. Consider, for example, the case in which $\mathscr{X} = \mathbb{R}^d$, $\nu \in \mathscr{P}(\mathbb{R}^d)$ satisfies $\int_{\mathbb{R}^d} e^{\langle\alpha,y\rangle}\nu(dy) < \infty$ for all $\alpha \in \mathbb{R}^d$, and for some $d \times d$ matrix A and a vector $w \in \mathbb{R}^d$,

$$G(x,y) = Ay + y\,\langle x, w\rangle\,.$$

Then G satisfies parts (c) and (d) of Condition 10.1, as well as Condition 10.3. In the general case (in which ν need not be a probability measure), note that the local rate function corresponding to just the jump part of the SDE (10.2) would be

$$L(x,\beta) = \inf\left[\int_{\mathscr{X}} \ell(g(y))\nu(dy) : \int_{\mathscr{X}} G(x,y)g(y)\nu(dy) = \beta\right]. \tag{10.4}$$

Using convex duality and that $\ell(b)$ is dual to $(e^a - 1)$, for this to be superlinear in β [a condition needed for the rate function on path space to have compact level sets in the usual topology of $\mathscr{D}([0,T] : \mathbb{R}^d)$], one needs

$$\int_{\mathscr{X}} \left[e^{\langle \alpha, G(x,y)\rangle} - 1\right]\nu(dy) < \infty \text{ for all } \alpha \in \mathbb{R}^d. \tag{10.5}$$

However, this follows under Conditions 10.1 and 10.3. One can break the integral in (10.5) according to $\{y : M_G(y) > 1\}$ and $\{y : M_G(y) \le 1\}$. Then $M_G \in \mathscr{L}^1(\nu)$ implies $\nu(\{y : M_G(y) > 1\}) < \infty$, and the integral over $\{y : M_G(y) > 1\}$ is finite due to Condition 10.3. The mean value theorem gives the bound $\|\alpha\| e^{\|\alpha\|} M_G(y)$ for the integrand on $\{y : M_G(y) \le 1\}$, and finiteness for the corresponding integral follows from $M_G \in \mathscr{L}^1(\nu)$.

We recall that $S \doteq \cup_{n\in\mathbb{N}} S_n$, with each $S_n \doteq S_n^W \times S_n^N$ compact in the appropriate topology. The proof of the following theorem proceeds using a standard argument and is given after Lemma 10.8.

Theorem 10.5 *Fix $x_0 \in \mathbb{R}^d$, and assume Conditions 10.1 and 10.3. Then for each $q = (f,g) \in S$, there is a unique $\xi = \xi_q \in \mathscr{C}([0,T] : \mathbb{R}^d)$ such that for all $t \in [0,T]$,*

$$\begin{aligned}\xi(t) = x_0 &+ \int_0^t b(\xi(s))ds + \int_0^t \sigma(\xi(s))f(s)ds \\ &+ \int_{\mathscr{X}_t} G(\xi(s),y)g(s,y)\nu(dy)ds.\end{aligned} \tag{10.6}$$

For $q = (f,g) \in S$, let $\xi = \xi_q$ denote the solution of (10.6). Let $I : \mathscr{D}([0,T] : \mathbb{R}^d) \to [0,\infty]$ be defined by

$$I(\phi) \doteq \inf_{q\in S:\phi=\xi_q} \bar{L}_T(q), \tag{10.7}$$

where $\bar{L}_T(q) \doteq L_T^W(f) + L_T^N(g)$, with the individual costs defined as in (9.1).

Theorem 10.6 *Assume Conditions 10.1 and 10.3. Then I is a rate function on $\mathscr{D}([0,T] : \mathbb{R}^d)$ and $\{X^\varepsilon\}_{\varepsilon>0}$ satisfies a large deviation principle on $\mathscr{D}([0,T] : \mathbb{R}^d)$ with rate function I.*

Following our standard convention, the proof is given for $T = 1$. Before proceeding with the proof, we present two lemmas. The first will be used to prove tightness. Recall that $g \in S_n^N$ means that $\int_{[0,T]\times\mathscr{X}} \ell(g(u,y))\nu(dy)du \le n$. For a function $f : [0,1] \to \mathbb{R}^k$, define $\|f\|_{\infty,t} \doteq \sup_{0\le s\le t} \|f(s)\|$ for $t \in [0,1]$. Note that

the constant $c(\delta, n)$ appearing in the lemma may also depend on the function θ. However, when the lemma is used, this will be a fixed quantity, such as $M_G(y)$, that is associated with a particular process model under consideration.

Lemma 10.7 *Let $\theta \in \mathscr{L}_{exp}$ and suppose that $\nu(\{\theta > 1\}) < \infty$. Then for every $\delta > 0$ and $n \in \mathbb{N}$, there exists $c(\delta, n) \in (1, \infty)$ such that for all $\tilde{\theta} : \mathscr{X} \to \mathbb{R}_+$ satisfying $\tilde{\theta} \le \theta$, every measurable map $f : [0, 1] \to \mathbb{R}_+$, and all $0 \le s \le t \le 1$,*

$$\sup_{g \in S_n^N} \int_{(s,t]\times\mathscr{X}} f(u)\tilde{\theta}(y)g(u, y)\nu(dy)du \\ \le c(\delta, n)\left(\int_{\mathscr{X}} \tilde{\theta}(y)\nu(dy)\right)\left(\int_s^t f(u)du\right) + \delta \|f\|_{\infty,1}. \tag{10.8}$$

Proof Let $f : [0, 1] \to \mathbb{R}_+$, $g \in S_n^N$, and $\delta > 0$ be given. Then for each $m \in (0, \infty)$,

$$\int_{(s,t]\times\mathscr{X}} f(u)\tilde{\theta}(y)g(u, y)\nu(dy)\,du = T_1(m) + T_2(m), \tag{10.9}$$

where

$$T_1(m) \doteq \int_{(s,t]\times\{\theta\le m\}} f(u)\tilde{\theta}(y)g(u, y)\nu(dy)du,$$

and

$$T_2(m) \doteq \int_{(s,t]\times\{\theta> m\}} f(u)\tilde{\theta}(y)g(u, y)\nu(dy)du.$$

Using part (a) of Lemma 9.6 with $\sigma = k$, $a = \theta(y)$ and $b = g(u, y)$, for each $k \in \mathbb{N}$ we have the bound

$$T_2(m) \le \|f\|_{\infty,1}\left(\int_{\{\theta>m\}} e^{k\theta(y)}\nu(dy) + \frac{n}{k}\right).$$

Also, for each $\beta \in (1, \infty)$, $T_1(m)$ can be bounded by

$$T_1(m) \le T_3(m, \beta) + T_4(m, \beta),$$

where

$$T_3(m, \beta) \doteq \int_{E_1(m,\beta)} f(u)\tilde{\theta}(y)g(u, y)\nu(dy)du,$$
$$T_4(m, \beta) \doteq \int_{E_2(m,\beta)} f(u)\tilde{\theta}(y)g(u, y)\nu(dy)du,$$

and

$$E_1(m,\beta) \doteq \{(u,y) \in (s,t] \times \mathscr{X} : \theta(y) \le m \text{ and } g(u,y) \le \beta\},$$
$$E_2(m,\beta) \doteq \{(u,y) \in (s,t] \times \mathscr{X} : \theta(y) \le m \text{ and } g(u,y) > \beta\}.$$

Using part (b) of Lemma 9.6 and that $g \in S_n^N$, we obtain

$$T_3(m,\beta) + T_4(m,\beta) \le \beta \left(\int_{\mathscr{X}} \tilde{\theta}(y) \nu(dy) \right) \left(\int_s^t f(u) du \right) + \bar{\kappa}_1(\beta) mn \, \|f\|_{\infty,1} \,. \tag{10.10}$$

Thus the left side of (10.8) can be bounded by

$$\beta \left(\int_{\mathscr{X}} \tilde{\theta}(y) \nu(dy) \right) \left(\int_s^t f(u) du \right) + \|f\|_{\infty,1} \left(\bar{\kappa}_1(\beta) mn + \int_{\{\theta > m\}} e^{k\theta(y)} \nu(dy) + \frac{n}{k} \right).$$

Given $\delta > 0$, choose $k \in \mathbb{N}$ such that $n/k < \delta/3$. Then use that $\theta \in \mathscr{L}_{\exp}$ to choose $m \in (0,\infty)$ such that $\int_{\{\theta>m\}} e^{k\theta(y)} \nu(dy) < \delta/3$. This is possible, since $\nu(\{\theta > 1\}) < \infty$. Finally, choose $\beta \in (1,\infty)$ such that $\bar{\kappa}_1(\beta) mn < \delta/3$. The result now follows by taking $c(\delta, n) = \beta$. □

The following lemma is proved similarly to Lemma 10.7 and therefore only a sketch is provided. Recall that $\mathscr{X}_1 \doteq [0,1] \times \mathscr{X}$ and $\nu_1 \in \mathscr{P}(\mathscr{X}_1)$ is the product measure $\nu_1(ds \times dy) = \nu(dy)ds$.

Lemma 10.8 *Let $\theta \in \mathscr{L}_{exp}^{\rho} \cap \mathscr{L}^1(\nu)$ for some $\rho \in (0,\infty)$. Then for every $n \in \mathbb{N}$,*

$$\sup_{g \in S_n^N} \int_{\mathscr{X}_1} \theta(y) g(u,y) \nu(dy) du < \infty.$$

Proof Consider the equality in (10.9) with $m = 1$, $s = 0$, $t = 1$, $f = 1$, and $\tilde{\theta} = \theta$. Then as in the proof of Lemma 10.7,

$$T_2(1) \le \int_{\{\theta>1\}} e^{\rho\theta(y)} \nu(dy) + \frac{n}{\rho}.$$

Also, as with the proof of (10.10),

$$T_1(1) \le \int_{\mathscr{X}} \theta(y) \nu(dy) + \bar{\kappa}_1(1) n.$$

The result follows by combining the two estimates. □

Proof of Theorem 10.5. Fix $q = (f,g) \in S$ and let $k \in \mathbb{N}$ be such that $q \in S_k$. We first prove the existence of a solution to (10.6). Consider a sequence $\{\phi_n\}$ in $\mathscr{C}([0,1] : \mathbb{R}^d)$

constructed recursively as follows. Define $\phi_1(t) \doteq x_0$ for all $t \in [0,1]$, and for $n \in \mathbb{N}$, let

$$\phi_{n+1}(t) \doteq x_0 + \int_0^t b(\phi_n(s))ds + \int_0^t \sigma(\phi_n(s))f(s)ds + \int_{\mathscr{X}_t} G(\phi_n(s), y)g(s, y)\nu(dy)ds, \ t \in [0,1]. \tag{10.11}$$

Using the growth conditions on b, σ, and G, we have that there is a $c_1 \in (0, \infty)$ such that for all $n \in \mathbb{N}$ and $t \in [0,1]$,

$$\|\phi_{n+1}\|_{\infty,t} \le \|x_0\| + c_1 \int_0^t (1 + \|\phi_n\|_{\infty,s})(1 + |f(s)|)ds + c_1 \int_{\mathscr{X}_t} M_G(y)(1 + \|\phi_n\|_{\infty,s})g(s, y)\nu(dy)ds.$$

Thus, with $h(s) \doteq (1 + |f(s)| + \int_{\mathscr{X}} M_G(y)g(s, y)\nu(dy))$, for some $c_2 \in (0, \infty)$, we have

$$\|\phi_{n+1}\|_{\infty,t} \le c_2 \left(1 + \int_0^t \|\phi_n\|_{\infty,s} h(s)ds\right), \ t \in [0,1], n \in \mathbb{N}.$$

From Lemma 10.8, we obtain $\int_{[0,1]} h(s)ds < \infty$. A standard recursive argument now shows that for all n, $\|\phi_n\|_{\infty,1} \le c_2 \exp \int_0^1 h(s)ds < \infty$.

Using Lemma 10.7, it is easily seen that for each $n \in \mathbb{N}$, $\phi_n \in \mathscr{C}([0,1] : \mathbb{R}^d)$. Indeed, the continuity of the last term in (10.11) follows on observing that for every $\delta > 0$ and $0 \le s \le t \le 1$,

$$\int_{(s,t]\times\mathscr{X}} \|G(\phi_n(u), y)\|g(u, y)\nu(dy)du \le (1 + \|\phi_n\|_{\infty,1}) \left(c(\delta, k)(t - s) \int_{\mathscr{X}} M_G(y)\nu(dy) + \delta\right).$$

For $n \in \mathbb{N}$ and $t \in [0,1]$, let $a_n(t) \doteq \|\phi_{n+1} - \phi_n\|_{\infty,t}$. Then there exists $c_3 \in (0, \infty)$ such that for all $n \ge 2$ and $t \in [0,1]$,

$$a_n(t) \le c_3 \int_0^t a_{n-1}(s)ds + c_3 \int_0^t a_{n-1}(s)|f(s)|ds + \int_0^t a_{n-1}(s) \left(\int_{\mathscr{X}} L_G(y)g(s, y)\nu(dy)\right) ds.$$

Thus, with $m(s) \doteq c_3(1 + |f(s)|) + \int_{\mathscr{X}} L_G(y)g(s, y)\nu(dy)$, we have for all $t \in [0,1]$ and $n \ge 2$ that

$$a_n(t) \le \int_0^t a_{n-1}(s)m(s)ds.$$

This shows that for all $n \in \mathbb{N}$,

$$a_{n+1}(1) \le a_1(1)\frac{\left(\int_0^1 m(s)ds\right)^n}{n!}.$$

Lemma 10.8 implies $\int_{\mathscr{X}_1} L_G(y)g(s,y)\nu(dy)ds < \infty$, and thus $\int_0^1 m(s)ds < \infty$. From this it follows that $\{\phi_n\}$ is a Cauchy sequence in $\mathscr{C}([0,1]:\mathbb{R}^d)$ and therefore must converge to some $\phi \in \mathscr{C}([0,1]:\mathbb{R}^d)$. From the continuity of b and σ it follows that for every $t \in [0,1]$,

$$\int_0^t b(\phi_n(s))ds + \int_0^t \sigma(\phi_n(s))f(s)ds \to \int_0^t b(\phi(s))ds + \int_0^t \sigma(\phi(s))f(s)ds.$$

Also,

$$\begin{aligned}\int_{\mathscr{X}_1} &\|G(\phi_n(s),y) - G(\phi(s),y)\|g(s,y)\nu(dy)ds \\ &\le \|\phi_n - \phi\|_{\infty,1}\int_{\mathscr{X}_1} L_G(y)g(s,y)\nu(dy)ds.\end{aligned}$$

Since $\int_{\mathscr{X}_1} L_G(y)g(s,y)\nu(dy)ds < \infty$, the right-hand side in the last display converges to 0 as $n \to \infty$. Combining these observations, we have that ϕ solves (10.6), proving the existence of solutions.

We now consider uniqueness. Suppose that ϕ_1, ϕ_2 are two solutions of (10.6) in $\mathscr{C}([0,1]:\mathbb{R}^d)$. Then using the Lipschitz property of b, σ, and G, we have that for some $c_4 \in (0,\infty)$ and all $t \in [0,1]$,

$$\begin{aligned}\|\phi_1 - \phi_2\|_{\infty,t} \le c_4\int_0^t \|\phi_1 - \phi_2\|_{\infty,s}ds + c_4\int_0^t \|\phi_1 - \phi_2\|_{\infty,s}|f(s)|ds \\ + \int_0^t \|\phi_1 - \phi_2\|_{\infty,s}\left(\int_{\mathscr{X}} L_G(y)g(s,y)\nu(dy)\right)ds.\end{aligned}$$

Thus

$$\|\phi_1 - \phi_2\|_{\infty,t} \le \int_0^t \|\phi_1 - \phi_2\|_{\infty,s}\left(c_4 + c_4|f(s)| + \int_{\mathscr{X}} L_G(y)g(s,y)\nu(dy)\right)ds.$$

Recalling that $\int_{\mathscr{X}_1} L_G(y)g(s,y)\nu(dy) < \infty$, an application of Gronwall's lemma implies $\phi_1 = \phi_2$. □

The following lemma is useful in characterizing the limit points of weakly converging controlled jump processes. It will be used in the proof of Theorem 10.6.

Recall that as discussed in Sect. 9.1, $g_k \to g$ in the topology of S_n^N if $\int_{\mathscr{X}_1} f g_k d\nu_1 \to \int_{\mathscr{X}_1} f g d\nu_1$ for bounded continuous f with compact support.

Lemma 10.9 *Fix $n \in \mathbb{N}$ and let K be a compact subset of $\mathscr{X}_1$. Let $g, g_k \in S_n^N$, $k \in \mathbb{N}$ be such that $g_k \to g$. Also, let $\gamma : \mathscr{X}_1 \to \mathbb{R}$ be a bounded measurable function. Then as $k \to \infty$,*

$$\int_K \gamma(s,y) g_k(s,y) \nu(dy) ds \to \int_K \gamma(s,y) g(s,y) \nu(dy) ds. \tag{10.12}$$

Proof We assume that $\nu_1(K) \neq 0$, since otherwise, the result is trivially true. By replacing, if needed, g_k and g with $g_k + 1_K$ and $g + 1_K$ respectively, we can assume without loss of generality that $\int_K g(s,y)\nu_1(ds \times dy) \neq 0$ and $\int_K g_k(s,y)\nu_1(ds \times dy) \neq 0$ for all $k \in \mathbb{N}$. If (10.12) holds with K replaced by K_1, where $K \subset K_1$ for some compact K_1, then by taking $\tilde{\gamma} = \gamma 1_K$, we see that (10.12) also holds with the compact set K. Also, since ν is finite on every compact set, we can always find a compact set $K_1 \supset K$ such that $\nu_1(\partial K_1) = 0$. Hence in proving the lemma, we can assume without loss of generality that $\nu_1(\partial K) = 0$. Recall from Chap. 9 that for $g \in S_n^N$, ν_1^g is defined by setting $\nu_1^g(A) = \int_A g(s,y)\nu_1(ds \times dy)$ for $A \in \mathscr{B}(\mathscr{X}_1)$. Define probability measures $\tilde{\nu}^k$ and $\tilde{\nu}$ as follows:

$$\tilde{\nu}^k(\cdot) \doteq \frac{\nu_1^{g_k}(\cdot \cap K)}{m_k}, \quad \tilde{\nu}(\cdot) \doteq \frac{\nu_1^{g}(\cdot \cap K)}{m},$$

where $m \doteq \nu_1^g(K)$ and $m_k \doteq \nu_1^{g_k}(K)$. Let $\theta(\cdot) \doteq \nu_1(\cdot \cap K)/\nu_1(K)$. Then

$$\begin{aligned} R(\tilde{\nu}^k \| \theta) &= \int_K \log\left(\frac{\nu_1(K)}{m_k} g_k(s,y)\right) \frac{1}{m_k} g_k(s,y)\nu_1(ds \times dy) \\ &= \frac{1}{m_k}\int_K [\ell(g_k(s,y)) + g_k(s,y) - 1]\,\nu_1(ds \times dy) + \log\frac{\nu_1(K)}{m_k} \\ &\leq \frac{n}{m_k} + 1 - \frac{\nu_1(K)}{m_k} + \log\frac{\nu_1(K)}{m_k}. \end{aligned}$$

Since $g_k \to g$ and $\nu_1(\partial K) = 0$, it follows that $m_k \to m$, and therefore the last display implies $\sup_{k\in\mathbb{N}} R(\tilde{\nu}^k \| \theta) < \infty$. Also note that $\tilde{\nu}^k$ converges weakly to $\tilde{\nu}$. From Lemma 2.5, it follows that

$$\frac{1}{m_k}\int_K \gamma(s,y) g_k(s,y)\nu_1(ds \times dy) \to \frac{1}{m}\int_K \gamma(s,y) g(s,y)\nu_1(ds \times dy),$$

which proves (10.12). □

10.2.1 Proof of the Large Deviation Principle

In this section we prove Theorem 10.6. Theorem 10.2 shows that the solution to the SDE (10.2) can be expressed as a measurable mapping on the input noises: $X^\varepsilon = \mathscr{G}^\varepsilon(\sqrt{\varepsilon}W, \varepsilon N^{1/\varepsilon})$. We now verify that $\mathscr{G}^\varepsilon$ satisfies Condition 9.1, which by Theorem 9.2 will complete the proof of the LDP. The notation used is that of Sect. 9.2. In particular, the spaces $\mathbb{V}$ and $\bar{\mathbb{V}}$ are the canonical spaces for a Brownian motion and PRM on $\mathscr{X}_1$ and $\mathscr{Y}_1$, respectively.

Define $\mathscr{G}^0 : \mathbb{V} \to \mathscr{D}([0,T] : \mathbb{R}^d)$ as follows. If $(w,m) \in \mathbb{V}$ is of the form $(w,m) = (\int_0^\cdot f(s)ds, \nu_1^g)$ for some $q = (f,g) \in S$, set

$$\mathscr{G}^0(w,m) = \mathscr{G}^0\left(\int_0^\cdot f(s)ds, \nu_1^g\right) = \xi_q,$$

where ξ_q is the unique solution of (10.6). For all other $(w,m) \in \mathbb{V}$ set $\mathscr{G}^0(w,m) = 0$. Since $\bar{L}_T(q) = \infty$ for such q, with this definition, I defined in (10.7) is the same as the function I defined in (9.4).

We will show that part (b) of Condition 9.1, which is the weak convergence of the controlled processes as $\varepsilon \to 0$, holds with this choice of $\mathscr{G}^0$. Part (a) of the condition follows if we prove continuity of $q \mapsto \mathscr{G}^0(q)$ for q such that $\bar{L}_T(q) \le n$ (recall that the initial condition has been assumed fixed). This is in fact an easier deterministic analogue of the proof of part (b), and hence omitted (see, for example, the proof of Theorem 11.25). Fix $n \in \mathbb{N}$ and let $u^\varepsilon = (\psi^\varepsilon, \varphi^\varepsilon) \in \bar{\mathscr{A}}_{b,n}$, $u = (\psi, \varphi) \in \bar{\mathscr{A}}_{b,n}$ be such that u^ε converges in distribution to u as $\varepsilon \to 0$. We recall that this implies the a.s. bounds

$$\int_0^1 \|\psi^\varepsilon(s)\|^2 ds \le n \text{ and } \int_{\mathscr{X}_1} \ell(\varphi^\varepsilon(s,x))\nu_1(ds \times dx) \le n. \tag{10.13}$$

Furthermore, almost surely $\varphi^\varepsilon(s,x)$ has upper and lower bounds of the form $1/\delta$ and δ for all x in some compact set K, and $\varphi^\varepsilon(s,x) = 1$ for $x \notin K$ (where $\delta > 0$ and K depend on φ^ε). The analogous statements also hold for (ψ, φ), and as we will see, the ability to restrict to controls with such nice properties greatly simplifies the arguments. Almost all of the difficulties in the proof are due to the jump term [for comparison, one can consider the proof of the analogous diffusion model in Theorem 3.19].

Let $\tilde{\varphi}^\varepsilon = 1/\varphi^\varepsilon$ and for $t \in [0,1]$ define

$$\mathscr{E}_1^\varepsilon(t) \doteq \exp\left[\int_{\mathscr{X}_t \times [0,\infty)} 1_{[0,\varphi^\varepsilon(s,y)/\varepsilon]}(r)\log(\tilde{\varphi}^\varepsilon(s,y))\bar{N}(ds \times dy \times dr) \right.$$
$$\left. + \int_{\mathscr{X}_t \times [0,\infty)} 1_{[0,\varphi^\varepsilon(s,y)/\varepsilon]}(r)(-\tilde{\varphi}^\varepsilon(s,y)+1)\bar{\nu}_1(ds \times dy \times dr)\right]$$

and

$$\mathscr{E}_2^\varepsilon(t) \doteq \exp\left[-\frac{1}{\sqrt{\varepsilon}}\int_0^t \psi^\varepsilon(s)dW(s) - \frac{1}{2\varepsilon}\int_0^t \|\psi^\varepsilon(s)\|^2 ds\right].$$

Let $\mathscr{E}^\varepsilon(t) = \mathscr{E}_1^\varepsilon(t)\mathscr{E}_2^\varepsilon(t)$. Then using the independence between the Brownian and Poisson noises, it follows that $\{\mathscr{E}^\varepsilon(t)\}_{0\le t\le 1}$ is a $\mathscr{F}_t$-martingale, and consequently

$$\bar{Q}^\varepsilon(A) = \int_A \mathscr{E}^\varepsilon(1)dP, \ \ A \in \mathscr{B}(\bar{\mathbb{V}})$$

defines a probability measure on $\bar{\mathbb{V}}$. The bounds (10.13) on ψ^ε and φ^ε along with the properties of φ^ε in terms of some compact set K noted below (10.13) imply that P and $\bar{Q}^\varepsilon$ are mutually absolutely continuous, and by Girsanov's theorem [see Theorem D.3],

$$\left(\sqrt{\varepsilon}W^{\psi^\varepsilon/\sqrt{\varepsilon}}, \varepsilon N^{\varphi^\varepsilon/\varepsilon}\right)$$

under $\bar{Q}^\varepsilon$ has the same probability law as $(\sqrt{\varepsilon}W, \varepsilon N^{1/\varepsilon})$ under P, where we recall $W^{\psi/\sqrt{\varepsilon}} \doteq W + \int_0^\cdot \psi(s)ds/\sqrt{\varepsilon}$. Thus it follows that $\bar{X}^\varepsilon = \mathscr{G}^\varepsilon(\sqrt{\varepsilon}W^{\psi^\varepsilon/\sqrt{\varepsilon}}, \varepsilon N^{\varphi^\varepsilon/\varepsilon})$ is the unique solution, under both $\bar{Q}^\varepsilon$ and P, of the controlled SDE given by $\bar{X}^\varepsilon(0) = x_0$ and

$$\begin{aligned} d\bar{X}^\varepsilon(t) &= \left[b(\bar{X}^\varepsilon(t)) + \sigma(\bar{X}^\varepsilon(t))\psi^\varepsilon(t)\right]dt + \sqrt{\varepsilon}\sigma(\bar{X}^\varepsilon(t))dW(t) \\ &\quad + \varepsilon\int_{\mathscr{X}} G(\bar{X}^\varepsilon(t-), x)N^{\varphi^\varepsilon/\varepsilon}(dt \times dx). \end{aligned} \tag{10.14}$$

We end this section by stating martingale bounds that will be useful in the sequel. Recall the definition of $\bar{\mathscr{A}}^N$ from Sect. 8.3. Also recall that for $\varphi \in \bar{\mathscr{A}}^N$, N_c^φ denotes the compensated form of N^φ.

Lemma 10.10 *Let $\varphi \in \bar{\mathscr{A}}^N$, and assume that $\psi : [0,1] \times \Omega \times \mathscr{X} \to \mathbb{R}$ is $\mathscr{P}\mathscr{F} \otimes \mathscr{B}(\mathscr{X})/\mathscr{B}(\mathbb{R})$-measurable and*

$$E\int_{\mathscr{X}_1} (|\psi(s,x)| \vee |\psi(s,x)|^2)\varphi(s,x)\nu_1(ds \times dx) < \infty.$$

Then there exists $C \in (0,\infty)$ such that for all $t \in [0,T]$

$$E\left[\sup_{0\le s\le t}\left|\int_{\mathscr{X}_t} \psi(s,x)N_c^\varphi(ds \times dx)\right|\right] \le CE\left[\int_{\mathscr{X}_t} \psi(s,x)^2\varphi(s,x)\nu_1(ds \times dx)\right]^{\frac{1}{2}},$$

and there is also the bound

$$E\left[\sup_{0\le s\le t}\left|\int_{\mathscr{X}_t} \psi(s,x)N_c^\varphi(ds \times dx)\right|^2\right] \le 4E\left[\int_{\mathscr{X}_t} \psi(s,x)^2\varphi(s,x)\nu_1(ds \times dx)\right].$$

Proof The bounds follow from Doob's maximal inequality and the Lenglart–Lepingle–Pratelli inequality [(D.2) and (D.4) of Appendix D], and from expressions for the quantities on the right-hand sides that are stated in Sect. D.2.2. In particular, in applying (D.2), we use that the expected quadratic variation of the stochastic integral in the last display is

$$E\int_{\mathscr{X}_t}\psi(s,x)^2N^{\varphi}(ds\times dx)=E\int_{\mathscr{X}_t}\psi(s,x)^2\varphi(s,x)\nu_1(ds\times dx).$$

□

10.2.1.1 Tightness

Lemma 10.11 *Assume Conditions 10.1 and 10.3. Given controls* $(\psi^\varepsilon,\varphi^\varepsilon)\in\bar{\mathscr{A}}_{b,n}$, *let* $\bar{X}^\varepsilon$ *be the corresponding unique solution to (10.14). Then* $\{\bar{X}^\varepsilon\}$ *is a tight family of* $\mathscr{D}([0,1]:\mathbb{R}^d)$*-valued random variables.*

Proof We begin with an estimate on the supremum of $\bar{X}^\varepsilon(t)$. Recalling $\|x\|_{\infty,t}\doteq\sup_{0\le s\le t}\|x(s)\|$, by Condition 10.1 we have that for suitable $c_1<\infty$,

$$\begin{aligned}\|\bar{X}^\varepsilon\|_{\infty,t}\le{}&\|x_0\|+c_1\int_0^t(1+\|\bar{X}^\varepsilon\|_{\infty,s})(1+\|\psi^\varepsilon(s)\|)ds\\&+\sqrt{\varepsilon}\left\|\int_0^{\cdot}\sigma(\bar{X}^\varepsilon(s))dW(s)\right\|_{\infty,t}\\&+\varepsilon\left\|\int_{\mathscr{X}_{\cdot}}M_G(y)(1+\|\bar{X}^\varepsilon(s-)\|)N_c^{\varphi^\varepsilon/\varepsilon}(ds\times dy)\right\|_{\infty,t}\\&+\int_{\mathscr{X}_t}M_G(y)(1+\|\bar{X}^\varepsilon\|_{\infty,s})\varphi^\varepsilon(s,y)\nu(dy)ds,\end{aligned}$$

where $N_c^{\varphi^\varepsilon/\varepsilon}(ds\times dy)=N^{\varphi^\varepsilon/\varepsilon}(ds\times dy)-\varepsilon^{-1}\varphi^\varepsilon(s,y)\nu_1(ds\times dy)$. Let

$$\begin{aligned}\mathscr{R}_t^\varepsilon\doteq{}&\sqrt{\varepsilon}\left\|\int_0^{\cdot}\sigma(\bar{X}^\varepsilon(s))dW(s)\right\|_{\infty,t}\\&+\varepsilon\left\|\int_{\mathscr{X}_{\cdot}}M_G(y)(1+\|\bar{X}^\varepsilon(s-)\|)N_c^{\varphi^\varepsilon/\varepsilon}(ds\times dy)\right\|_{\infty,t}.\end{aligned}\tag{10.15}$$

Using the bound (10.13) for ψ^ε and Hölder's inequality, Gronwall's inequality [Lemma E.2] gives

$$\begin{aligned}1+\|\bar{X}^\varepsilon\|_{\infty,1}&\le(1+\|x_0\|+\mathscr{R}_1^\varepsilon)\exp\left\{c_1(1+\sqrt{n})+\int_{\mathscr{X}_1}M_G(y)\varphi^\varepsilon(s,y)\nu(dy)ds\right\}\\&\le c_2(1+\|x_0\|+\mathscr{R}_1^\varepsilon),\end{aligned}\tag{10.16}$$

where, using Lemma 10.7 with $\delta = 1$, $f \equiv 1$ and $g = \varphi^\varepsilon$ and the fact that $M_G \in \mathscr{L}^1(\nu) \cap \mathscr{L}_{\exp}$, we obtain

$$c_2 \doteq \exp\left\{c_1(1+\sqrt{n}) + c(1,n)\int_{\mathscr{X}} M_G(y)\nu(dy) + 1\right\} < \infty.$$

Also, Condition 10.1 implies that for some $c_3 < \infty$,

$$\sqrt{\varepsilon}E\left\|\int_0^{\cdot}\sigma(\bar{X}^\varepsilon(s))dW(s)\right\|_{\infty,1} \le c_3\sqrt{\varepsilon}\left(E\|\bar{X}^\varepsilon\|_{\infty,1}+1\right).$$

Let $m \in (0,\infty)$. Then the expectation of the second term in the definition of $\mathscr{R}_t^\varepsilon$ can be bounded by the sum of

$$T_1^\varepsilon \doteq \varepsilon E\left\|\int_{\mathscr{X}_\cdot} M_G(y)1_{\{M_G\le m\}}(1+\|\bar{X}^\varepsilon(s-)\|)N_c^{\varphi^\varepsilon/\varepsilon}(ds\times dy)\right\|_{\infty,1} \tag{10.17}$$

and

$$T_2^\varepsilon \doteq 2E\int_{\mathscr{X}_t} M_G(y)1_{\{M_G\ge m\}}(1+\|\bar{X}^\varepsilon(s)\|)\varphi^\varepsilon(s,y)\nu(dy)ds. \tag{10.18}$$

For (10.18), we use the representation $N_c^{\varphi^\varepsilon/\varepsilon} = N^{\varphi^\varepsilon/\varepsilon} - \varepsilon^{-1}\varphi^\varepsilon\nu_1$, that the corresponding integrals are almost surely nondecreasing in t, and the identity

$$\begin{aligned}&\varepsilon E\int_{\mathscr{X}_t} M_G(y)1_{\{M_G\ge m\}}(1+\|\bar{X}^\varepsilon(s-)\|)N^{\varphi^\varepsilon/\varepsilon}(ds\times dy)\\ &\quad = E\int_{\mathscr{X}_t} M_G(y)1_{\{M_G\ge m\}}(1+\|\bar{X}^\varepsilon(s)\|)\varphi^\varepsilon(s,y)\nu(dy)ds.\end{aligned}$$

An application of Lemmas 10.7 and 10.10 as used before yield that for some $c_4 < \infty$,

$$\begin{aligned}T_1^\varepsilon &\le c_4\sqrt{\varepsilon m}E\left[(1+\|\bar{X}^\varepsilon\|_{\infty,1})\int_{\mathscr{X}_1} M_G(y)\varphi^\varepsilon(s,y)\nu(dy)ds\right]^{1/2}\\ &\le c_4\sqrt{\varepsilon m}(1+E\|\bar{X}^\varepsilon\|_{\infty,1})\left(c(1,n)\int_{\mathscr{X}} M_G(y)\nu(dy)+1\right)^{1/2}.\end{aligned}$$

Also, for every $\delta > 0$,

$$T_2^\varepsilon \le 2(1+E\|\bar{X}^\varepsilon\|_{\infty,1})\left(c(\delta,n)\int_{\mathscr{X}} M_G(y)1_{\{M_G\ge m\}}\nu(dy)+\delta\right).$$

Choosing $\delta > 0$ sufficiently small and then $m < \infty$ sufficiently large, there is $\varepsilon_0 > 0$ such that for all $\varepsilon \le \varepsilon_0$,

$$E\mathscr{R}_1^\varepsilon = c_3\sqrt{\varepsilon}\left(E\|\bar{X}^\varepsilon\|_{\infty,1} + 1\right) + T_1^\varepsilon + T_2^\varepsilon \le \frac{1}{2c_2}\left(E\|\bar{X}^\varepsilon\|_{\infty,1} + 1\right).$$

Using this in (10.16) then gives that for $\varepsilon \le \varepsilon_0$,

$$\bar{E}\|\bar{X}^\varepsilon\|_{\infty,1} \le 2c_2(\|x_0\| + 1),$$

and therefore

$$\sup_{\varepsilon\le\varepsilon_0} \bar{E}\|\bar{X}^\varepsilon\|_{\infty,1} < \infty. \tag{10.19}$$

Henceforth we consider only $\varepsilon < \varepsilon_0$. We next argue that $\mathscr{R}_1^\varepsilon$ defined in (10.15) converges to 0 in probability. The term with the Brownian motion is easy. Using Condition 10.1, the estimate in (10.19), and the Burkholder–Davis–Gundy inequality, it follows that

$$\sqrt{\varepsilon}\left\|\int_{[0,\cdot]} \sigma(\bar{X}^\varepsilon(s))dW(s)\right\|_{\infty,1} \to 0 \text{ in probability as } \varepsilon \to 0. \tag{10.20}$$

Next we consider the Poisson term, and write

$$\begin{aligned}&\varepsilon\int_{\mathscr{X}_t} G(\bar{X}^\varepsilon(s-), y)N^{\varphi^\varepsilon/\varepsilon}(ds\times dy) \\ &\quad= \varepsilon\int_{\mathscr{X}_t} G(\bar{X}^\varepsilon(s-), y)N_c^{\varphi^\varepsilon/\varepsilon}(ds\times dy) + \int_{\mathscr{X}_t} G(\bar{X}^\varepsilon(s), y)\varphi^\varepsilon(s, y)\nu(dy)ds.\end{aligned} \tag{10.21}$$

Consider the first term on the right side of (10.21). For $\alpha \in (0,\infty)$, define the stopping time $\tau_\alpha^\varepsilon \doteq \inf\{s : \|\bar{X}^\varepsilon(s)\| > \alpha\}$. We first show that

$$T^{\alpha,\varepsilon} \doteq \varepsilon\left\|\int_{(0,\tau_\alpha^\varepsilon\wedge\cdot]\times\mathscr{X}} G(\bar{X}^\varepsilon(s-), y)N_c^{\varphi^\varepsilon/\varepsilon}(ds\times dy)\right\|_{\infty,1}$$

converges to zero in probability as $\varepsilon \to 0$. To do this, note that for every $r \in (0,\infty)$,

$$T^{\alpha,\varepsilon} \le T_{\le r}^{\alpha,\varepsilon} + T_{>r}^{\alpha,\varepsilon},$$

where

$$T_{\le r}^{\alpha,\varepsilon} \doteq \varepsilon\left\|\int_{(0,\tau_\alpha^\varepsilon\wedge\cdot]\times\{M_G(y)\le r\}} G(\bar{X}^\varepsilon(s-), y)N_c^{\varphi^\varepsilon/\varepsilon}(ds\times dy)\right\|_{\infty,1},$$

$$T^{\alpha,\varepsilon}_{>r} \doteq \varepsilon \left\| \int_{(0,\tau^\varepsilon_\alpha \wedge \cdot] \times \{M_G(y) > r\}} G(\bar{X}^\varepsilon(s-), y) N_c^{\varphi^\varepsilon/\varepsilon}(ds \times dy) \right\|_{\infty,1}.$$

By Lemma 10.10, there is $c_5 \in (0, \infty)$ such that

$$E(T^{\alpha,\varepsilon}_{\leq r})^2 \leq \varepsilon c_5 E \int_{(0,\tau^\varepsilon_\alpha \wedge 1] \times \{M_G(y) \leq r\}} M_G^2(y)(1 + \|\bar{X}^\varepsilon(s)\|^2)\varphi^\varepsilon(s, y)\nu(dy)ds.$$

We then use that $\|\bar{X}^\varepsilon(s)\| \leq \alpha$ for $s \in (0, \tau^\varepsilon_\alpha \wedge 1)$ and write $(0, \tau^\varepsilon_\alpha \wedge 1] \times \{M_G(y) \leq r\}$ as the disjoint union of two sets, the first for which $\varphi^\varepsilon(s, y) < \beta$ and the second for which $\varphi^\varepsilon(s, y) \geq \beta$, and finally apply part (b) of Lemma 9.6 and use (10.13) to get

$$E(T^{\alpha,\varepsilon}_{\leq r})^2 \leq \varepsilon c_5(1 + \alpha^2)(\beta r \|M_G\|_1 + r^2 n \bar{\kappa}_1(\beta)).$$

To deal with the term $T^{\alpha,\varepsilon}_{>r}$, note that for every $k \in \mathbb{N}$,

$$\begin{aligned} ET^{\alpha,\varepsilon}_{>r} &\leq 2E \int_{(0,\tau^\varepsilon_\alpha \wedge \cdot] \times \{M_G(y) > r\}} \|G(\bar{X}^\varepsilon(s-), y)\|\varphi^\varepsilon(s, y)\nu(dy)ds \\ &\leq 2(1 + \alpha)E \int_{[0,1] \times \{M_G(y) > r\}} M_G(y)\varphi^\varepsilon(s, y)\nu(dy)ds \\ &\leq 2(1 + \alpha)\left(\int_{\{M_G(y) > r\}} e^{kM_G(y)}\nu(dy) + \frac{n}{k} \right), \end{aligned}$$

where the second inequality follows on using the growth bound stated in part (d) of Condition 10.1 and recalling the definition of τ^ε_α, and the last inequality is a consequence of part (a) of Lemma 9.6 with $\sigma = k$, $a = M_G(y)$ and $b = \varphi^\varepsilon(s, y)$ and again the fact that φ^ε takes values in S^N_n. Since $M_G \in \mathscr{L}^1(\nu) \cap \mathscr{L}_{\exp}$, for every $k \in \mathbb{N}$, $\int_{\{M_G(y) > r\}} e^{kM_G(y)}\nu(dy) \to 0$ as $r \to \infty$. Combining these two bounds and sending $\varepsilon \to 0$, $r \to \infty$, $k \to \infty$ in that order shows that for each $\alpha \in (0, \infty)$,

$$T^{\alpha,\varepsilon} \to 0 \text{ in probability as } \varepsilon \to 0. \tag{10.22}$$

Next let

$$T^\varepsilon \doteq \varepsilon \left\| \int_{(0,\cdot] \times \mathscr{X}} G(\bar{X}^\varepsilon(s-), y) N_c^{\varphi^\varepsilon/\varepsilon}(ds \times dy) \right\|_{\infty,1},$$

where the restriction on the time variable in $T^{\alpha,\varepsilon}$ has been dropped. Defining $A_\alpha \doteq \{\|\bar{X}^\varepsilon\|_{\infty,1} < \alpha\}$, for all $\eta > 0$,

$$\begin{aligned} P(T^\varepsilon > \eta) &= P(\{T^\varepsilon > \eta\} \cap A_\alpha) + P(\{T^\varepsilon > \eta\} \cap A^c_\alpha) \\ &\leq P(T^{\alpha,\varepsilon}_1 > \eta) + P(A^c_\alpha). \end{aligned}$$

Combining this last bound with (10.22) and (10.19), we see that $T^\varepsilon \to 0$ in probability as $\varepsilon \to 0$. Together with (10.20), this shows that

$$\mathscr{R}_1^\varepsilon \to 0 \text{ in probability as } \varepsilon \to 0.$$

Thus

$$\begin{aligned}\bar{X}^\varepsilon(t) = x_0 &+ \int_0^t b(\bar{X}^\varepsilon(s))ds + \int_0^t \sigma(\bar{X}^\varepsilon(s))\psi^\varepsilon(s)ds \\ &+ \int_{\mathscr{X}_t} G(\bar{X}^\varepsilon(s), y)\varphi^\varepsilon(s, y)\nu(dy)\, ds + \bar{\mathscr{R}}^\varepsilon(t),\end{aligned} \tag{10.23}$$

where $\|\bar{\mathscr{R}}^\varepsilon\|_{\infty,1} \le \mathscr{R}_1^\varepsilon$ converges to 0 in probability as $\varepsilon \to 0$. Tightness of the terms in (10.23) that involve b or σ follows from standard estimates, and are the same as calculations used for the small noise diffusion model in Chap. 3. Thus in order to prove tightness of $\{\bar{X}^\varepsilon\}$, it suffices to argue that

$$\xi^\varepsilon(t) = \int_{\mathscr{X}_t} G(\bar{X}^\varepsilon(s), y)\varphi^\varepsilon(s, y)\nu(dy)ds,\ t \in [0, 1]$$

is tight in $\mathbb{U} = \mathscr{D}([0, 1] : \mathbb{R}^d)$. By Lemma 10.7, for every $\delta > 0$ and $0 \le s \le t \le 1$,

$$\begin{aligned}\|\xi^\varepsilon(t) - \xi^\varepsilon(s)\| &\le \int_{[s,t]\times\mathscr{X}} (1 + \|\bar{X}^\varepsilon\|_{\infty,u})M_G(y)\varphi^\varepsilon(u, y)\nu(dy)du \\ &\le (1 + \|\bar{X}^\varepsilon\|_{\infty,1})c(\delta, n)(t - s)\|M_G\|_1 + \delta(1 + \|\bar{X}^\varepsilon\|_{\infty,1}),\end{aligned}$$

where as before, $\|\cdot\|_1$ is the norm in $\mathscr{L}^1(\nu)$. Tightness of ξ^ε is now a consequence of (10.19). Thus we have shown that $\{\bar{X}^\varepsilon\}$ is tight in $\mathscr{D}([0, 1] : \mathbb{R}^d)$. □

10.2.1.2 Identification of Limits

The following lemma completes the verification of part (b) of Condition 9.1, and hence the proof of Theorem 10.6.

Lemma 10.12 *Assume Conditions 10.1 and 10.3. Given controls $(\psi^\varepsilon, \varphi^\varepsilon) \in \bar{\mathscr{A}}_{b,n}$, let $\bar{X}^\varepsilon$ be the corresponding unique solution to (10.14). Assume that $(\psi^\varepsilon, \varphi^\varepsilon)$ converges in distribution to (ψ, φ). Then $\bar{X}^\varepsilon$ converges in distribution to the unique solution to (10.6) with $(f, g) = (\psi, \varphi)$.*

Proof From Lemma 10.11 it follows that if for some fixed n, the controls $(\psi^\varepsilon, \varphi^\varepsilon)$ are in $\bar{\mathscr{A}}_{b,n}$ for every $\varepsilon > 0$, then $\{\bar{X}^\varepsilon\}_{\varepsilon>0}$ is a tight collection of $\mathscr{D}([0, 1] : \mathbb{R}^d)$-valued random variables. It was also shown in the proof of the lemma that $\|\bar{\mathscr{R}}^\varepsilon\|_{\infty,1}$ appearing in (10.23) converges to 0 in probability as $\varepsilon \to 0$. It follows from this last property and (10.23) that $\bar{X}$ has continuous sample paths a.s. By appealing to the

Skorohod representation theorem, we assume without loss of generality the almost sure convergence $(\bar{X}^\varepsilon, \psi^\varepsilon, \varphi^\varepsilon, \bar{\mathscr{R}}^\varepsilon) \to (\bar{X}, \psi, \varphi, 0)$. Using the assumed conditions on b and σ, it is straightforward [see, for example, the proof of Lemma 3.21] using Hölder's inequality and the dominated convergence theorem to show that for every t, the sum of the first three terms on the right side of (10.23) converges a.s. to

$$x_0 + \int_0^t b(\bar{X}(s))ds + \int_0^t \sigma(\bar{X}(s))\psi(s)ds.$$

In view of the unique solvability of (10.6), to complete the verification that $\mathscr{G}^\varepsilon$ satisfies part (b) of Condition 9.1, it then suffices to show that for all $t \in [0, 1]$,

$$\int_{\mathscr{X}_t} G(\bar{X}^\varepsilon(s), y)\varphi^\varepsilon(s, y)\nu(dy)ds - \int_{\mathscr{X}_t} G(\bar{X}(s), y)\varphi(s, y)\nu(dy)ds \to 0 \tag{10.24}$$

as $\varepsilon \to 0$.

We write the expression in (10.24) as $T_3^\varepsilon(t) + T_4^\varepsilon(t)$, where

$$T_3^\varepsilon(t) \doteq \int_{\mathscr{X}_t} (G(\bar{X}^\varepsilon(s), y) - G(\bar{X}(s), y))\varphi^\varepsilon(s, y)\nu(dy)ds,$$
$$T_4^\varepsilon(t) \doteq \int_{\mathscr{X}_t} G(\bar{X}(s), y)(\varphi^\varepsilon(s, y) - \varphi(s, y))\nu(dy)ds.$$

Using Condition 10.1, we obtain

$$T_3^\varepsilon(t) \le \|\bar{X}^\varepsilon - \bar{X}\|_{\infty,1} \int_{\mathscr{X}_1} L_G(y)\varphi^\varepsilon(s, y)\nu(dy)ds.$$

Since $L_G \in \mathscr{L}^1(\nu) \cap \mathscr{L}^\rho_{\exp}$, we see from Lemma 10.8 that $T_3^\varepsilon(t) \to 0$ a.s. as $\varepsilon \to 0$. Let $\{K_r\}_{r\in\mathbb{N}}$ be a sequence of compact subsets of $\mathscr{X}$ such that $K_r \uparrow \mathscr{X}$ as $r \to \infty$, and let $E_r \doteq K_r \cap \{M_G \le r\}$. Write $T_4^\varepsilon(t) = T_{4,r\le}^\varepsilon(t) + T_{4,r>}^\varepsilon(t)$, where

$$T_{4,r\le}^\varepsilon(t) \doteq \int_{\mathscr{X}_t} G(\bar{X}(s), y)1_{E_r}(y)(\varphi^\varepsilon(s, y) - \varphi(s, y))\nu(dy)ds$$
$$T_{4,r>}^\varepsilon(t) \doteq \int_{\mathscr{X}_t} G(\bar{X}(s), y)1_{E_r^c}(y)(\varphi^\varepsilon(s, y) - \varphi(s, y))\nu(dy)ds.$$

Using Lemma 10.9, for every $r \in (0, \infty)$, $T_{4,r\le}^\varepsilon(t) \to 0$ as $\varepsilon \to 0$. Also, using Lemma 10.7 again, for every $\delta > 0$,

$$T^{\varepsilon}_{4,r>}(t) \le (1 + \|\bar{X}\|_{\infty,1}) \int_{\mathscr{X}_t} M_G(y) 1_{E_r^c}(y)(\varphi^{\varepsilon}(s,y) + \varphi(s,y))\nu(dy)ds$$
$$\le 2(1 + \|\bar{X}\|_{\infty,1}) \left(c(\delta,n)t \int_{\mathscr{X}} M_G(y) 1_{E_r^c}(y)\nu(dy) + \delta \right).$$

Since $M_G \in L^1(\nu)$, it follows that $\sup_{\varepsilon\in(0,\varepsilon_0)} T^{\varepsilon}_{4,r>}(t) \to 0$ if we first send $r \to \infty$ and then $\delta \to 0$. Combining these two estimates, we have that $T^{\varepsilon}_4(t)$ converges to 0 as $\varepsilon \to 0$. Thus we have proved (10.24), which completes the proof of the lemma and therefore, as noted previously, also the proof of Theorem 10.6. □

10.3 An MDP for Small Noise Jump-Diffusions

Throughout this section we assume Condition 10.1, which implies Lipschitz continuity and linear growth conditions on b, σ, and G in the x variable. Let $X^0 \in \mathscr{C}([0,T] : \mathbb{R}^d)$ be the unique solution of the equation

$$X^0(t) = x_0 + \int_0^t b(X^0(s))ds + \int_{\mathscr{X}_t} G(X^0(s), y)\nu(dy)\, ds,\ t \in [0,T].$$

We now establish a Laplace principle for $\{Y^{\varepsilon}\}$ with scaling function $\varkappa(\varepsilon)$, where

$$Y^{\varepsilon} = \frac{1}{a(\varepsilon)}(X^{\varepsilon} - X^0),$$

and as in (9.6), $a(\varepsilon)$ satisfies $a(\varepsilon) \to 0$ and $\varkappa(\varepsilon) \doteq \varepsilon/a^2(\varepsilon) \to 0$. For the MDP we assume some additional smoothness on the coefficients. For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}^d$ let $Df(x) = \left(\partial f_i(x)/\partial x_j\right)_{i,j}$. Following our convention, for matrices we use the operator norm, so that $\|Df(x)\| \doteq \sup_{w\in\mathbb{R}^d:\|w\|=1} \|Df(x)w\|$. Similarly, if $g : \mathbb{R}^d \times \mathscr{X} \to \mathbb{R}^d$ is differentiable in x for each fixed $y \in \mathscr{X}$ and $D_x g(x,y) = \left(\partial g_i(x,y)/\partial x_j\right)_{i,j}$, then $\|D_x g(x,y)\|$ denotes the norm of this matrix.

For the MDP, the integrability assumption on M_G in Condition 10.3 can be weakened, analogous to the corresponding weakening in going from the LDP to MDP in the setting of discrete time models (Chaps. 4 and 5). The following is the only assumption besides Condition 10.1 needed for the MDP.

Condition 10.13 (a) *The functions L_G and M_G are in $\mathscr{L}^1(\nu) \cap \mathscr{L}^{\rho}_{exp}$ for some $\rho \in (0,\infty)$.*

(b) *For every $y \in \mathscr{X}$, the maps $x \mapsto b(x)$ and $x \mapsto G(x,y)$ are differentiable. For some $L_{Db} \in (0,\infty)$,*

$$\|Db(x) - Db(\bar{x})\| \le L_{Db}\|x - \bar{x}\|,\ x, \bar{x} \in \mathbb{R}^d;$$

for some $L_{DG} \in \mathscr{L}^1(\nu)$,

$$\|D_x G(x, y) - D_x G(\bar{x}, y)\| \le L_{DG}(y)\,\|x - \bar{x}\|\,,\ x, \bar{x} \in \mathbb{R}^d,\ y \in \mathscr{X};$$

and lastly,

$$\sup_{\{x \in \mathbb{R}^d : \|x\| \le \|X^0\|_{\infty,T}\}} \int_{\mathscr{X}} \|D_x G(x, y)\| \nu(dy) < \infty.$$

Recall from Chap. 9 that we define $\mathscr{L}^2 \doteq \mathscr{L}^2([0, T] : \mathscr{H}_0) \times \mathscr{L}^2(\nu_T)$, and that in this chapter, $\mathscr{H}_0 = \mathbb{R}^d$. For $q = (f_1, f_2) \in \mathscr{L}^2$, consider the equation

$$\begin{aligned}\eta(t) = &\int_0^t [Db(X^0(s))]\eta(s)ds + \int_{\mathscr{X}_t} [D_x G(X^0(s), y)]\eta(s)\nu(dy)ds \\ &+ \int_0^t \sigma(X^0(s)) f_1(s)ds + \int_{\mathscr{X}_t} G(X^0(s), y) f_2(s, y)\nu(dy)ds. \end{aligned} \tag{10.25}$$

Since $M_G \in \mathscr{L}^1(\nu) \cap \mathscr{L}^{\rho}_{\exp} \subset \mathscr{L}^2(\nu)$, the last integral on the right side is finite by Hölder's inequality, and so under Condition 10.13, (10.25) has a unique solution $\eta_q \in \mathscr{C}([0, T] : \mathbb{R}^d)$. For $\eta \in \mathscr{D}([0, T] : \mathbb{R}^d)$, let

$$\bar{I}(\eta) \doteq \inf_{q=(f_1,f_2)\in\mathscr{L}^2 : \eta = \eta_q} \left[\frac{1}{2}\left(\|f_1\|^2_{W,2} + \|f_2\|^2_{N,2}\right)\right].$$

In particular, $\bar{I}(\eta) = \infty$ for all $\eta \in \mathscr{D}([0, T] : \mathbb{R}^d) \setminus \mathscr{C}([0, T] : \mathbb{R}^d)$.

Theorem 10.14 *Assume Conditions 10.1 and 10.13. Then $\{Y^\varepsilon\}_{\varepsilon>0}$ satisfies the Laplace principle in $\mathscr{D}([0, T] : \mathbb{R}^d)$ with scaling function $\varkappa(\varepsilon)$ and rate function $\bar{I}$.*

The following theorem gives an alternative expression for the rate function. From part (d) of Condition 10.1 and part (a) of Condition 10.13, it follows that $y \mapsto G_i(X^0(s), y)$ is in $\mathscr{L}^2(\nu)$ for all $s \in [0, T]$ and $i = 1, \ldots, d$, where $G = (G_1, \ldots, G_d)^T$. For $i = 1, \ldots, d$, let $e_i : \mathscr{X}_T \to \mathbb{R}$ be measurable functions such that for each $s \in [0, T]$, $\{e_i(s, \cdot)\}_{i=1}^d$ is an orthonormal collection in $\mathscr{L}^2(\nu)$ and the linear span of the collection contains that of $\{G_i(X^0(s), \cdot)\}_{i=1}^d$. Define $\bar{b}(x) \doteq \int_{\mathscr{X}} D_x G(x, y)\nu(dy)$, $x \in \mathbb{R}^d$, and define also $A : [0, T] \to \mathbb{R}^{d\times d}$ by

$$A_{ij}(s) \doteq \langle G_i(X^0(s), \cdot), e_j(s, \cdot)\rangle_{\mathscr{L}^2(\nu)},\ i, j = 1, \ldots, d,\ s \in [0, T], \tag{10.26}$$

where $\langle \cdot, \cdot \rangle_{\mathscr{L}^2(\nu)}$ is the inner product in $\mathscr{L}^2(\nu)$.

For $\eta \in \mathscr{D}([0, T] : \mathbb{R}^d)$, let

$$I(\eta) = \inf_{\tilde{q}=(\tilde{f}_1,\tilde{f}_2)} \left[\frac{1}{2}\left(\|\tilde{f}_1\|^2_2 + \|\tilde{f}_2\|^2_2\right)\right],$$

where the infimum is taken over all $\tilde{q} = (\tilde{f}_1, \tilde{f}_2)$, $\tilde{f}_1, \tilde{f}_2 \in \mathscr{L}^2([0, T] : \mathbb{R}^d)$ such that for $t \in [0, T]$,

$$\eta(t) = \int_0^t [Db(X^0(s)) + \bar{b}(X^0(s))]\eta(s)ds + \int_0^t \sigma(X^0(s))\tilde{f}_1(s)ds + \int_0^t A(s)\tilde{f}_2(s)ds. \tag{10.27}$$

Here $\|\cdot\|_2$ is the usual norm on $\mathscr{L}^2([0,T]:\mathbb{R}^d)$, and thus the same as $\|\cdot\|_{W,2}$. The proof of the following theorem is given in Sect. 10.3.3.

Theorem 10.15 *Under the conditions of Theorem 10.14, $I = \bar{I}$.*

Remark 10.16 Theorem 10.15 in particular says that the rate function for $\{Y^\varepsilon\}$ is the same as that appearing in the large deviation principle with scaling function ε for the Gaussian process

$$dZ^\varepsilon(t) = B(t)Z^\varepsilon(t)dt + \sqrt{\varepsilon}A(t)dW_1(t) + \sqrt{\varepsilon}\sigma(X^0(t))dW_2(t),\ \ Z^\varepsilon(0) = 0,$$

where W_1, W_2 are independent standard d-dimensional Brownian motions and $B(t) = Db(X^0(t)) + \bar{b}(X^0(t))$.

10.3.1 Some Preparatory Results

Following our standard convention, the proof is given for $T = 1$, and thus $\mathbb{U} = \mathscr{D}([0,1]:\mathbb{R}^d)$. From Theorem 10.2, it follows that there exists a measurable map $\mathscr{G}^\varepsilon : \mathbb{V} \to \mathbb{U}$ such that $X^\varepsilon = \mathscr{G}^\varepsilon(\sqrt{\varepsilon}W, \varepsilon N^{1/\varepsilon})$. Using $Y^\varepsilon = (X^\varepsilon - X^0)/a(\varepsilon)$, there is measurable $\mathscr{K}^\varepsilon$ such that $Y^\varepsilon = \mathscr{K}^\varepsilon(\sqrt{\varepsilon}W, \varepsilon N^{1/\varepsilon})$. Define $\mathscr{K}^0 : \mathscr{L}^2 \to \mathbb{U}$ by $\mathscr{K}^0(q) = \eta$ if η solves (10.25) for $q = (f_1, f_2) \in \mathscr{L}^2$. In order to prove Theorem 10.14, we will verify that Condition 9.8 holds with these choices of $\mathscr{K}^\varepsilon$ and $\mathscr{K}^0$.

The following lemma verifies a continuity property of $\mathscr{K}^0$. Recall the space $\hat{S}_n \doteq \{(f_1, f_2) \in \mathscr{L}^2 : \|f_1\|^2_{W,2} + \|f_2\|^2_{N,2} \le n\}$ introduced above Condition 9.8. This is viewed as a subset of the Hilbert space $\mathscr{L}^2$ defined there, and with respect to the topology of weak convergence in $\mathscr{L}^2$ is a compact Polish space. Together with the continuity established in Lemma 10.17, the compactness of $\hat{S}_n$ implies part (a) of Condition 9.8.

Lemma 10.17 *Suppose Condition 10.1 holds and $M_G \in \mathscr{L}^2(\nu)$. Fix $n \in (0,\infty)$ and let $q^k, q \in \hat{S}_n$, $k \in \mathbb{N}$ be such that $q^k \to q$. Let $\mathscr{K}^0(q) = \eta$, where η solves (10.25). Then $\mathscr{K}^0(q^k) \to \mathscr{K}^0(q)$.*

Proof Note that from part (d) of Condition 10.1 and since $M_G \in \mathscr{L}^2(\nu)$, the map $(s,y) \mapsto G(X^0(s), y)1_{[0,t]}(s)$ is in $\mathscr{L}^2(\nu_1)$. Let $q^k = (f_1^k, f_2^k)$ and $q = (f_1, f_2)$. Since $f_2^k \to f_2$, we have for every $t \in [0,1]$ that

$$\int_{\mathscr{X}_t} f_2^k(s,y)G(X^0(s),y)\nu(dy)ds \to \int_{\mathscr{X}_t} f_2(s,y)G(X^0(s),y)\nu(dy)ds. \tag{10.28}$$

We argue that the convergence is in fact uniform in t. Note that for $0 \le s \le t \le 1$,

$$\begin{aligned}
&\left\| \int_{[s,t]\times\mathscr{X}} f_2^k(u,y) G(X^0(u),y)\nu(dy)du \right\| \\
&\quad \le \left(1 + \|X^0\|_{\infty,1}\right) \int_{[s,t]\times\mathscr{X}} M_G(y)|f_2^k(u,y)|\nu(dy)du \\
&\quad \le \left(1 + \|X^0\|_{\infty,1}\right) |t-s|^{1/2}\sqrt{n}\|M_G\|_{\mathscr{L}^2(\nu)}, \qquad (10.29)
\end{aligned}$$

where $\|\cdot\|_{\mathscr{L}^2(\nu)}$ denotes the norm in $\mathscr{L}^2(\nu)$. This implies equicontinuity, and hence the convergence in (10.28) is uniform in $t \in [0,1]$.

Next, since $f_1^k \to f_1$, and since Condition 10.1 implies that $\sigma(\cdot)$ is continuous, it follows that for every $t \in [0,1]$,

$$\int_0^t \sigma(X^0(s)) f_1^k(s)ds \to \int_0^t \sigma(X^0(s)) f_1(s)ds.$$

Once again an equicontinuity estimate similar to (10.29) shows that the convergence is uniform. The conclusion of the lemma now follows from Gronwall's inequality. □

In order to verify part (b) of Condition 9.8, we first prove some a priori estimates. Recall the spaces $\mathscr{L}^\rho_{\exp}$ introduced in (10.3) and $S^{N,\varepsilon}_{n,+}$ and $S^{N,\varepsilon}_n$ in (9.7). Here $S^{N,\varepsilon}_{n,+}$ are controls for the Poisson noise with cost bounded by $na^2(\varepsilon)$, the scaling that is appropriate for an MDP, and $S^{N,\varepsilon}_n$ are the centered and rescaled versions of elements of $S^{N,\varepsilon}_{n,+}$.

Lemma 10.18 *Let $h \in \mathscr{L}^1(\nu) \cap \mathscr{L}^\rho_{exp}$ for some $\rho > 0$ and let I be a measurable subset of $[0,1]$. Let $n \in (0,\infty)$. Then there exist maps ϑ, ξ, ζ from $(0,\infty)$ to $(0,\infty)$ such that $\vartheta(u) \to 0$ as $u \to \infty$ and $\xi(u) \to 0$ as $u \to 0$, and for all $\varepsilon, \beta \in (0,\infty)$,*

$$\sup_{f\in S^{N,\varepsilon}_n} \int_{I\times\mathscr{X}} h(y)|f(s,y)|1_{\{|f|\ge\beta/a(\varepsilon)\}}\nu(dy)ds \le \sqrt{a(\varepsilon)}\vartheta(\beta) + (1+\lambda_1(I))\xi(\varepsilon)$$

and

$$\sup_{f\in S^{N,\varepsilon}_n} \int_{I\times\mathscr{X}} h(y)|f(s,y)|\nu(dy)\,ds \le \zeta(\beta)\lambda_1(I)^{1/2} + \sqrt{a(\varepsilon)}\vartheta(\beta) + (1+\lambda_1(I))\xi(\varepsilon),$$

where $\lambda_1(I)$ denotes the Lebesgue measure of I.

Proof Let $f \in S_n^{N,\varepsilon}$ and $\beta \in (0,\infty)$. Then

$$\int_{I\times\mathscr{X}} h(y)|f(s,y)|\nu(dy)ds \le \int_{I\times\mathscr{X}} h(y)|f(s,y)|1_{\{|f|\le\beta/a(\varepsilon)\}}\nu(dy)ds + \int_{I\times\mathscr{X}} h(y)|f(s,y)|1_{\{|f|\ge\beta/a(\varepsilon)\}}\nu(dy)ds. \tag{10.30}$$

Recall that $\mathscr{L}^1(\nu)\cap\mathscr{L}^\rho_{\exp}\subset\mathscr{L}^p(\nu)$ for all $p\ge 1$. By the Cauchy-Schwarz inequality and part (c) of Lemma 9.7, we have

$$\begin{aligned}&\int_{I\times\mathscr{X}} h(y)|f(s,y)|1_{\{|f|\le\beta/a(\varepsilon)\}}\nu(dy)ds \\ &\le \left(\lambda_1(I)\|h\|_2^2\int_{\mathscr{X}_1} f(s,y)^2 1_{\{|f|\le\beta/a(\varepsilon)\}}\nu(dy)ds\right)^{1/2} \\ &\le \|h\|_2(n\kappa_2(\beta))^{1/2}\lambda_1(I)^{1/2}.\end{aligned} \tag{10.31}$$

We now consider the second term on the right side of (10.30). We decompose h as $h1_{\{h\le 1/a(\varepsilon)^{1/2}\}} + h1_{\{h>1/a(\varepsilon)^{1/2}\}}$. Then using part (a) of Lemma 9.7, we obtain

$$\begin{aligned}\int_{I\times\mathscr{X}} h(y)1_{\{h\le 1/a(\varepsilon)^{1/2}\}}|f(s,y)|1_{\{|f|\ge\beta/a(\varepsilon)\}}\nu(dy)\,ds &\le \frac{1}{\sqrt{a(\varepsilon)}}na(\varepsilon)\kappa_1(\beta) \\ &= n\sqrt{a(\varepsilon)}\kappa_1(\beta).\end{aligned}$$

Also, letting $g = a(\varepsilon)f + 1$ and noting that the definition of $S_n^{N,\varepsilon}$ implies $g\ge 0$, we have

$$\begin{aligned}&\int_{I\times\mathscr{X}} h(y)1_{\{h>1/a(\varepsilon)^{1/2}\}}|f(s,y)|1_{\{|f|\ge\beta/a(\varepsilon)\}}\nu(dy)\,ds \\ &\le \frac{\lambda_1(I)}{a(\varepsilon)}\int_{\mathscr{X}} h(y)1_{\{h>1/a(\varepsilon)^{1/2}\}}\nu(dy) \\ &\quad + \frac{1}{a(\varepsilon)}\int_{I\times\mathscr{X}} h(y)1_{\{h>1/a(\varepsilon)^{1/2}\}}g(s,y)\nu(dy)\,ds.\end{aligned} \tag{10.32}$$

The first term on the right side can be bounded by

$$\lambda_1(I)C_1(\varepsilon) \doteq \lambda_1(I)\int_{\mathscr{X}} h(y)^3 1_{\{h>1/a(\varepsilon)^{1/2}\}}\nu(dy),$$

where $h\in\mathscr{L}^\rho_{\exp}\cap\mathscr{L}^1(\nu)$ implies $C_1(\varepsilon)\to 0$ as $\varepsilon\to 0$. The second term on the right side of (10.32) can be bounded, using part (a) of Lemma 9.6 with $a = \rho h(y)/2$, $b = g(s,y)$, $\sigma = 1$, and also using that $g\in S_{n,+}^{N,\varepsilon}$, by

$$\begin{aligned}
&\frac{2\lambda_1(I)}{\rho a(\varepsilon)} \int_{\mathscr{X}} e^{\rho h(y)/2} 1_{\{h>1/a(\varepsilon)^{1/2}\}} \nu(dy) + \frac{2}{\rho a(\varepsilon)} n a^2(\varepsilon) \\
&\quad \le \frac{2\lambda_1(I)}{\rho} \int_{\mathscr{X}} h(y)^2 e^{\rho h(y)/2} 1_{\{h>1/a(\varepsilon)^{1/2}\}} \nu(dy) + \frac{2na(\varepsilon)}{\rho} \\
&\quad = \lambda_1(I) C_2(\varepsilon) + \frac{2na(\varepsilon)}{\rho},
\end{aligned}$$

where $C_2(\varepsilon)$ converges to 0 as $\varepsilon \to 0$. Thus the second term on the right side of (10.30) can be bounded by

$$\sqrt{a(\varepsilon)}\vartheta(\beta) + (1 + \lambda_1(I))\xi(\varepsilon),$$

where

$$\vartheta(\beta) = n\kappa_1(\beta), \ \xi(\varepsilon) = C_1(\varepsilon) + C_2(\varepsilon) + \frac{2na(\varepsilon)}{\rho}.$$

This gives the first bound of the lemma. The second bound also follows with these choices of ξ, ϑ, and $\zeta(\beta) = \|h\|_2 (n\kappa_2(\beta))^{1/2}$ using (10.31). □

The following lemma is proved in a fashion similar to that of Lemma 10.7, and so only a sketch is given.

Lemma 10.19 *Let $\theta \in \mathscr{L}^{\rho}_{exp}$ for some $\rho > 0$ and suppose that $\nu(\{\theta > 1\}) < \infty$. Then for every $\delta > 0$ and $n \in \mathbb{N}$, there exists $\tilde{c}(\delta, n) \in (0, \infty)$ such that for all measurable maps $\tilde{\theta} : \mathscr{X} \to \mathbb{R}_+$ satisfying $\tilde{\theta} \le \theta$, any measurable $f : [0, 1] \to \mathbb{R}_+$, and all $0 \le s \le t \le 1$,*

$$\begin{aligned}
&\sup_{g \in S^{N,\varepsilon}_{n,+}} \int_{(s,t]\times\mathscr{X}} f(u)\tilde{\theta}(y) g(u, y) \nu(dy) du \\
&\quad \le \tilde{c}(\delta, n) \left(\int_{\mathscr{X}} \tilde{\theta}(y) \nu(dy) \right) \left(\int_s^t f(u) du \right) + (\delta + n\rho^{-1} a^2(\varepsilon)) \|f\|_{\infty,1}.
\end{aligned}$$

Proof Let $f : [0, 1] \to \mathbb{R}_+$ and $g \in S^{N,\varepsilon}_{n,+}$. For $m \in (0, \infty)$, let $T_i(m)$, $i = 1, 2$, be as in the proof of Lemma 10.7. Then using part (a) of Lemma 9.6 with $a = \rho\theta(y)$, $b = g(u, y)$, and $\sigma = 1$, we can bound $T_2(m)$ as

$$T_2(m) \le \frac{\|f\|_{\infty,1}}{\rho} \left(\int_{\{\theta>m\}} e^{\rho\theta(y)} \nu(dy) + n a^2(\varepsilon) \right).$$

Also, as in the proof of Lemma 10.7,

$$T_1(m) \le \beta \left(\int_{\mathscr{X}} \tilde{\theta}(y) \nu(dy) \right) \left(\int_s^t f(u) du \right) + \bar{\kappa}_1(\beta) m n a^2(\varepsilon) \|f\|_{\infty,1},$$

where $\bar{\kappa}_1(\beta) \to 0$ as $\beta \to \infty$. The result now follows on recalling (10.9) and choosing first m sufficiently large and then β sufficiently large. □

Recall the map $\mathscr{K}^\varepsilon$ introduced at the beginning of Sect. 10.3.1 and the definition of $\mathscr{U}^\varepsilon_{n,+}$ in (9.8), and recall that by (9.7), $\mathscr{U}^\varepsilon_{n,+}$ is the class of controls for both types of noise for which the cost scales proportionally with $a^2(\varepsilon)$. Since $\mathscr{U}^\varepsilon_{n,+}$ is contained in $\bar{\mathscr{A}}_{b,\bar{n}}$ for some $\bar{n} \in \mathbb{N}$, it follows from Sect. 10.2.1 that for $n \in \mathbb{N}$ and $u = (\zeta, \varphi) \in \mathscr{U}^\varepsilon_{n,+}$, the equation

$$d\bar{X}^\varepsilon(t) = \big[b(\bar{X}^\varepsilon(t)) + \sigma(\bar{X}^\varepsilon(t))\zeta(t)\big]\,dt + \sqrt{\varepsilon}\sigma(\bar{X}^\varepsilon(t))dW(t) + \varepsilon \int_{\mathscr{X}} G(\bar{X}^\varepsilon(t-), x) N^{\varphi/\varepsilon}(dt \times dx),$$

$\bar{X}^\varepsilon(0) = x_0$ has a unique solution.

Define $\bar{Y}^\varepsilon = \mathscr{K}^\varepsilon(\sqrt{\varepsilon}W^{\zeta/\sqrt{\varepsilon}}, \varepsilon N^{\varphi/\varepsilon})$, and note that using Girsanov's theorem [Theorem D.3] as in the proof of Theorem 3.19 yields

$$\bar{Y}^\varepsilon = \frac{1}{a(\varepsilon)}(\bar{X}^\varepsilon - X^0). \tag{10.33}$$

The following moment bound on $\bar{X}^\varepsilon$ follows along the lines of the proof of (10.19).

Lemma 10.20 *Assume Conditions 10.1 and 10.13. For every $n \in \mathbb{N}$, there exists an $\varepsilon_0 \in (0, 1)$ such that*

$$\sup_{\varepsilon \in (0,\varepsilon_0)} \sup_{u=(\zeta,\varphi)\in \mathscr{U}^\varepsilon_{n,+}} E\|\bar{X}^\varepsilon\|_{\infty,1} < \infty.$$

Proof Using the same argument as that used to establish (10.16), for all $\varepsilon \in (0, 1)$, we have

$$\|\bar{X}^\varepsilon\|_{\infty,1} \le (1 + \|x_0\| + \tilde{\mathscr{R}}^\varepsilon_1) \exp\left\{c_1(1 + \sqrt{n}a_1) + \int_{\mathscr{X}_1} M_G(y)\varphi(s, y)\nu(dy)\,ds\right\}, \tag{10.34}$$

where $a_1 \doteq \sup_{\varepsilon\in(0,1)} a(\varepsilon)$ and

$$\tilde{\mathscr{R}}^\varepsilon_1 \doteq \sqrt{\varepsilon}\left\|\int_0^\cdot \sigma(\bar{X}^\varepsilon(s))dW(s)\right\|_{\infty,1} + \varepsilon\left\|\int_{\mathscr{X}_\cdot} M_G(y)(1 + \|\bar{X}^\varepsilon(s-)\|)N^{\varphi/\varepsilon}_c(ds \times dy)\right\|_{\infty,1}.$$

Using Lemma 10.19 with $\delta = 1$, $f \equiv 1$ and recalling that $\varphi \in S^{N,\varepsilon}_{n,+}$ and the fact that $M_G \in \mathscr{L}^1(\nu) \cap \mathscr{L}_{\exp}$, we obtain

$$\|\bar{X}^\varepsilon\|_{\infty,1} \le c_2(1 + \|x_0\| + \tilde{\mathscr{R}}_1^\varepsilon), \tag{10.35}$$

where

$$c_2 \doteq \exp\left\{c_1(1+\sqrt{n}a_1) + \tilde{c}(1,n)\int_{\mathscr{X}} M_G(y)\nu(dy) + 1 + \frac{n}{\rho}a_1^2\right\} < \infty.$$

We split the expected value of the second term in the definition of $\tilde{\mathscr{R}}_1^\varepsilon$ as $T_1^\varepsilon + T_2^\varepsilon$, where T_i^ε are just as in (10.17)–(10.18) in the proof of the corresponding LDP, and then follow the same procedure as that below (10.17)–(10.18) to bound the two terms. In this case, however, we use Lemma 10.19 rather than Lemma 10.7, and find that for some $c_3 \in (0,\infty)$,

$$T_1^\varepsilon \le c_3\sqrt{\varepsilon}m(1 + E\|\bar{X}^\varepsilon\|_{\infty,1})\left(\tilde{c}(1,n)\int_{\mathscr{X}} M_G(y)\nu(dy) + 1 + \frac{n}{\rho}a_1^2\right)^{1/2},$$

and for every $\delta > 0$ and $\varepsilon \in (0,1)$,

$$T_2^\varepsilon \le 2(1 + E\|\bar{X}^\varepsilon\|_{\infty,1})\left(\tilde{c}(\delta,n)\int_{\mathscr{X}} M_G(y)1_{\{M_G\ge m\}}\nu(dy) + \delta + \frac{n}{\rho}a^2(\varepsilon)\right).$$

Choosing first δ sufficiently small, next m sufficiently large, and finally ε_0 sufficiently small, we have for all $\varepsilon \le \varepsilon_0$ that

$$E\tilde{\mathscr{R}}_1^\varepsilon \le \frac{1}{2c_2}\left(E\|\bar{X}^\varepsilon\|_{\infty,1} + 1\right).$$

The result now follows on using this estimate in (10.35). □

The following tightness property plays a key role in the proof of Theorem 10.14.

Lemma 10.21 *Suppose Conditions 10.1 and 10.13 hold, and define $\bar{Y}^\varepsilon$ by (10.33). For every $n \in \mathbb{N}$, there exists an $\varepsilon_1 \in (0,1)$ such that*

$$\left\{\|\bar{Y}^\varepsilon\|_{\infty,1},\ u \in \mathscr{U}_{n,+}^\varepsilon,\ \varepsilon \in (0,\varepsilon_1)\right\}$$

is a tight collection of $\mathbb{R}_+$-valued random variables.

Proof Let $u = (\zeta,\varphi) \in \mathscr{U}_{n,+}^\varepsilon$ and let $\psi \doteq (\varphi - 1)/a(\varepsilon)$. Then

$$\begin{aligned}\bar{X}^\varepsilon(t) - X^0(t) = &\int_0^t \left(b(\bar{X}^\varepsilon(s)) - b(X^0(s))\right)ds + \sqrt{\varepsilon}\int_0^t \sigma(\bar{X}^\varepsilon(s))dW(s)\\ &+ \int_{\mathscr{X}_t} \varepsilon G(\bar{X}^\varepsilon(s-),y)N_c^{\varphi/\varepsilon}(ds\times dy)\\ &+ \int_{\mathscr{X}_t}\left(G(\bar{X}^\varepsilon(s),y) - G(X^0(s),y)\right)\varphi(s,y)\nu(dy)ds\end{aligned}$$

$$+\int_{\mathscr{X}_t} G(X^0(s),y)(\varphi(s,y)-1)\nu(dy)ds$$
$$+\int_0^t \sigma(\bar{X}^\varepsilon(s))\zeta(s)ds.$$

Write $\bar{Y}^\varepsilon = (\bar{X}^\varepsilon - X^0)/a(\varepsilon)$ as

$$\bar{Y}^\varepsilon = M^\varepsilon + A^\varepsilon + B^\varepsilon + \mathscr{E}^\varepsilon + C^\varepsilon, \tag{10.36}$$

where for $t \in [0,1]$,

$$M^\varepsilon(t) \doteq \frac{\varepsilon}{a(\varepsilon)}\int_{\mathscr{X}_t} G(\bar{X}^\varepsilon(s-),y)N_c^{\varphi/\varepsilon}(ds\times dy) + \left(\frac{\varepsilon}{a(\varepsilon)}\right)^{1/2}\int_0^t \sigma(\bar{X}^\varepsilon(s))dW(s)$$
$$A^\varepsilon(t) \doteq \frac{1}{a(\varepsilon)}\int_0^t \left(b(\bar{X}^\varepsilon(s)) - b(X^0(s))\right)ds,$$
$$B^\varepsilon(t) \doteq \frac{1}{a(\varepsilon)}\int_{\mathscr{X}_t} \left(G(\bar{X}^\varepsilon(s),y) - G(X^0(s),y)\right)\nu(dy)ds,$$
$$\mathscr{E}^\varepsilon(t) \doteq \int_{\mathscr{X}_t} \left(G(\bar{X}^\varepsilon(s),y) - G(X^0(s),y)\right)\psi(s,y)\nu(dy)ds,$$
$$C^\varepsilon(t) \doteq \int_{\mathscr{X}_t} G(X^0(s),y)\psi(s,y)\nu(dy)ds + \frac{1}{a(\varepsilon)}\int_0^t \sigma(\bar{X}^\varepsilon(s))\zeta(s)ds.$$

With ε_0 as in Lemma 10.20, we have from the Burkholder–Davis–Gundy inequality (see Sect. D.1) that

$$\left\{\left\|\int_0^\cdot \sigma(\bar{X}^\varepsilon(s))dW(s)\right\|_{\infty,1}\right\}_{\varepsilon\le\varepsilon_0}$$

is tight. Also, as in the proof of Lemma 10.20, for some $c_1 \in (0,\infty)$, we have

$$E\left\|\int_{\mathscr{X}_\cdot} G(\bar{X}^\varepsilon(s-),y)N_c^{\varphi/\varepsilon}(ds\times dy)\right\|_{\infty,1} \le (1+E\|\bar{X}^\varepsilon\|_{\infty,1})$$
$$\times\left(c_1\sqrt{\varepsilon}m + 2\tilde{c}(\delta,n)\int_{\mathscr{X}} M_G(y)1_{\{M_G\ge m\}}\nu(dy) + 2\delta + \frac{2n}{\rho}a^2(\varepsilon)\right)$$

for every $\delta>0$, $\varepsilon\in(0,\varepsilon_0)$, and $u\in\mathscr{U}^\varepsilon_{n,+}$. Combining these two estimates, we see that $\left\{\|M^\varepsilon\|_{\infty,1}\right\}_{\varepsilon\le\varepsilon_0}$ is tight, and since $\varepsilon/a(\varepsilon)\to 0$

$$\|M^\varepsilon\|_{\infty,1} \to 0 \text{ in probability} \tag{10.37}$$

as $\varepsilon\to 0$.

In the rest of the proof, we will show upper bounds of the form c, $c\int_0^t \|\bar{Y}^\varepsilon\|_{\infty,s}ds$ or $ca(\varepsilon)\|\bar{Y}^\varepsilon\|_{\infty,1}$ for each of the remaining terms in (10.36). By the Lipschitz condition

on G [part (c) of Condition 10.1 and part (a) of Condition 10.13], there is $c_2 \in (0, \infty)$ such that for all $t \in [0, 1]$, $u \in \mathscr{U}^{\varepsilon}_{n,+}$, we have

$$\begin{aligned}\|\mathscr{E}^{\varepsilon}\|_{\infty,t} &\le a(\varepsilon) \int_{\mathscr{X}_t} L_G(y) \|\bar{Y}^{\varepsilon}(s)\| \, |\psi(s, y)| \nu(dy) ds \\ &\le a(\varepsilon) \|\bar{Y}^{\varepsilon}\|_{\infty,t} \int_{\mathscr{X}_t} L_G(y) |\psi(s, y)| \nu(dy) ds \\ &\le c_2 a(\varepsilon) \|\bar{Y}^{\varepsilon}\|_{\infty,t}, \end{aligned} \tag{10.38}$$

where the last inequality follows from the second bound in Lemma 10.18. Again using the Lipschitz condition on G, we have for all $t \in [0, 1]$ that

$$\|B^{\varepsilon}\|_{\infty,t} \le \|L_G\|_1 \int_0^t \|\bar{Y}^{\varepsilon}(s)\| ds.$$

Similarly, the Lipschitz condition on b gives

$$\|A^{\varepsilon}\|_{\infty,t} \le L_b \int_0^t \|\bar{Y}^{\varepsilon}(s)\| ds.$$

Finally, we come to the term C^{ε}. Again using the second bound in Lemma 10.18, we have that for some $c_3 \in (0, \infty)$ and all $u \in \mathscr{U}^{\varepsilon}_{n,+}$, with $u = (\zeta, \varphi)$, $\psi = (\varphi - 1)/a(\varepsilon)$,

$$\left\| \int_{\mathscr{X}_\cdot} G(X^0(s), y) \psi(s, y) \nu(dy) ds \right\|_{\infty,1} \le c_3.$$

Since $(\zeta, \varphi) \in \mathscr{U}^{\varepsilon}_{n,+}$, the bound $\int_0^1 \|\zeta(s)\|^2 ds \le n a^2(\varepsilon)$ applies. Thus for $t \in [0, 1]$,

$$\frac{1}{a(\varepsilon)} \int_0^t \sigma(\bar{X}^{\varepsilon}(s)) \zeta(s) ds = \frac{1}{a(\varepsilon)} \int_0^t \sigma(X^0(s)) \zeta(s) ds + \frac{1}{a(\varepsilon)} \mathscr{R}^{\varepsilon}_1(t),$$

where

$$\|\mathscr{R}^{\varepsilon}_1\|_{\infty,1} \le a(\varepsilon) L_\sigma \|\bar{Y}^{\varepsilon}\|_{\infty,1} \sqrt{n}, \tag{10.39}$$

and

$$\left\| \frac{1}{a(\varepsilon)} \int_0^t \sigma(X^0(s)) \zeta(s) ds \right\|_{\infty,1} \le (\|X^0\|_{\infty,1} L_\sigma + \|\sigma(0)\|) \sqrt{n}.$$

Bringing terms in (10.36) of the form $c a(\varepsilon) \|\bar{Y}^{\varepsilon}\|_{\infty,1}$ to the left side and renormalizing for a coefficient of one, we have, for some $c_4 \in (0, \infty)$, $\tilde{\varepsilon}_0 \in (0, \varepsilon_0)$, and all $u \in \mathscr{U}^{\varepsilon}_{n,+}$, $t \in [0, 1]$, and $\varepsilon \le \tilde{\varepsilon}_0$, that

$$\|\bar{Y}^{\varepsilon}\|_{\infty,t} \le c_4 \left(1 + \int_0^t \|\bar{Y}^{\varepsilon}\|_{\infty,s} ds \right) + Z^{\varepsilon},$$

where $\{Z^\varepsilon\}_{\varepsilon \le \tilde{\varepsilon}_0}$ is tight. The result now follows by an application of Gronwall's inequality. □

The next two lemmas will be needed in the weak convergence arguments of Lemma 10.24. The first is an immediate consequence of Lemma 10.18.

Lemma 10.22 *Let $h \in \mathscr{L}^1(\nu) \cap \mathscr{L}^\rho_{exp}$ for some $\rho > 0$. Then for all $n \in \mathbb{N}$, and $\beta \in (0, \infty)$,*

$$\sup_{f \in S_n^{N,\varepsilon}} \int_{\mathscr{X}_1} h(y)\, |f(s,y)| 1_{\{|f| \ge \beta/a(\varepsilon)\}} \nu(dy)ds \to 0 \quad \text{as } \varepsilon \to 0.$$

For $n \in (0, \infty)$, let $\hat{S}_n^N \doteq \{f \in \mathscr{L}^2(\nu_1) : \|f\|^2_{N,2} \le n\}$.

Lemma 10.23 *Let $n \in \mathbb{N}$, $\varepsilon > 0$, and $f^\varepsilon \in S_n^{N,\varepsilon}$. Let $\eta : \mathscr{X}_1 \to \mathbb{R}^d$ be a measurable function such that*

$$|\eta(s,y)| \le h(y) \text{ for } y \in \mathscr{X}, s \in [0,1],$$

where $h \in \mathscr{L}^1(\nu) \cap \mathscr{L}^\rho_{exp}$ for some $\rho > 0$. Suppose there is $\beta \in (0,1]$ such that $f^\varepsilon 1_{\{|f^\varepsilon| \le \beta/a(\varepsilon)\}}$ converges in $\hat{S}^N_{n\kappa_2(1)}$ to f. Then for all $t \in [0,1]$,

$$\int_{\mathscr{X}_t} \eta(s,y) f^\varepsilon(s,y) \nu_1(ds \times dy) \to \int_{\mathscr{X}_t} \eta(s,y) f(s,y) \nu_1(ds \times dy).$$

Proof It follows from Lemma 10.22 that

$$\int_{\mathscr{X}_1} |\eta(s,y) f^\varepsilon(s,y)| 1_{\{|f^\varepsilon| \ge \beta/a(\varepsilon)\}} \nu_1(ds \times dy) \to 0 \text{ as } \varepsilon \to 0.$$

Also, since $\eta 1_{[0,t]} \in \mathscr{L}^2(\nu_1)$ for all $t \in [0,1]$ and $f^\varepsilon 1_{\{|f^\varepsilon| \le \beta/a(\varepsilon)\}} \to f$, we have

$$\int_{\mathscr{X}_t} \eta(s,y) f^\varepsilon(s,y) 1_{\{|f^\varepsilon| \le \beta/a(\varepsilon)\}} \nu_1(ds \times dy) \to \int_{\mathscr{X}_t} \eta(s,y) f(s,y) \nu_1(ds \times dy).$$

The result follows on combining the last two displays. □

10.3.2 Proof of the Moderate Deviation Principle

The following is the key result needed in the proof of Theorem 10.14. It gives tightness of the joint distribution of controls and controlled processes, and indicates how limits of these two quantities are related. Recall $\hat{S}_n^N \doteq \{f \in \mathscr{L}^2(\nu_1) : \|f\|^2_{N,2} \le n\}$.

Lemma 10.24 *Suppose Conditions 10.1 and 10.13 hold. Let $n \in \mathbb{N}$, $\varepsilon > 0$, and $u^\varepsilon = (\zeta^\varepsilon, \varphi^\varepsilon) \in \mathscr{U}^\varepsilon_{n,+}$. Let $\psi^\varepsilon \doteq (\varphi^\varepsilon - 1)/a(\varepsilon)$, suppose $\beta \in (0, 1]$, and set $\bar{Y}^\varepsilon \doteq \mathscr{K}^\varepsilon(\sqrt{\varepsilon} W^{\zeta^\varepsilon/\sqrt{\varepsilon}}, \varepsilon N^{\varphi^\varepsilon/\varepsilon})$. Then $\{(\bar{Y}^\varepsilon, \psi^\varepsilon 1_{\{|\psi^\varepsilon| \le \beta/a(\varepsilon)\}}, \zeta^\varepsilon/a(\varepsilon))\}_{\varepsilon>0}$ is tight in*

$$\mathscr{D}([0,1] : \mathbb{R}^d) \times \hat{S}_{n(\kappa_2(1)+1)},$$

and any limit point $(\bar{Y}, \psi, \zeta)$ satisfies (10.25) a.s., with η replaced by $\bar{Y}$ and (f_1, f_2) replaced by (ψ, ζ).

Proof We use the notation from the proof of Lemma 10.21 but replace (ζ, φ) throughout by $(\zeta^\varepsilon, \varphi^\varepsilon)$. Assume without loss of generality that $\varepsilon \le \varepsilon_0$. From (10.37) we have that $\|M^\varepsilon\|_{\infty,1} \to 0$ in probability as $\varepsilon \to 0$. Also, since from Lemma 10.21 $\{\|\bar{Y}^\varepsilon\|_{\infty,1}\}_{\varepsilon \le \varepsilon_0}$ is tight, (10.38) implies that $\|\mathscr{E}^\varepsilon\|_{\infty,1} \to 0$ in probability.

Noting that $\bar{X}^\varepsilon(t) = X^0(t) + a(\varepsilon)\bar{Y}^\varepsilon(t)$, we have by Taylor's formula,

$$G(\bar{X}^\varepsilon(s), y) - G(X^0(s), y) = a(\varepsilon) D_x G(X^0(s), y)\bar{Y}^\varepsilon(t) + R^\varepsilon(s, y),$$

where

$$\|R^\varepsilon(s, y)\| \le L_{DG}(y) a^2(\varepsilon) \|\bar{Y}^\varepsilon(s)\|^2.$$

Hence

$$B^\varepsilon(t) = \int_{\mathscr{X}_t} D_x G(X^0(s), y)\bar{Y}^\varepsilon(s)\nu(dy)\, ds + T_1^\varepsilon(t),$$

where

$$\|T_1^\varepsilon\|_{\infty,1} \le \|L_{DG}\|_1 a(\varepsilon) \int_0^1 \|\bar{Y}^\varepsilon(s)\|^2 ds.$$

Thus using Lemma 10.21 again, we have that $\|T_1^\varepsilon\|_{\infty,1} \to 0$ in probability. Similarly,

$$A^\varepsilon(t) = \int_0^t Db(X^0(s))\bar{Y}^\varepsilon(s) ds + T_2^\varepsilon(t),$$

where $\|T_2^\varepsilon\|_{\infty,1} \to 0$ in probability. Also, from (10.39) $\|\mathscr{R}_1^\varepsilon\|_{\infty,1} \to 0$ in probability. Putting these estimates together, we have from (10.36) that

$$\begin{aligned} \bar{Y}^\varepsilon(t) = T_3^\varepsilon(t) + \int_0^t Db(X^0(s))\bar{Y}^\varepsilon(s) ds + \frac{1}{a(\varepsilon)} \int_0^t \sigma(X^0(s))\zeta^\varepsilon(s) ds \qquad (10.40) \\ + \int_{\mathscr{X}_t} D_x G(X^0(s), y)\bar{Y}^\varepsilon(s)\nu(dy) ds + \int_{\mathscr{X}_t} G(X^0(s), y)\psi^\varepsilon(s, y)\nu(dy) ds, \end{aligned}$$

where

$$T_3^\varepsilon \doteq M^\varepsilon + \mathscr{E}^\varepsilon + T_1^\varepsilon + T_2^\varepsilon + \mathscr{R}_1^\varepsilon \Rightarrow 0.$$

We now prove tightness of

$$\tilde{A}^\varepsilon(\cdot) \doteq \int_0^\cdot Db(X^0(s))\bar{Y}^\varepsilon(s)ds,\ \ \tilde{B}^\varepsilon(\cdot) \doteq \int_{\mathscr{X}\times[0,\cdot]} D_x G(X^0(s), y)\bar{Y}^\varepsilon(s)\nu(dy)ds,$$
$$\tilde{C}^\varepsilon(\cdot) \doteq \int_{\mathscr{X}\times[0,\cdot]} G(X^0(s), y)\psi^\varepsilon(s, y)\nu(dy)ds,\ \ \tilde{D}^\varepsilon(\cdot) \doteq \frac{1}{a(\varepsilon)}\int_0^\cdot \sigma(X^0(s))\zeta^\varepsilon(s)ds.$$

Applying Lemma 10.18 with $h = M_G$, for every $\beta \in (0, 1]$ and $\delta \in (0, 1)$, we have

$$\begin{aligned}\|\tilde{C}^\varepsilon(t+\delta) - \tilde{C}^\varepsilon(t)\| &= \int_{[t,t+\delta]\times\mathscr{X}} \|G(X^0(s), y)\|\,|\psi^\varepsilon(s, y)|\nu(dy)ds \\ &\le \left(1 + \|X^0\|_{\infty,1}\right)\int_{[t,t+\delta]\times\mathscr{X}} M_G(y)|\psi^\varepsilon(s, y)|\nu(dy)ds \\ &\le \left(1 + \|X^0\|_{\infty,1}\right)(\zeta(\beta)\delta^{1/2} + \sqrt{a(\varepsilon)}\vartheta(\beta) + 2\xi(\varepsilon)).\end{aligned}$$

Tightness of $\{\tilde{C}^\varepsilon\}_{\varepsilon>0}$ in $\mathscr{C}([0, 1] : \mathbb{R}^d)$ is now immediate from the properties of ϑ and ξ.

Next we argue the tightness of $\tilde{B}^\varepsilon$. For $0 \le t \le t + \delta \le 1$, we have

$$\begin{aligned}\|\tilde{B}^\varepsilon(t+\delta) - \tilde{B}^\varepsilon(t)\| &\le \int_{[t,t+\delta]\times\mathscr{X}} \|D_x G(X^0(s), y)\bar{Y}^\varepsilon(s)\|\nu(dy)ds \\ &\le \left(\sup_{\|x\|\le\|X^0\|_{\infty,1}} \int_{\mathscr{X}} \|D_x G(x, y)\|\nu(dy)\right)\int_{[t,t+\delta]} \|\bar{Y}^\varepsilon(s)\|ds \\ &\le c_1\|\bar{Y}^\varepsilon\|_{\infty,1}\delta,\end{aligned}$$

where $c_1 \doteq \sup_{\|x\|\le\|X^0\|_{\infty,1}} \int_{\mathscr{X}} \|D_x G(x, y)\|\nu(dy)$ is finite by part (b) of Condition 10.13. Tightness of $\{\tilde{B}^\varepsilon\}_{\varepsilon>0}$ in $\mathscr{C}([0, 1] : \mathbb{R}^d)$ now follows as a consequence of Lemma 10.21. Similarly, it can be seen that $\tilde{A}^\varepsilon$ is tight in $\mathscr{C}([0, 1] : \mathbb{R}^d)$. Finally, since $\zeta^\varepsilon \in S^W_{na^2(\varepsilon)}$ implies the bound $na^2(\varepsilon)$ for $\int_0^1 \|\zeta^\varepsilon\|^2 ds$, it follows that for $0 \le t \le t + \delta \le 1$, we have

$$\begin{aligned}\|\tilde{D}^\varepsilon(t+\delta) - \tilde{D}^\varepsilon(t)\| &\le \frac{1}{a(\varepsilon)}\int_t^{t+\delta} \|\sigma(X^0(s))\zeta^\varepsilon(s)\|ds \\ &\le \sqrt{\delta}(\|X^0\|_{\infty,1}L_\sigma + \|\sigma(0)\|)\sqrt{n}.\end{aligned}$$

Tightness of $\{\tilde{D}^\varepsilon\}_{\varepsilon>0}$ in $\mathscr{C}([0, 1] : \mathbb{R}^d)$ is now immediate. Since each of these terms is tight, $\{\bar{Y}^\varepsilon\}_{\varepsilon>0}$ is tight in $\mathscr{D}([0, 1] : \mathbb{R}^d)$. Also, from part (c) of Lemma 9.7,

$$\left(\psi^\varepsilon 1_{\{|\psi^\varepsilon|\le\beta/a(\varepsilon)\}}, \frac{1}{a(\varepsilon)}\zeta^\varepsilon\right)$$

takes values in the compact space $\hat{S}_{n(\kappa_2(1)+1)}$ for all $\varepsilon > 0$ and is therefore automatically tight. This completes the proof of the first part of the lemma.

Suppose now that

$$\left(\bar{Y}^\varepsilon, \psi^\varepsilon 1_{\{|\psi^\varepsilon| \le \beta/a(\varepsilon)\}}, \frac{1}{a(\varepsilon)}\zeta^\varepsilon\right)$$

converges in distribution along a subsequence to $(\bar{Y}, \psi, \zeta)$. From Lemma 10.23 and the tightness of $\tilde{C}^\varepsilon$, $\tilde{D}^\varepsilon$ established above, we have that

$$\left(\bar{Y}^\varepsilon, \int_{\mathscr{X}_\cdot} G(X^0(s), y)\psi^\varepsilon(s, y)\nu(dy)ds, \frac{1}{a(\varepsilon)}\int_0^\cdot \sigma(X^0(s))\zeta^\varepsilon(s)ds\right)$$

converges in distribution, in $\mathscr{D}([0, 1] : \mathbb{R}^{3d})$, to

$$\left(\bar{Y}, \int_{\mathscr{X}_\cdot} G(X^0(s), y)\psi(s, y)\nu(dy)ds, \int_0^\cdot \sigma(X^0(s))\zeta(s)ds\right).$$

The result now follows on using this convergence in (10.40) and recalling that $T_3^\varepsilon \Rightarrow 0$. □

We now complete the proof of the moderate deviation principle.

Proof (*of Theorem* 10.14) It suffices to show that Condition 9.8 holds with $\mathscr{K}^\varepsilon$ and $\mathscr{K}^0$ defined as at the beginning of Sect. 10.3.1. Part (a) of the condition was verified in Lemma 10.17. Consider now part (b). Fix $n \in \mathbb{N}$ and $\beta \in (0, 1]$. Let $(\zeta^\varepsilon, \varphi^\varepsilon) \in \mathscr{U}^\varepsilon_{n,+}$ and $\psi^\varepsilon = (\varphi^\varepsilon - 1)/a(\varepsilon)$. Suppose that for some $\beta \in (0, 1]$, $(\psi^\varepsilon 1_{\{|\psi^\varepsilon| \le \beta/a(\varepsilon)\}}, \zeta^\varepsilon/a(\varepsilon)) \Rightarrow (\psi, \zeta)$. To complete the proof, we need to show that

$$\mathscr{K}^\varepsilon\left(\sqrt{\varepsilon}W^{\zeta^\varepsilon/\sqrt{\varepsilon}}, \ \varepsilon N^{\varphi^\varepsilon/\varepsilon}\right) \Rightarrow \mathscr{K}^0(\zeta, \psi). \tag{10.41}$$

Recall that the left side of (10.41) equals $\bar{Y}^\varepsilon$ defined in (10.33). From Lemma 10.24, $\{(\bar{Y}^\varepsilon, \psi^\varepsilon 1_{\{|\psi^\varepsilon| \le \beta/a(\varepsilon)\}}, \zeta^\varepsilon/a(\varepsilon))\}$ is tight in $\mathscr{D}([0, 1] : \mathbb{R}^d) \times \hat{S}_{n(\kappa_2(1)+1)}$, and every limit point $(\bar{Y}, \psi, \zeta)$ satisfies (10.25) a.s., with η replaced by $\bar{Y}$ and (f_1, f_2) with $(\bar{\psi}, \bar{\zeta})$. Since (10.25) has a unique solution, $\mathscr{K}^0(f_1, f_2)$ for every $f = (f_1, f_2) \in \mathscr{L}^2$ [recall $\mathscr{L}^2 \doteq \mathscr{L}^2([0, 1] : \mathbb{R}^d) \times \mathscr{L}^2(\nu_T)$], (ζ, ψ) has the same law as $(\bar{\zeta}, \bar{\psi})$, and every limit point of $\bar{Y}^\varepsilon$ must have the same distribution as $\mathscr{K}^0(\zeta, \psi)$. The result follows. □

10.3.3 Equivalence of Two Rate Functions

In this section we present the proof of Theorem 10.15. To simplify notation, suppose without loss that $T = 1$. Fix $\eta \in \mathscr{C}([0, 1] : \mathbb{R}^d)$ and $\delta > 0$. Let $\tilde{f} = (\tilde{f}_1, \tilde{f}_2)$, where $\tilde{f}_i \in \mathscr{L}^2([0, 1] : \mathbb{R}^d)$, $i = 1, 2$, be such that

$$\frac{1}{2}\int_0^1 \left(\|\tilde{f}_1(s)\|^2 + \|\tilde{f}_2(s)\|^2\right) ds \le I(\eta) + \delta$$

and $(\eta, \tilde{f})$ satisfy (10.27). Define $f_2 : \mathscr{X}_1 \to \mathbb{R}$ by

$$f_2(s, y) \doteq \sum_{i=1}^{d} \tilde{f}_{2,i}(s) e_i(s, y), \ \ (s, y) \in \mathscr{X}_1, \tag{10.42}$$

where the e_i were introduced just before the statement of Theorem 10.15. From the orthonormality of $e_i(s, \cdot)$, it follows that

$$\frac{1}{2}\int_{\mathscr{X}_1} |f_2(s, y)|^2 \nu_1(ds \times dy) = \frac{1}{2}\int_0^1 \|\tilde{f}_2(s)\|^2 ds. \tag{10.43}$$

Also, with $A(s)$ defined as in (10.26), we have

$$\begin{aligned}
[A(s)\tilde{f}_2(s)]_i &= \sum_{j=1}^{d} \langle G_i(X^0(s), \cdot), e_j(s, \cdot)\rangle_{\mathscr{L}^2(\nu)} \tilde{f}_{2,j}(s) \\
&= \left\langle G_i(X^0(s), \cdot), \sum_{j=1}^{d} e_j(s, \cdot)\tilde{f}_{2,j}(s) \right\rangle_{\mathscr{L}^2(\nu)} \\
&= \langle G_i(X^0(s), \cdot), f_2(s, \cdot)\rangle_{\mathscr{L}^2(\nu)},
\end{aligned}$$

so that $A(s)\tilde{f}_2(s) = \int_{\mathscr{X}} f_2(s, y) G(X^0(s), y)\nu(dy)$. Consequently, η satisfies (10.25) with f_2 as in (10.42) and $f_1 = \tilde{f}_1$. Combining this with (10.43), we have

$$\begin{aligned}
\bar{I}(\eta) &\le \frac{1}{2}\int_0^1 \|\tilde{f}_1(s)\|^2 ds + \frac{1}{2}\int_{\mathscr{X}_1} |f_2(s, y)|^2 \nu_2(ds \times dy) \\
&= \frac{1}{2}\int_0^1 \left(\|\tilde{f}_1(s)\|^2 + \|\tilde{f}_2(s)\|^2\right) ds \\
&\le I(\eta) + \delta.
\end{aligned}$$

Since $\delta > 0$ is arbitrary, we have $\bar{I}(\eta) \le I(\eta)$.

Conversely, suppose $\delta > 0$ and $q = (f_1, f_2) \in \mathscr{L}^2$ is such that

$$\frac{1}{2}\int_{\mathscr{X}_1} |f_2(s, y)|^2 \nu_1(ds \times dy) + \frac{1}{2}\int_0^1 \|f_1(s)\|^2 ds \le \bar{I}(\eta) + \delta$$

and (10.25) holds. For $i = 1, \ldots, d$, define $\tilde{f}_{2,i} : [0, 1] \to \mathbb{R}$ by

$$\tilde{f}_{2,i}(s) = \langle f_2(s, \cdot), e_i(s, \cdot)\rangle_{\mathscr{L}^2(\nu)}.$$

For $s \in [0,1]$, let $\{e_j(s,\cdot)\}_{j=d+1}^{\infty}$ be defined in such a manner that $\{e_j(s,\cdot)\}_{j=1}^{\infty}$ is a complete orthonormal system in $\mathscr{L}^2(\nu)$. Then for every $s \in [0,1]$, $i = 1, \ldots, d$,

$$\begin{aligned}
[A(s)\tilde{f}_2(s)]_i &= \sum_{j=1}^{d} \langle G_i(X^0(s),\cdot), e_j(s,\cdot)\rangle_{\mathscr{L}^2(\nu)} \langle f_2(s,\cdot), e_j(\cdot,s)\rangle_{\mathscr{L}^2(\nu)} \\
&= \sum_{j=1}^{\infty} \langle G_i(X^0(s),\cdot), e_j(\cdot,s)\rangle_{\mathscr{L}^2(\nu)} \langle f_2(s,\cdot), e_j(s,\cdot)\rangle_{\mathscr{L}^2(\nu)} \\
&= \langle G_i(X^0(s),\cdot), f_2(s,\cdot)\rangle_{\mathscr{L}^2(\nu)},
\end{aligned}$$

where the second equality follows on observing that $G_i(X^0(s),\cdot)$ is in the linear span of $\{e_j(s,\cdot)\}_{j=1}^{d}$ for $i = 1, \ldots, d$. Thus $A(s)\tilde{f}_2(s) = \int_{\mathscr{X}} f_2(s,y)G(X^0(s),y)\nu(dy)$, and therefore $(\eta, \tilde{f})$ satisfy (10.27) with $\tilde{f}_2$ defined as above and $\tilde{f}_1 = f_1$. Note that with $\tilde{f}_2 = (\tilde{f}_{2,1}, \ldots, \tilde{f}_{2,d})$,

$$\begin{aligned}
\frac{1}{2}\int_0^1 \|\tilde{f}_2(s)\|^2 ds &= \frac{1}{2}\int_0^1 \sum_{j=1}^{d} \langle f_2(s,\cdot), e_j(s,\cdot)\rangle^2_{\mathscr{L}^2(\nu)} ds \\
&\le \frac{1}{2}\int_0^1 \int_{\mathscr{X}} f_2^2(s,y)\nu(dy)ds.
\end{aligned}$$

Thus

$$\begin{aligned}
I(\eta) &\le \frac{1}{2}\int_0^1 \left(\|\tilde{f}_1(s)\|^2 + \|\tilde{f}_2(s)\|^2\right) ds \\
&\le \frac{1}{2}\int_0^1 \|f_1(s)\|^2 ds + \frac{1}{2}\int_0^1 \int_{\mathscr{X}} f_2^2(s,y)\nu(dy)ds \\
&\le \bar{I}(\eta) + \delta.
\end{aligned}$$

Since $\delta > 0$ is arbitrary, $I(\eta) \le \bar{I}(\eta)$, which completes the proof. □

10.4 Notes

The first results for a general class of continuous time small noise Markov processes appear to be those of Wentzell [245–248]. The class covered includes both Gaussian and Poisson driving noises, and the proof uses approximation by discrete time processes.

Large deviation principles for small noise infinite dimensional stochastic differential equations driven by PRM are considered in [38]. Although this paper considers a considerably more complex setting than the one studied in the current chapter, it

assumed a somewhat stronger condition on the coefficient function that plays the role of G in this chapter. Specifically, it required the functions M_G and L_G to satisfy a more stringent integrability condition than the one used here (Condition 10.3). Some of the results used to deal with these weaker integrability conditions are from [47]. Another distinction between this chapter and [38] is that here we consider systems driven by both Gaussian and Poisson noise, and consider both large and moderate deviations. A moderate deviation principle of the form given in Sect. 10.3, applicable to both finite and infinite dimensional SDE, was presented in [41] for the setting in which the driving noise is only Poisson. It is worth noting that most of the technical details in this chapter arise from the treatment of the PRM term.

There is an important distinction between the types of processes one can represent using PRMs and their large deviation analysis. For one class, which includes the models considered in this chapter as well as the example in Sect. 3.3, different points in the point space of the underlying PRM are used to model different types of jumps in the solution to the SDE. The conditions placed on the coefficient that modulates the impact of the noise on the state (G in this chapter) tend to be continuity-type conditions, analogous to those one places on the diffusion coefficient of an SDE driven by Brownian motion, though stated in terms of integration over the space $\mathscr{X}$. These continuity properties are used to establish uniqueness of the map that takes the controls into the state for the limit dynamics, under the assumption that the cost of the controls is bounded. With the second class, there are only finitely many different types of jumps of the state, and the role of G is simply to "thin" the PRM to produce state-dependent jump rates. Examples of this type include the process models of Chap. 13 as well as those in [23, 42]. Owing to its role in thinning, G is typically not continuous, and one does not expect uniqueness of the limiting deterministic map that takes controls to the state process. However, as we will see, it is in fact sufficient to prove the following restricted uniqueness: given a state trajectory for which there is a corresponding control with finite cost, find a control that produces the same state trajectory with nearly the same cost and for which there is uniqueness. These points are illustrated in the analysis carried out in Chap. 13.

Chapter 11
Systems Driven by an Infinite Dimensional Brownian Noise

In Chap. 8 we gave a representation for positive functionals of a Hilbert space valued Brownian motion. This chapter will apply the representation to study the large deviation properties of infinite dimensional small noise stochastic dynamical systems. In the application, the driving noise is given by a Brownian sheet, and so in this chapter we will present a sufficient condition analogous to Condition 9.1 (but there will be no Poisson noise in this chapter) that covers the setting of such noise processes (see Condition 11.15). Another formulation of an infinite dimensional Brownian motion that will be needed in Chap. 12 is as a sequence of independent Brownian motions regarded as a $\mathscr{C}([0, T] : \mathbb{R}^{\infty})$-valued random variable. We also present the analogous sufficient condition (Condition 11.12) for an LDP to hold for this type of driving noise.

To illustrate the approach we consider a class of reaction–diffusion stochastic partial differential equations (SPDE), for which well-posedness has been studied in [174]. Previous works that prove an LDP for this SPDE include [170, 235]. The proof of the Laplace principle proceeds by verification of Condition 11.15. Just as in Chap. 10, the key ingredients in the verification of this condition are the well-posedness and compactness for sequences of controlled versions of the original SPDE [Theorems 11.23, 11.24, and 11.25]. Also as in Chap. 10, the techniques and estimates used to prove such properties for the original (uncontrolled) stochastic model can be applied here as well, and indeed proofs for the controlled SPDEs proceed in very much the same way as those of their uncontrolled counterparts.

The chapter is organized as follows. In Sect. 11.1 we recall some common formulations of an infinite dimensional Brownian motion and relations between them. Starting from the variational representation for a Hilbert space valued Brownian motion from Chap. 8, we present analogous representations for these equivalent formulations of infinite dimensional Brownian motion. Then starting from the sufficient condition for Hilbert space valued Brownian motion given in Chap. 9, we state the corresponding sufficient conditions for a uniform Laplace principle to hold for these

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_11

other formulations in Sect. 11.2. The illustration of how the conditions are verified is given in Sect. 11.3, which studies the large deviation properties for a family of stochastic reaction–diffusion equations.

11.1 Formulations of Infinite Dimensional Brownian Motion

An infinite dimensional Brownian motion arises in a natural fashion in the study of stochastic processes with a spatial parameter. We refer the reader to [69, 169, 243] for numerous examples in the physical sciences in which infinite dimensional Brownian motions are used to model the driving noise for stochastic dynamical systems. Depending on the application of interest, the infinite dimensional nature of the driving noise may be expressed in a variety of forms. Some examples include Hilbert space valued Brownian motion (as was considered in Chap. 8); cylindrical Brownian motion; an infinite sequence of iid standard (1-dimensional) Brownian motions; and space-time Brownian sheets. In what follows, we describe these various formulations and explain how they relate to each other. We will be concerned only with processes defined over a fixed time horizon, and thus fix $T \in (0, \infty)$, and all filtrations and stochastic processes will be defined over the horizon $[0, T]$. Reference to T will be omitted unless essential. Let $(\Omega, \mathscr{F}, P)$ be a probability space with a filtration $\{\mathscr{F}_t\}_{0 \le t \le T}$ satisfying the usual conditions. Let $(\mathscr{H}, \langle \cdot, \cdot \rangle)$ be a real separable Hilbert space. Let Λ be a symmetric strictly positive trace class operator on $\mathscr{H}$. Recall that an $\mathscr{H}$-valued continuous stochastic process $\{W(t)\}_{0 \le t \le T}$ defined on $(\Omega, \mathscr{F}, P)$ is called a Λ-Wiener process with respect to $\{\mathscr{F}_t\}$ if for every nonzero $h \in \mathscr{H}$, $\{\langle \Lambda h, h \rangle^{-1/2} \langle W(t), h \rangle\}$ is a one-dimensional standard $\{\mathscr{F}_t\}$-Wiener process.

Another formulation for an infinite dimensional Brownian motion, which will be used in Chap. 12 for the study of stochastic flows of diffeomorphisms, is as follows. Let $\{\beta_i\}_{i \in \mathbb{N}}$ be an infinite sequence of independent standard one-dimensional, $\{\mathscr{F}_t\}$-Brownian motions. We denote the product space of countably infinite copies of the real line by $\mathbb{R}^\infty$. Note that a sequence of independent standard Brownian motions $\{\beta_i\}_{i \in \mathbb{N}}$ can be regarded as a random variable with values in $\mathscr{C}([0, T] : \mathbb{R}^\infty)$, where $\mathbb{R}^\infty$ is equipped with the usual topology of coordinatewise convergence, which can be metrized using the distance

$$d(u, v) \doteq \sum_{k=1}^{\infty} \frac{|u_k - v_k| \wedge 1}{2^k}.$$

It is easily checked that with this metric, $\mathbb{R}^\infty$ is a Polish space. Thus $\beta = \{\beta_i\}_{i \in \mathbb{N}}$ is a random variable with values in the Polish space $\mathscr{C}([0, T] : \mathbb{R}^\infty)$, and can be regarded as another model of an infinite dimensional Brownian motion.

Let $\{e_i\}_{i \in \mathbb{N}}$ be a complete orthonormal system (CONS) for the Hilbert space $\mathscr{H}$ such that $\Lambda e_i = \lambda_i e_i$, where λ_i is the strictly positive ith eigenvalue of Λ, which

corresponds to the eigenvector e_i. Since Λ is a trace class operator, $\sum_{i=1}^{\infty} \lambda_i < \infty$. Define

$$\beta_i(t) \doteq \frac{1}{\sqrt{\lambda_i}} \langle W(t), e_i \rangle, \ 0 \le t \le T, \ i \in \mathbb{N},$$

where W as before is a Λ-Wiener process with respect to $\{\mathscr{F}_t\}$. It is easy to check that $\{\beta_i\}$ is a sequence of independent standard $\{\mathscr{F}_t\}$-Brownian motions. Thus starting from a Λ-Wiener process, one can produce an infinite collection of independent standard Brownian motions in a straightforward manner. Conversely, given a collection of independent standard Brownian motions $\{\beta_i\}_{i\in\mathbb{N}}$ and $(\Lambda, \{e_i, \lambda_i\})$ as above, one can obtain a Λ-Wiener process W by setting

$$W(t) \doteq \sum_{i=1}^{\infty} \sqrt{\lambda_i} \beta_i(t) e_i, \ 0 \le t \le T. \tag{11.1}$$

The right side of (11.1) clearly converges in $\mathscr{L}^2(P)$ for each fixed t. Furthermore, one can check that the series also converges in $\mathscr{C}([0,T] : \mathscr{H})$ almost surely [69, Theorem 4.3]. These observations lead to the following result.

Proposition 11.1 *There exist measurable maps* $f : \mathscr{C}([0,T] : \mathbb{R}^\infty) \to \mathscr{C}([0,T] : \mathscr{H})$ *and* $g : \mathscr{C}([0,T] : \mathscr{H}) \to \mathscr{C}([0,T] : \mathbb{R}^\infty)$ *such that* $f(\beta) = W$ *and* $g(W) = \beta$ *a.s.*

Remark 11.2 Consider the Hilbert space $l_2 \doteq \{x = (x_1, x_2, \ldots) : x_i \in \mathbb{R}$ and $\sum_{i=1}^{\infty} x_i^2 < \infty\}$ with the inner product $\langle x, y \rangle_0 \doteq \sum_{i=1}^{\infty} x_i y_i$. Let $\{\lambda_i\}_{i\in\mathbb{N}}$ be a sequence of strictly positive numbers such that $\sum_{i=1}^{\infty} \lambda_i < \infty$. Then the Hilbert space $\bar{l}_2 \doteq \{x = (x_1, x_2, \ldots) : x_i \in \mathbb{R}$ and $\sum_{i=1}^{\infty} \lambda_i x_i^2 < \infty\}$ with the inner product $\langle x, y \rangle \doteq \sum_{i=1}^{\infty} \lambda_i x_i y_i$ contains l_2, and the embedding map $\iota : l_2 \to \bar{l}_2, \iota(x) = x$ is Hilbert–Schmidt. Furthermore, the infinite sequence of real Brownian motions β takes values in $\bar{l}_2$ almost surely and can be regarded as a $\bar{l}_2$-valued Λ-Wiener process, where Λ is defined by $\langle \Lambda x, y \rangle = \sum_{i=1}^{\infty} \lambda_i^2 x_i y_i$, $x, y \in \bar{l}_2$.

Equation (11.1) above can be interpreted as saying that the sequence $\{\lambda_i\}$ (or equivalently the trace class operator Λ) injects a "coloring" to a white noise such that the resulting process has greater regularity. In some models of interest, such coloring is obtained indirectly in terms of (state-dependent) diffusion coefficients. It is natural in such situations to consider the driving noise a "cylindrical Brownian motion" rather than a Hilbert space valued Brownian motion. Let $(\mathscr{H}, \langle \cdot, \cdot \rangle)$ as before be a real separable Hilbert space and fix a probability space and a filtration as above.

Definition 11.3 A family $\{B_t(h) = B(t,h) : t \in [0,T], h \in \mathscr{H}\}$ of real random variables is said to be an $\{\mathscr{F}_t\}$**-cylindrical Brownian motion** if the following hold.

(a) For every $h \in \mathscr{H}$ with $\|h\| = 1$, $\{B(t,h)\}$ is a standard $\mathscr{F}_t$-Wiener process.

(b) For every $t \in [0,T]$, $a_1, a_2 \in \mathbb{R}$ and $f_1, f_2 \in \mathscr{H}$,

$$B(t, a_1 f_1 + a_2 f_2) = a_1 B(t, f_1) + a_2 B(t, f_2) \ \text{ a.s.}$$

If $\{B_t(h)\}$ is a cylindrical Brownian motion and $\{e_i\}$ is a CONS in $\mathscr{H}$, setting $\beta_i(t) \doteq B(t, e_i)$, we see that $\{\beta_i\}$ is a sequence of independent standard one-dimensional $\{\mathscr{F}_t\}$-Brownian motions. Conversely, given a sequence $\{\beta_i\}_{i\in\mathbb{N}}$ of independent standard one-dimensional $\{\mathscr{F}_t\}$-Brownian motions,

$$B_t(h) \doteq \sum_{i=1}^{\infty} \beta_i(t)\langle e_i, h\rangle \tag{11.2}$$

defines a cylindrical Brownian motion on $\mathscr{H}$. For each $h \in \mathscr{H}$, the series in (11.2) converges in $\mathscr{L}^2(P)$ and a.s. in $\mathscr{C}([0, T] : \mathbb{R})$.

Proposition 11.4 *Let B be a cylindrical Brownian motion as in Definition 11.3 and let β be as constructed in the last paragraph. Then $\sigma\{B_s(h) : 0 \le s \le t, h \in \mathscr{H}\} = \sigma\{\beta(s) : 0 \le s \le t\}$ for all $t \in [0, T]$. In particular, if X is a $\sigma\{B(s, h) : 0 \le s \le T, h \in \mathscr{H}\}$-measurable random variable, then there exists a measurable map $g : \mathscr{C}([0, T] : \mathbb{R}^\infty) \to \mathbb{R}$ such that $g(\beta) = X$ a.s.*

In yet other stochastic dynamical systems, the driving noise is given as a space-time white noise process, also referred to as a Brownian sheet. In what follows, we introduce this stochastic process and describe its relationship with the formulations considered above. Let $(\Omega, \mathscr{F}, P, \{\mathscr{F}_t\})$ be as before and fix a bounded open subset $O \subset \mathbb{R}^d$. We follow standard usage and denote both cylindrical Brownian motions by B [more precisely by $B_t(h)$] and also Brownian sheets by B [in this case $B(t, x)$]. The intended use should be clear from context.

Definition 11.5 A family of real-valued Gaussian random variables

$$\{B(t, x),\ (t, x) \in [0, T] \times O\}$$

is called a **Brownian sheet** if the following hold.

(a) If $(t, x) \in [0, T] \times O$, then $E(B(t, x)) = 0$.

(b) If $0 \le s \le t \le T$, then $\{B(t, x) - B(s, x), x \in O\}$ is independent of $\mathscr{F}_s$.

(c) $\mathrm{Cov}\,(B(t, x), B(s, y)) = \lambda(A_{t,x} \cap A_{s,y})$, where λ is Lebesgue measure on $[0, T] \times O$ and

$$A_{t,x} \doteq \left\{(s, y) \in [0, T] \times O : 0 \le s \le t \ \text{ and } \ y_j \le x_j,\ j = 1, \ldots, d\right\}.$$

(d) The map $(t, u) \mapsto B(t, u)$ from $[0, T] \times O$ to $\mathbb{R}$ is uniformly continuous a.s.

Due to the uniform continuity property of part (d), $B = \{B(t, x),\ (t, x) \in [0, T] \times O\}$ can be regarded as a random variable with values in the Polish space $\mathscr{C}([0, T] \times \bar{O} : \mathbb{R})$, the space of continuous functions from $[0, T] \times \bar{O}$ to $\mathbb{R}$, equipped with the uniform topology.

To introduce stochastic integrals with respect to a Brownian sheet, we need the following definitions and notation, which are largely taken from [169].

Definition 11.6 (*Elementary and simple functions*) A function $f : O \times [0,T] \times \Omega \to \mathbb{R}$ is elementary if there exist $a, b \in [0,T]$, $a \le b$, a bounded $\{\mathscr{F}_a\}$-measurable random variable X, and $A \in \mathscr{B}(O)$ such that

$$f(x,s,\omega) = X(\omega)1_{(a,b]}(s)1_A(x).$$

A finite sum of elementary functions is referred to as a simple function. We denote by $\bar{\mathscr{S}}$ the class of all simple functions.

We now introduce the $\{\mathscr{F}_t\}$-predictable σ-field on $\Omega \times [0,T] \times O$. The definition is analogous to that of a predictable σ-field on $\Omega \times [0,T]$ introduced in Chap. 8 and is denoted by the same symbol.

Definition 11.7 (*Predictable σ-field*) The $\{\mathscr{F}_t\}$-predictable σ-field $\mathscr{P}\mathscr{F}$ on $\Omega \times [0,T] \times O$ is the σ-field generated by $\bar{\mathscr{S}}$. A function $f : \Omega \times [0,T] \times O \to \mathbb{R}$ is called an $\{\mathscr{F}_t\}$-predictable process if it is $\mathscr{P}\mathscr{F}$-measurable.

Remark 11.8 In Chap. 8 we considered a probability space supporting a Hilbert space valued Wiener process and defined the classes of integrands/controls $\mathscr{A}_b$, $\mathscr{A}$, $\bar{\mathscr{A}}_b$, and $\bar{\mathscr{A}}$. The first two are predictable with respect to the filtration generated by the Wiener process and either have a finite $\mathscr{L}^2$ norm a.s. ($\mathscr{A}$) or satisfy a uniform bound on this norm a.s. ($\mathscr{A}_b$), and the last two are analogous, save being $\{\mathscr{F}_t\}$-predictable (see the definitions given after Definition 8.2). In this chapter we will need the analogous processes for a number of alternative formulations of infinite dimensional Brownian motion. With some abuse of notation, we use the same symbols to denote the classes with the analogous predictability and boundedness properties for all these different formulations. The class intended in any circumstance will be clear from the context.

Thus analogous to the class of integrands $\bar{\mathscr{A}}$ introduced in Chap. 8, consider the class of all $\{\mathscr{F}_t\}$-predictable processes f such that $\int_{[0,T]\times O} f^2(s,x)ds\,dx < \infty$ a.s., and denote this class by $\bar{\mathscr{A}}$. Classes $\mathscr{A}_b$, $\mathscr{A}$, and $\bar{\mathscr{A}}_b$ are defined similarly. For all $f \in \bar{\mathscr{A}}$, the stochastic integral $M_t(f) \doteq \int_{[0,t]\times O} f(s,u)B(ds \times du)$, $t \in [0,T]$, is well defined as in Chap. 2 of [243]. Furthermore, for all $f \in \bar{\mathscr{A}}$, $\{M_t(f)\}_{0\le t\le T}$ is a continuous $\{\mathscr{F}_t\}$-local martingale. More properties of the stochastic integral can be found in Appendix D.2.4, and in much greater detail in [243].

Consider the Hilbert space $\mathscr{L}^2(O) \doteq \{f : O \to \mathbb{R} : \int_O f^2(x)dx < \infty\}$ equipped with the usual inner product. Let $\{\phi_i\}_{i\in\mathbb{N}}$ be a CONS in $\mathscr{L}^2(O)$. Then it is easy to verify that $\beta = \{\beta_i\}_{i\in\mathbb{N}}$ defined by $\beta_i(t) \doteq \int_{[0,t]\times O} \phi_i(x)B(ds \times dx)$, $i \in \mathbb{N}$, $t \in [0,T]$ is a sequence of independent standard real Brownian motions. Also, for $(t,x) \in [0,T] \times O$,

$$B(t,x) = \sum_{i=1}^{\infty} \beta_i(t) \int_O \phi_i(y)1_{(-\infty,x]}(y)dy \tag{11.3}$$

(where $(-\infty,x] = \{y : y_i \le x_i \text{ for all } i = 1,\dots,d\}$), and the series in (11.3) converges in $\mathscr{L}^2(P)$ for each (t,x). From these considerations, it follows that

$$\sigma\{B(t,x),\ t\in[0,T],\ x\in O\}=\sigma\{\beta_i(t),\ i\in\mathbb{N},\ t\in[0,T]\}. \tag{11.4}$$

As a consequence of (11.4) and Lemma E.1 in the appendix, we have the following result.

Proposition 11.9 *There exists a measurable map $g:\mathscr{C}([0,T]:\mathbb{R}^\infty)\to\mathscr{C}([0,T]\times\bar{O}:\mathbb{R})$ such that $B=g(\beta)$ a.s., where β is as defined by $\beta_i(t)\doteq\int_{[0,t]\times O}\phi_i(x)\,B(ds\times dx)$.*

11.1.1 The Representations

In Chap. 8 we presented a variational representation for positive functionals of a Hilbert space valued Brownian motion. Using this representation and the results of Sect. 11.1, we can obtain analogous representations for other formulations of an infinite dimensional Brownian motion. Let $(\Omega,\mathscr{F},P,\{\mathscr{F}_t\})$ and $\beta=\{\beta_i\}$ be as in Sect. 11.1. Recall that β is a $\mathscr{C}([0,T]:\mathbb{R}^\infty)$-valued random variable.

Let $\mathscr{P}\mathscr{F}$ be the $\{\mathscr{F}_t\}$-predictable σ-field on $[0,T]\times\Omega$ as introduced in Definition 8.2. For a Hilbert space $\mathscr{H}_0$, let $\bar{\mathscr{A}}$ as in Chap. 8 be the collection of all $\mathscr{H}_0$-valued $\{\mathscr{F}_t\}$-predictable processes for which $\int_0^T\|\phi(s)\|_0^2ds<\infty$ a.s., where $\|\cdot\|_0$ is the norm in the Hilbert space $\mathscr{H}_0$. We also recall the classes $\mathscr{A}_b$, $\mathscr{A}$ and $\bar{\mathscr{A}}_b$ introduced in Chap. 8. Note that when $\mathscr{H}_0=l_2$, every $u\in\bar{\mathscr{A}}$ can be written as $u=\{u_i\}_{i\in\mathbb{N}}$, where for each i, u_i is a real-valued $\{\mathscr{F}_t\}$-predictable process and $\sum_{i=1}^\infty\int_0^T|u_i(s)|^2ds<\infty$ a.s. The following result is a consequence of Theorem 8.3, Proposition 11.1, and Remark 11.2.

Theorem 11.10 *Let G be a bounded measurable function mapping $\mathscr{C}([0,T]:\mathbb{R}^\infty)$ into $\mathbb{R}$. Then with $\mathscr{H}_0=l_2$, we have*

$$-\log Ee^{-G(\beta)}=\inf_{u=\{u_i\}\in\mathscr{R}}E\left[\frac{1}{2}\int_0^T\sum_{i=1}^\infty|u_i(s)|^2ds+G\left(\beta^u\right)\right],$$

where $\beta_i^{u_i}=\beta_i+\int_0^\cdot u_i(s)ds$, $i\in\mathbb{N}$, $\beta^u\doteq\{\beta_i^{u_i}\}_{i\in\mathbb{N}}$, and $\mathscr{R}$ can be $\bar{\mathscr{A}}$, $\bar{\mathscr{A}}_b$, $\mathscr{A}$, or $\mathscr{A}_b$.

Proof Taking $\mathscr{H}=\bar{l}_2$ introduced in Remark 11.2, it follows from Proposition 11.1 that there is a measurable map $g:\mathscr{C}([0,T]:\mathscr{H})\to\mathscr{C}([0,T]:\mathbb{R}^\infty)$ such that $\beta=g(W)$, where W is as defined in (11.1) with $\{\lambda_i\}$ as in Remark 11.2 and e_i as the vector with the ith coordinate $1/\sqrt{\lambda_i}$ and remaining coordinates 0. Note that the function g can be explicitly written as

$$[g(x)]_i(t)=\frac{1}{\sqrt{\lambda_i}}\langle x(t),e_i\rangle=x_i(t),\quad x\in\mathscr{C}([0,T]:\mathscr{H}),\ i\in\mathbb{N},\ t\in[0,T].$$

From Theorem 8.3, we then have

$$-\log Ee^{-G(\beta)} = -\log Ee^{-G(g(W))}$$
$$= \inf_{u=\{u_i\}\in\mathscr{R}} E\left[\frac{1}{2}\int_0^T \sum_{i=1}^{\infty} |u_i(s)|^2 ds + G\left(g(W^u)\right)\right],$$

where $W^u(t) \doteq W(t) + \int_0^t u(s)ds$. The result now follows on observing that for all $u \in \mathscr{R}$, $g(W^u) = \beta^u$. □

We next note the representation theorem for a Brownian sheet that follows from Proposition 11.9, Theorem 11.10, and an application of Girsanov's theorem. In the statement below, $\bar{\mathscr{A}}$, $\mathscr{A}_b$, $\mathscr{A}$, and $\bar{\mathscr{A}}_b$ are as introduced below Definition 11.7.

Theorem 11.11 *Let $G : \mathscr{C}([0,T] \times \bar{O} : \mathbb{R}) \to \mathbb{R}$ be a bounded measurable map. Let B be a Brownian sheet as in Definition 11.5. Then*

$$-\log Ee^{-G(B)} = \inf_{u\in\mathscr{R}} E\left[\frac{1}{2}\int_0^T \int_O u^2(s,r)drds + G(B^u)\right],$$

where $B^u(t,x) = B(t,x) + \int_0^t \int_{(-\infty,x]\cap O} u(s,y)dyds$ and $\mathscr{R}$ can be $\bar{\mathscr{A}}$, $\bar{\mathscr{A}}_b$, $\mathscr{A}$, or $\mathscr{A}_b$.

Proof We consider only the case $\mathscr{R} = \mathscr{A}_b$, and note that all remaining cases can be treated similarly. Let g be as in Proposition 11.9. To apply the proposition, we need to refer to the analogous set of control processes used in Theorem 11.10, which we denote by $\mathscr{A}_b^\beta$. Then with β as defined above (11.3), we have

$$-\log Ee^{-G(B)} = -\log Ee^{-G(g(\beta))}$$
$$= \inf_{\hat{u}=\{\hat{u}_i\}\in\mathscr{A}_b^\beta} E\left[\frac{1}{2}\int_0^T \sum_{i=1}^{\infty} |\hat{u}_i(s)|^2 ds + G\left(g(\beta^{\hat{u}})\right)\right]. \tag{11.5}$$

Note that there is a one-to-one correspondence between elements of $\mathscr{A}_b$ and $\mathscr{A}_b^\beta$ given through the relations

$$u(t,x) \doteq \sum_{i=1}^{\infty} \hat{u}_i(t)\phi_i(x), \ \ (t,x) \in [0,T]\times O \text{ for } \{\hat{u}_i\} \in \mathscr{A}_b^\beta,$$
$$\hat{u}_i(t) \doteq \int_O u(t,x)\phi_i(x)dx, \ \ t \in [0,T] \text{ for } u \in \mathscr{A}_b.$$

Furthermore,

$$\int_0^T \sum_{i=1}^{\infty} |\hat{u}_i(s)|^2 ds = \int_0^T \int_O u^2(s,r)drds. \tag{11.6}$$

Finally, from Girsanov's theorem, with any u and $\hat{u}$ given by the above relations there is a measure Q that is mutually absolutely continuous with respect to P and is

such that under Q, $(\beta^{\hat{u}}, B^u)$ have the same law as (β, B) under P. Thus $G(g(\beta^{\hat{u}})) = G(B^u)$ a.s., and the result now follows from (11.5) and (11.6). □

The analogous representation holds for cylindrical Brownian motion, with a similar proof. We omit the details.

11.2 General Sufficient Condition for an LDP

In this section we will present sufficient conditions for a uniform Laplace principle that are similar to those presented in Sect. 9.2.1 but with the driving noise a Brownian sheet or an infinite sequence of real Brownian motions (i.e., the $\mathbb{R}^\infty$-valued random variable β), rather than a Hilbert space valued Brownian motion. For simplicity, we do not include a Poisson noise here, although that setting can be covered in a similar manner.

Let, as in Sect. 11.1, $\beta = \{\beta_i\}$ be a sequence of independent standard real $\{\mathscr{F}_t\}$-Brownian motions on $(\Omega, \mathscr{F}, P, \{\mathscr{F}_t\})$. Recall that β is a $\mathscr{C}([0,T] : \mathbb{R}^\infty)$-valued random variable. For each $\varepsilon > 0$, let $\mathscr{G}^\varepsilon : \mathscr{Z} \times \mathscr{C}([0,T] : \mathbb{R}^\infty) \to \mathscr{E}$ be a measurable map, where $\mathscr{Z}$ and $\mathscr{E}$ are Polish spaces, and define

$$X^{\varepsilon,z} \doteq \mathscr{G}^\varepsilon(z, \sqrt{\varepsilon}\beta). \tag{11.7}$$

We now consider the Laplace principle for the family $\{X^{\varepsilon,z}\}$ and introduce the analogue of Condition 9.1 for this setting. In the assumption, S_M and $\bar{\mathscr{A}}_{b,M}$ (the deterministic controls with squared $\mathscr{L}^2$ norm bounded by M and $\{\mathscr{F}_t\}$-predictable processes that take values in S_M, respectively) are defined as in (8.1) and below (8.2), with $\mathscr{H}_0$ there replaced by the Hilbert space l_2.

Condition 11.12 *There exists a measurable map $\mathscr{G}^0 : \mathscr{Z} \times \mathscr{C}([0,T] : \mathbb{R}^\infty) \to \mathscr{E}$ such that the following hold.*

(a) *For every $M < \infty$ and compact set $K \subset \mathscr{Z}$, the set*

$$\Gamma_{M,K} \doteq \left\{ \mathscr{G}^0\left(z, \int_0^\cdot u(s)ds\right) : u \in S_M, z \in K \right\}$$

is a compact subset of $\mathscr{E}$.

(b) *Consider $M < \infty$ and families $\{u^\varepsilon\} \subset \bar{\mathscr{A}}_{b,M}$ and $\{z^\varepsilon\} \subset \mathscr{Z}$ such that u^ε converges in distribution (as S_M-valued random elements) to u and $z^\varepsilon \to z$ as $\varepsilon \to 0$. Then*

$$\mathscr{G}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}\beta + \int_0^\cdot u^\varepsilon(s)ds\right) \to \mathscr{G}^0\left(z, \int_0^\cdot u(s)ds\right),$$

as $\varepsilon \to 0$ in distribution.

The proof of the following uses a straightforward reduction to Theorem 9.2.

Theorem 11.13 *Let $X^{\varepsilon,z}$ be as in (11.7) and suppose that Condition 11.12 holds. For $z \in \mathscr{Z}$ and $\phi \in \mathscr{E}$ let*

$$I_z(\phi) \doteq \inf_{\{u \in \mathscr{L}^2([0,T]:l_2):\phi = \mathscr{G}^0\left(z, \int_0^\cdot u(s)ds\right)\}} \left[\frac{1}{2}\sum_{i=1}^{\infty}\int_0^T |u_i(s)|^2 ds\right]. \tag{11.8}$$

Suppose that for all $\phi \in \mathscr{E}$, $z \mapsto I_z(\phi)$ is a lower semicontinuous map from $\mathscr{Z}$ to $[0,\infty]$. Then for all $z \in \mathscr{E}_0$, $\phi \mapsto I_z(\phi)$ is a rate function on $\mathscr{E}$, and the family $\{I_z(\cdot),\ z \in \mathscr{Z}\}$ of rate functions has compact level sets on compacts. Furthermore, the family $\{X^{\varepsilon,z}\}$ satisfies the Laplace principle on $\mathscr{E}$ with rate function I_z, uniformly on compact subsets of $\mathscr{Z}$.

Proof From Remark 11.2 we can regard β as an $\mathscr{H}$-valued Λ-Wiener process, where $\mathscr{H} = \bar{l}_2$ and Λ is a trace class operator, as defined in Remark 11.2. Also, one can check that $\mathscr{H}_0 \doteq \Lambda^{1/2}\mathscr{H} = l_2$. Since the embedding map $\iota : \mathscr{C}([0,T] : \bar{l}_2) \to \mathscr{C}([0,T] : \mathbb{R}^\infty)$ is measurable (in fact continuous), $\hat{\mathscr{G}}^\varepsilon : \mathscr{Z} \times \mathscr{C}([0,T] : \bar{l}_2) \to \mathscr{E}$ defined by $\hat{\mathscr{G}}^\varepsilon(z,v) \doteq \mathscr{G}^\varepsilon(z, \iota(v))$, $(z,v) \in \mathscr{Z} \times \mathscr{C}([0,T] : \bar{l}_2)$ is a measurable map for every $\varepsilon \geq 0$. Note also that for $\varepsilon > 0$, $X^{\varepsilon,z} = \hat{\mathscr{G}}^\varepsilon(z, \sqrt{\varepsilon}\beta)$ a.s. Since Condition 11.12 holds, we have that parts (a) and (b) of Condition 9.1 are satisfied with $\mathscr{G}^\varepsilon$ there replaced by $\hat{\mathscr{G}}^\varepsilon$ for $\varepsilon \geq 0$ (note that there is no Poisson noise here) and with W replaced with β. Define $\hat{I}_z(\phi)$ by the right side of (9.4) but with $\mathscr{G}^0$ replaced by $\hat{\mathscr{G}}^0$, $S^{\hat{\mathscr{G}}}_{z,\phi} \doteq \left\{f \in \mathscr{L}^2([0,T] : \mathscr{H}_0) : \phi = \mathscr{G}^0(z, \int_0^\cdot f(s)ds)\right\}$, and $\bar{L}_T(q)$ replaced by $\frac{1}{2}\int_0^T \|f(s)\|_0^2 ds$, so that

$$\hat{I}_z(\phi) = \inf_{f \in S^{\hat{\mathscr{G}}}_{z,\phi}} \left[\frac{1}{2}\int_0^T \|f(s)\|_0^2 ds\right].$$

Clearly $I_z(\phi) = \hat{I}_z(\phi)$ for all $(z,\phi) \in \mathscr{Z} \times \mathscr{E}$. The result is now an immediate consequence of Theorem 9.2. □

Remark 11.14 Since for $t \in (0,T)$, $\sum_{i=1}^\infty (\beta_i(t))^2 = \infty$ a.s., the $\mathbb{R}^\infty$-valued random variable $\beta(t)$ does not lie in the subset l_2 of $\mathbb{R}^\infty$. However, for any sequence $\{\lambda_i\}$ as in Remark 11.2, $\sum_{i=1}^\infty \lambda_i(\beta_i(t))^2 < \infty$ a.s., which shows that the support of $\beta(t)$ does lie in the larger Hilbert space $\bar{l}_2$. In fact, $t \mapsto \beta(t)$ is a.s. a continuous map from $[0,T]$ to $\bar{l}_2$, and it is easily checked that it defines a Λ-Wiener process with sample paths in $\bar{l}_2$. This identification of β with a Hilbert space valued Wiener process allows us to leverage Theorem 9.2 in establishing Theorem 11.13. Note that there are many different possible choices of sequences $\{\lambda_i\}$ (and corresponding Hilbert spaces $\bar{l}_2$) and any of them can be used to prove the theorem, which itself does not involve any specific Hilbert space.

To close this section, we consider the Laplace principle for functionals of a Brownian sheet. Let B be a Brownian sheet as in Definition 11.5. Let $\mathscr{G}^\varepsilon : \mathscr{Z} \times \mathscr{C}([0,T] \times \bar{O} : \mathbb{R}) \to \mathscr{E}$, $\varepsilon > 0$, be a family of measurable maps. Define $X^{\varepsilon,z} \doteq \mathscr{G}^\varepsilon(z, \sqrt{\varepsilon}B)$.

We now provide sufficient conditions for a Laplace principle to hold for the family $\{X^{\varepsilon,z}\}$.

Analogous to the classes defined in (8.1), we introduce for $N \in (0, \infty)$,

$$S_N \doteq \left\{\phi \in \mathscr{L}^2([0,T] \times O) : \int_{[0,T]\times O} \phi^2(s,r)dsdr \leq N\right\},$$

$$\bar{\mathscr{A}}_{b,N} \doteq \{u \in \bar{\mathscr{A}} : u(\omega) \in S_N, P\text{-a.s.}\}. \tag{11.9}$$

Once more, S_N is endowed with the weak topology on $\mathscr{L}^2([0,T] \times O)$, under which it is a compact metric space. For $u \in \mathscr{L}^2([0,T] \times O)$, define $\mathrm{Int}(u) \in \mathscr{C}([0,T] \times O : \mathbb{R})$ by

$$\mathrm{Int}(u)(t,x) \doteq \int_{[0,t]\times(O\cap(-\infty,x])} u(s,y)dsdy, \tag{11.10}$$

where as before, $(-\infty, x] \doteq \{y : y_i \leq x_i \text{ for all } i = 1, \ldots, d\}$.

Condition 11.15 *There exists a measurable map $\mathscr{G}^0 : \mathscr{Z} \times \mathscr{C}([0,T] \times O : \mathbb{R}) \to \mathscr{E}$ such that the following hold.*

(a) *For every $M < \infty$ and compact set $K \subset \mathscr{Z}$, the set*

$$\Gamma_{M,K} \doteq \left\{\mathscr{G}^0(z, \mathrm{Int}(u)) : u \in S_M,\ z \in K\right\}$$

is a compact subset of $\mathscr{E}$, where Int(u) *is as in* (11.10).

(b) *Consider $M < \infty$ and families $\{u^\varepsilon\} \subset \bar{\mathscr{A}}_{b,M}$ and $\{z^\varepsilon\} \subset \mathscr{Z}$ such that u^ε converges in distribution (as S_M-valued random elements) to u and $z^\varepsilon \to z$ as $\varepsilon \to 0$. Then*

$$\mathscr{G}^\varepsilon\left(z^\varepsilon, \sqrt{\varepsilon}B + \mathrm{Int}(u^\varepsilon)\right) \to \mathscr{G}^0\left(z, \mathrm{Int}(u)\right)$$

in distribution as $\varepsilon \to 0$.

For $f \in \mathscr{E}$ and $z \in \mathscr{Z}$, define

$$I_z(f) = \inf_{\{u\in\mathscr{L}^2([0,T]\times O):\, f=\mathscr{G}^0(z,\mathrm{Int}(u))\}} \left[\frac{1}{2}\int_{[0,T]\times O} u^2(s,r)drds\right]. \tag{11.11}$$

Theorem 11.16 *Let $\mathscr{G}^0 : \mathscr{Z} \times \mathscr{C}([0,T] \times O : \mathbb{R}) \to \mathscr{E}$ be a measurable map satisfying Condition 11.15. Suppose that for all $f \in \mathscr{E}$, $z \mapsto I_z(f)$ is a lower semicontinuous map from $\mathscr{Z}$ to $[0,\infty]$. Then for every $z \in \mathscr{Z}$, $I_z : \mathscr{E} \to [0,\infty]$, defined by (11.11), is a rate function on $\mathscr{E}$, and the family $\{I_z, z \in \mathscr{Z}\}$ of rate functions has compact level sets on compacts. Furthermore, the family $\{X^{z,\varepsilon}\}$ satisfies the Laplace principle on $\mathscr{E}$ with rate function I_z, uniformly for z in compact subsets of $\mathscr{Z}$.*

Proof Let $\{\phi_i\}_{i=1}^{\infty}$ be a CONS in $\mathscr{L}^2(O)$ and let

$$\beta_i(t) \doteq \int_{[0,t]\times O} \phi_i(x) B(ds \times dx), \quad t \in [0,T],\ i \in \mathbb{N}.$$

Then $\beta = \{\beta_i\}$ is a sequence of independent standard real Brownian motions, and it can be regarded as a $\mathscr{C}([0,T] : \mathbb{R}^{\infty})$-valued random variable. Furthermore, (11.3) is satisfied, and from Proposition 11.9, there is a measurable map $g : \mathscr{C}([0,T] : \mathbb{R}^{\infty}) \to \mathscr{C}([0,T] \times O : \mathbb{R})$ such that $g(\beta) = B$ a.s. For $\varepsilon > 0$, define $\hat{\mathscr{G}}^{\varepsilon} : \mathscr{Z} \times \mathscr{C}([0,T] : \mathbb{R}^{\infty}) \to \mathscr{E}$ by $\hat{\mathscr{G}}^{\varepsilon}(z, \sqrt{\varepsilon} v) \doteq \mathscr{G}^{\varepsilon}(z, \sqrt{\varepsilon} g(v))$, $(z, v) \in \mathscr{Z} \times \mathscr{C}([0,T] : \mathbb{R}^{\infty})$. Clearly, $\hat{\mathscr{G}}^{\varepsilon}$ is a measurable map and $\hat{\mathscr{G}}^{\varepsilon}(z, \sqrt{\varepsilon}\beta) = X^{\varepsilon,z}$ a.s. Next, note that

$$\left\{ v \in \mathscr{C}([0,T] : \mathbb{R}^{\infty}) : v(\cdot) = \int_0^{\cdot} \hat{u}(s) ds, \text{ for some } \hat{u} \in \mathscr{L}^2([0,T] : l_2) \right\}$$

is a measurable subset of $\mathscr{C}([0,T] : \mathbb{R}^{\infty})$. For $\hat{u} \in \mathscr{L}^2([0,T] : l_2)$, define $u_{\hat{u}} \in \mathscr{L}^2([0,T] \times O)$ by

$$u_{\hat{u}}(t,x) \doteq \sum_{i=1}^{\infty} \hat{u}_i(t) \phi_i(x), \quad (t,x) \in [0,T] \times O.$$

Define $\hat{\mathscr{G}}^0 : \mathscr{Z} \times \mathscr{C}([0,T] : \mathbb{R}^{\infty}) \to \mathscr{E}$ by

$$\hat{\mathscr{G}}^0(z,v) \doteq \mathscr{G}^0(z, \mathrm{Int}(u_{\hat{u}})) \quad \text{if } v = \int_0^{\cdot} \hat{u}(s) ds \text{ and } \hat{u} \in \mathscr{L}^2([0,T] : l_2),$$

and set $\hat{\mathscr{G}}^0(z,v) \doteq 0$ for all other (z,v). Note that

$$\left\{ \hat{\mathscr{G}}^0 \left(z, \int_0^{\cdot} \hat{u}(s) ds \right) : \hat{u} \in S_M, z \in K \right\} = \left\{ \mathscr{G}^0 \left(z, \mathrm{Int}(u) \right) : u \in S_M, z \in K \right\},$$

where S_M on the left side is the one introduced above Condition 11.12, and S_M on the right side is the one introduced above (11.9). Since Condition 11.15 holds, we have that part (a) of Condition 11.12 holds with $\mathscr{G}^0$ there replaced by $\hat{\mathscr{G}}^0$. Next, an application of Girsanov's theorem (see the proof of Theorem 11.11) gives that for every $\hat{u}^{\varepsilon} \in \bar{\mathscr{A}}_{b,M}$ (where the latter class is as in Condition 11.12),

$$g\left(\beta + \frac{1}{\sqrt{\varepsilon}} \int_0^{\cdot} \hat{u}^{\varepsilon}(s) ds \right) = B + \frac{1}{\sqrt{\varepsilon}} \mathrm{Int}(u_{\hat{u}^{\varepsilon}}),$$

a.s. In particular, for every $M < \infty$ and families $\{\hat{u}^{\varepsilon}\} \subset \bar{\mathscr{A}}_{b,M}$ and $\{z^{\varepsilon}\} \subset \mathscr{Z}$ such that $\hat{u}^{\varepsilon}$ converges in distribution (as S_M-valued random elements) to $\hat{u}$ and $z^{\varepsilon} \to z$, we have

$$\lim_{\varepsilon\to 0}\hat{\mathscr{G}}^{\varepsilon}\left(z^{\varepsilon},\sqrt{\varepsilon}\beta+\int_0^{\cdot}\hat{u}^{\varepsilon}(s)ds\right)=\lim_{\varepsilon\to 0}\mathscr{G}^{\varepsilon}\left(z^{\varepsilon},\sqrt{\varepsilon}B+\mathrm{Int}(u_{\hat{u}^{\varepsilon}})\right)$$
$$=\mathscr{G}^{0}\left(z,\mathrm{Int}(u_{\hat{u}})\right)$$
$$=\hat{\mathscr{G}}^{0}\left(z,\int_0^{\cdot}\hat{u}(s)ds\right).$$

Thus w.p.1, part (b) of Condition 11.12 is satisfied with $\mathscr{G}^{\varepsilon}$ replaced by $\hat{\mathscr{G}}^{\varepsilon}$, $\varepsilon\geq 0$. The result now follows on noting that if $\hat{I}_z(f)$ is defined by the right side of (11.8) but with $\mathscr{G}^0$ there replaced by $\hat{\mathscr{G}}^0$, then $\hat{I}_z(f)=I_z(f)$ for all $(z,f)\in\mathscr{Z}\times\mathscr{E}$. □

11.3 Reaction–Diffusion SPDE

In this section we will use results from Sect. 11.2, and in particular Theorem 11.16, to study the small noise large deviation principle for a class of SPDE that was considered in [174]. The class includes, as a special case, the reaction–diffusion SPDEs considered in [235] (see Remark 11.22). The main result of the section is Theorem 11.21, which establishes the uniform Laplace principle for such SPDE.

As the discussion at the beginning of this Part III of the book indicates, this is but one of many possible applications of the abstract LDP (Theorem 9.2), though for this particular application we of course use the version appropriate for a Brownian sheet (Theorem 11.16). A main purpose of the presentation is to illustrate the claim that the essential issue in proving an LDP is a good qualitative theory for controlled versions of the original system under a law of large numbers scaling. Since we do not wish to prove this qualitative theory again, in this section we extensively apply results proved elsewhere, and in that sense, this section is not self-contained. This situation illustrates the fact that in any particular application of Theorem 9.2, one needs a thorough understanding of the qualitative properties of the infinite dimensional system under consideration.

11.3.1 The Large Deviation Theorem

Let $(\Omega,\mathscr{F},P)$ be a probability space with a filtration $\{\mathscr{F}_t\}_{0\leq t\leq T}$ satisfying the usual conditions. Let $O\subset\mathbb{R}^d$ be a bounded open set and $\{B(t,x):(t,x)\in\mathbb{R}_+\times O\}$ a Brownian sheet on this filtered probability space. Consider the SPDE

$$dX(t,r)=[L(t)X(t,r)+R(t,r,X(t,r))]drdt+\sqrt{\varepsilon}A(t,r,X(t,r))\,B(dr\times dt) \tag{11.12}$$

with initial condition $X(0,r)=x(r)$. Here $\{L(t)\}_{0\leq t<\infty}$ is a family of linear, closed, densely defined operators on $\mathscr{C}(O)$ that generates a two-parameter strongly

continuous semigroup (see [174, Sect. 1]) $\{U(t,s)\}_{0\le s\le t}$ on $\mathscr{C}(O)$, with kernel function $G(t,s,r,q)$, $0 \le s \le t, r, q \in O$. Thus for $f \in \mathscr{C}(O)$,

$$[U(t,s)f](r) = \int_O G(t,s,r,q) f(q) dq, \ r \in O, \ 0 \le s \le t \le T.$$

Also, A and R are measurable maps from $[0,T] \times O \times \mathbb{R}$ to $\mathbb{R}$ and $\varepsilon \in (0,\infty)$. By a solution of the SPDE (11.12), we mean the following.

Definition 11.17 A random field $X = \{X(t,r) : t \in [0,T], r \in O\}$ is called a mild solution of the stochastic partial differential equation (11.12) with initial condition ξ if $(t,r) \mapsto X(t,r)$ is continuous, $X(t,r)$ is $\{\mathscr{F}_t\}$-measurable for all $t \in [0,T]$ and $r \in O$, and if a.s. for all $t \in [0,T]$,

$$\begin{aligned} X(t,r) = &\int_O G(t,0,r,q)x(q)dq + \int_0^t \int_O G(t,s,r,q) R(s,q,X(s,q))\, dqds \\ &+ \sqrt{\varepsilon} \int_0^t \int_O G(t,s,r,q) A(s,q,X(s,q))\, B(dq \times ds). \end{aligned} \tag{11.13}$$

Implicit in Definition 11.17 is the requirement that the integrals in (11.13) be well defined. We will shortly introduce conditions on G, A, and R that ensure that for a continuous adapted random field X, all the integrals in (11.13) are meaningful. As a convention, we take $G(t,s,r,q)$ to be zero when $0 \le t < s \le T, r, q \in O$.

For $u \in \bar{\mathscr{A}}_{b,N}$ [which was defined in (11.9)], the controlled analogue of (11.13) is

$$\begin{aligned} Y(t,r) = &\int_O G(t,0,r,q)x(q)dq + \int_0^t \int_O G(t,s,r,q) R(s,q,Y(s,q)) dqds \\ &+ \sqrt{\varepsilon} \int_0^t \int_O G(t,s,r,q) A(s,q,Y(s,q)) B(dq \times ds) \\ &+ \int_0^t \int_O G(t,s,r,q) A(s,q,Y(s,q)) u(s,q) dqds. \end{aligned} \tag{11.14}$$

The main work in proving an LDP for (11.13) is to prove qualitative properties (existence and uniqueness, tightness properties, and stability under perturbations) for solutions to (11.14). We begin by discussing the known qualitative theory for (11.13).

For $\alpha \in (0,\infty)$, let $\mathbb{B}_\alpha \doteq \{\psi \in \mathscr{C}(O) : \|\psi\|_\alpha < \infty\}$ be the Banach space with norm

$$\|\psi\|_\alpha \doteq \|\psi\|_0 + \sup_{r,q \in O, r \ne q} \frac{|\psi(r) - \psi(q)|}{\|r-q\|^\alpha}, \tag{11.15}$$

where $\|\psi\|_0 \doteq \sup_{r \in O} |\psi(r)|$. The Banach space $\mathbb{B}_\alpha([0,T] \times O)$ is defined as in (11.15) but with O replaced by $[0,T] \times O$, and for notational convenience we denote this space by $\mathbb{B}_\alpha^T$. For $\alpha = 0$, $\mathbb{B}_0^T$ is the space of all continuous maps from $[0,T] \times \bar{O}$ to $\mathbb{R}$ endowed with the sup-norm. The following will be a standing assumption

for this section. In the assumption, $\bar{\alpha}$ is a fixed constant, and the large deviation principle will be proved in the topology of $\mathscr{C}([0,T]:\mathbb{B}_\alpha)$, for any fixed $\alpha \in (0,\bar{\alpha})$. Using the contraction principle, this large deviation principle provides large deviation asymptotics for the evaluation $X(t,r)$ for every fixed $(t,r) \in [0,T]\times O$, and for many other functionals as well, e.g., $\sup_{t\in[0,T]} \|X(t,\cdot)\|_\alpha$, $\sup_{(t,r)\in[0,T]\times O} |X(t,r)|$. Recall that $O \subset \mathbb{R}^d$. The following condition is taken from [170].

Condition 11.18 *The following two conditions hold.*

(a) *There exist constants $K(T) < \infty$ and $\gamma \in (d,\infty)$ such that*

(i) *for all $t,s \in [0,T]$, $r \in O$,*

$$\int_O |G(t,s,r,q)|dq \le K(T); \tag{11.16}$$

(ii) *for all $0 \le s < t \le T$ and $r,q \in O$,*

$$|G(t,s,r,q)| \le K(T)(t-s)^{-\frac{d}{\gamma}}; \tag{11.17}$$

(iii) *if $\bar{\alpha} \doteq \frac{\gamma-d}{2\gamma}$, then for all $\alpha \in (0,\bar{\alpha})$ and for all $0 \le s < t_1 \le t_2 \le T$, $r_1, r_2, q \in O$,*

$$\begin{aligned} &|G(t_1,s,r_1,q) - G(t_2,s,r_2,q)| \\ &\quad \le K(T)\left[(t_2-t_1)^{1-\frac{d}{\gamma}}(t_1-s)^{-1} + |r_1-r_2|^{2\alpha}(t_1-s)^{-\frac{d+2\alpha}{\gamma}}\right]; \end{aligned} \tag{11.18}$$

(iv) *for all $z,y \in \mathbb{R}$, $r \in O$, and $0 \le t \le T$,*

$$|R(t,r,z) - R(t,r,y)| + |A(t,r,z) - A(t,r,y)| \le K(T)|z-y|$$

and

$$|R(t,r,z)| + |A(t,r,z)| \le K(T)(1+|z|). \tag{11.19}$$

(b) *For all $\alpha \in (0,\bar{\alpha})$ and $\xi \in \mathbb{B}_\alpha$, $\hat{\xi}(t) \doteq \int_O G(t,0,\cdot,q)\xi(q)dq$ belongs to $\mathbb{B}_\alpha$ and $\hat{\xi} \in \mathscr{C}([0,T]:\mathbb{B}_\alpha)$. The map $\xi \mapsto \hat{\xi}$ is a continuous map from $\mathbb{B}_\alpha$ to $\mathscr{C}([0,T]:\mathbb{B}_\alpha)$.*

Remark 11.19 (a) Note that the definition $\bar{\alpha} \doteq (\gamma-d)/2\gamma$ implies $\bar{\alpha} \in (0,1/2)$.

(b) We refer the reader to [169] for examples of families $\{L(t)\}_{t\ge 0}$ that satisfy Condition 11.18.

(c) Using (11.16) and (11.17), it follows that for all $0 \le s < t \le T$ and $r \in O$,

$$\int_O |G(t,s,r,q)|^2 dq \le K^2(T)(t-s)^{-\frac{d}{\gamma}}. \tag{11.20}$$

Since $\gamma > d$, the estimate (11.20) says that

$$\sup_{(r,t)\in O\times[0,T]} \int_{[0,t]\times O} |G(t,s,r,q)|^2 dq < \infty, \tag{11.21}$$

which in view of the linear growth assumption in (11.19) ensures that the stochastic integral in (11.13) is well defined.

(d) Lemma 4.1(ii) of [169] shows that under Condition 11.18, for every $\alpha < \bar{\alpha}$ there exists a constant $\tilde{K}(\alpha)$ such that for all $0 \le t_1 \le t_2 \le T$ and all $r_1, r_2 \in O$,

$$\int_0^T \int_O |G(t_1,s,r_1,q) - G(t_2,s,r_2,q)|^2 dq\, ds \le \tilde{K}(\alpha)\rho\left((t_1,r_1),(t_2,r_2)\right)^{2\alpha},$$

where ρ is the Euclidean distance in $[0,T]\times O \subset \mathbb{R}^{d+1}$. This estimate will be used in the proof of Lemma 11.28.

The following theorem is due to Kotelenez (see Theorems 2.1 and 3.4 in [174]; see also Theorem 3.1 in [169]).

Theorem 11.20 *Assume Condition 11.18 and fix $\alpha \in (0,\bar{\alpha})$. There exists a measurable function*

$$\mathscr{G}^\varepsilon : \mathbb{B}_\alpha \times \mathbb{B}_0^T \to \mathscr{C}([0,T]:\mathbb{B}_\alpha)$$

such that for every filtered probability space $(\Omega,\mathscr{F},P,\{\mathscr{F}_t\})$ with a Brownian sheet B, $X_x^\varepsilon \doteq \mathscr{G}^\varepsilon(x,\sqrt{\varepsilon}B)$ is the unique mild solution of (11.12) (with initial condition x), and it satisfies $\sup_{0\le t\le T} E\|X_x^\varepsilon(t)\|_0^p < \infty$ for all $p \in [0,\infty)$.

For the rest of the section we consider only $\alpha \in (0,\bar{\alpha})$. For $f \in \mathscr{C}([0,T]:\mathbb{B}_\alpha)$, define

$$I_x(f) \doteq \inf_u \int_{[0,T]\times O} u^2(s,q)dsdq, \tag{11.22}$$

where the infimum is taken over all $u \in \mathscr{L}^2([0,T]\times O)$ such that

$$\begin{aligned} f(t,r) = &\int_O G(t,0,r,q)x(q)dq + \int_{[0,t]\times O} G(t,s,r,q)R(s,q,f(s,q))dsdq \\ &+ \int_{[0,t]\times O} G(t,s,r,q)A(s,q,f(s,q))u(s,q)dsdq. \end{aligned} \tag{11.23}$$

The following is the main result of this section, which is a uniform Laplace principle for $\{X_x^\varepsilon\}$. The definition of a uniform Laplace principle was given in Chap. 1. There the dependence on the parameter over which uniformity is considered was noted in the expectation operator. In this chapter, however, it will be more convenient to work with a common probability measure (instead of a collection parametrized by $x \in \mathbb{B}_\alpha$) and instead note the dependence on x in the collection of random variables, i.e., we write X_x^ε to note this dependence.

Theorem 11.21 *Assume Condition 11.18, let $\alpha \in (0,\bar{\alpha})$, and let X_x^ε be as in Theorem 11.20. Then I_x defined by (11.22) is a rate function on $\mathscr{C}([0,T]:\mathbb{B}_\alpha)$, and the*

family $\{I_x, x \in \mathbb{B}_\alpha\}$ *of rate functions has compact level sets on compacts. Furthermore,* $\{X_x^\varepsilon\}$ *satisfies the Laplace principle on* $\mathscr{C}([0,T]:\mathbb{B}_\alpha)$ *with the rate function* I_x*, uniformly for* x *in compact subsets of* $\mathbb{B}_\alpha$*.*

Remark 11.22 (a) If part (b) of Condition 11.18 is weakened to merely the requirement that for every $\xi \in \mathbb{B}_\alpha$, $t \mapsto \int_O G(t,0,\cdot,q)\xi(q)dq$ be in $\mathscr{C}([0,T]:\mathbb{B}_\alpha)$, then the proof of Theorem 11.21 shows that for all $x \in \mathbb{B}_\alpha$, the large deviation principle for $\{X_x^\varepsilon\}$ on $\mathscr{C}([0,T]:\mathbb{B}_\alpha)$ holds (but not necessarily uniformly).

(b) The small noise LDP for a class of reaction–diffusion SPDEs, with $O = [0,1]$ and a bounded diffusion coefficient, has been studied in [235]. A difference in the conditions on the kernel G in [235] is that instead of (11.18), G satisfies the $\mathscr{L}^2$ estimate in Remark 11.19 (c) with $\bar{\alpha} = 1/4$. One finds that the proof of Lemma 11.28, which is at the heart of the proof of Theorem 11.21, uses only the $\mathscr{L}^2$ estimate rather than the condition (11.18). Using this observation and techniques in the proof of Theorem 11.21, one can extend results of [235] to the case in which the diffusion coefficient, instead of being bounded, satisfies the linear growth condition (11.19).

Since the proof of Theorem 11.21 relies on properties of the controlled process (11.14), the first step is to prove existence and uniqueness of solutions. This follows from a standard application of Girsanov's theorem. Following the convention used throughout the book, we denote the controlled version of the SPDE by an overbar.

Theorem 11.23 *Let* $\mathscr{G}^\varepsilon$ *be as in Theorem 11.20 and let* $u \in \bar{\mathscr{A}}_{b,N}$ *for some* $N \in \mathbb{N}$*, where* $\bar{\mathscr{A}}_{b,N}$ *is as defined in (11.9). For* $\varepsilon > 0$ *and* $x \in \mathbb{B}_\alpha$*, define*

$$\bar{X}_x^\varepsilon \doteq \mathscr{G}^\varepsilon\left(x, \sqrt{\varepsilon}B + \mathrm{Int}(u)\right),$$

where Int *is defined in (11.10). Then* $\bar{X}_x^\varepsilon$ *is the unique solution of (11.14).*

Proof Fix $u \in \bar{\mathscr{A}}_{b,N}$. Since

$$E\left[\exp\left\{-\frac{1}{\sqrt{\varepsilon}}\int_{[0,T]\times O} u(s,q)B(ds\times dq) - \frac{1}{2\varepsilon}\int_{[0,T]\times O} u^2(s,q)dsdq\right\}\right] = 1,$$

the measure $\gamma^{u,\varepsilon}$ defined by

$$d\gamma^{u,\varepsilon} = \exp\left\{-\frac{1}{\sqrt{\varepsilon}}\int_{[0,T]\times O} u(s,q)B(ds\times dq) - \frac{1}{2\varepsilon}\int_{[0,T]\times O} u^2(s,q)dsdq\right\} dP$$

is a probability measure on $(\Omega, \mathscr{F}, P)$. Furthermore, $\gamma^{u,\varepsilon}$ is mutually absolutely continuous with respect to P, and by Girsanov's theorem (Theorem D.2), the process $B^{u/\sqrt{\varepsilon}} \doteq B + \varepsilon^{-1/2}\mathrm{Int}(u)$ on $(\Omega, \mathscr{F}, \gamma^{u,\varepsilon}, \{\mathscr{F}_t\})$ is a Brownian sheet. Thus by Theorem 11.20, $\bar{X}_x^\varepsilon = \mathscr{G}^\varepsilon\left(x, \sqrt{\varepsilon}B + \mathrm{Int}(u)\right)$ is the unique solution of (11.13), with B there replaced by $B^{u/\sqrt{\varepsilon}}$, on $(\Omega, \mathscr{F}, \gamma^{u,\varepsilon}, \{\mathscr{F}_t\})$. However, equation (11.13) with $B^{u/\sqrt{\varepsilon}}$ is precisely the same as equation (11.14), and since $\gamma^{u,\varepsilon}$ and P are mutually absolutely continuous, we get that $\bar{X}_x^\varepsilon$ is the unique solution of (11.14) on $(\Omega, \mathscr{F}, P, \{\mathscr{F}_t\})$. This completes the proof. □

We next state two basic qualitative results regarding the processes $\bar{X}^\varepsilon_x$ that hold under Condition 11.18. The first is simply the controlled zero-noise version of the theorem just stated. Its proof follows from a simpler version of the arguments used in [174] to establish Theorem 11.20, and thus is omitted. The second is a standard convergence result, whose proof is given in Sect. 11.3.2.

Theorem 11.24 *Assume Condition 11.18, let $\alpha \in (0, \bar{\alpha})$, and fix $x \in \mathbb{B}_\alpha$ and $u \in \mathscr{L}^2([0,T] \times O)$. Then there is a unique function f in $\mathscr{C}([0,T] : \mathbb{B}_\alpha)$ that satisfies equation (11.23).*

In analogy with the notation $\bar{X}^\varepsilon_x$ for the solution of (11.14), we denote the unique solution f given by Theorem 11.24 by $\bar{X}^0_x$.

Theorem 11.25 *Assume Condition 11.18 and let $\alpha \in (0, \bar{\alpha})$. Let $M < \infty$, and suppose that $x^\varepsilon \to x$ and $u^\varepsilon \to u$ in distribution as $\varepsilon \to 0$ with $\{u^\varepsilon\} \subset \mathscr{A}_{b,M}$. Let $\bar{X}^\varepsilon_{x^\varepsilon}$ solve (11.14) with $u = u^\varepsilon$, and let $\bar{X}_x$ solve (11.14). Then $\bar{X}^\varepsilon_{x^\varepsilon} \to \bar{X}_x$ in distribution.*

Remark 11.26 As noted several times already in this book, the same analysis as that used to establish the large deviation bounds (and in particular the large deviation upper bound) typically yields compactness of level sets for the associated rate function. In the present setting, we note that the same argument used to prove Theorem 11.25 but with ε set to zero shows the following (under Condition 11.18). Suppose that $x_n \to x$ and $u_n \to u$ with $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{B}_\alpha$ and $\{u_n\}_{n\in\mathbb{N}} \subset S_M$, and that f_n solves (11.14) when (x,u) is replaced by (x_n, u_n). Then $f_n \to f$.

Proof (*of Theorem 11.21*) Define the map $\mathscr{G}^0 : \mathbb{B}_\alpha \times \mathbb{B}_0^T \to \mathscr{C}([0,T] : \mathbb{B}_\alpha)$ as follows. If $x \in \mathbb{B}_\alpha$ and $\phi \in \mathbb{B}_0^T$ is of the form $\phi(t,x) \doteq \mathrm{Int}(u)(t,x)$ for some $u \in \mathscr{L}^2([0,T] \times O)$, we define $\mathscr{G}^0(x,\phi)$ to be the solution f to (11.23). Let $\mathscr{G}^0(x,\phi) = 0$ for all other $\phi \in \mathbb{B}_0^T$. In view of Theorem 11.16, it suffices to show that $(\mathscr{G}^\varepsilon, \mathscr{G}^0)$ satisfy Condition 11.15 with $\mathscr{Z}$ and $\mathscr{E}$ there replaced by $\mathbb{B}_\alpha$ and $\mathscr{C}([0,T] : \mathbb{B}_\alpha)$, respectively, and that for all $f \in \mathscr{E}$, the map $x \mapsto I_x(f)$ is lower semicontinuous. The latter property and the first part of Condition 11.15 follow directly from Theorem 11.24 and Remark 11.26. The second part of Condition 11.15 follows from Theorem 11.25. □

Thus all that remains to complete the proof is to verify Theorem 11.25.

11.3.2 Qualitative Properties of Controlled Stochastic Reaction–Diffusion Equations

This section is devoted to the proof of Theorem 11.25. Throughout this section we assume Condition 11.18 and consider any fixed $\alpha \in (0, \bar{\alpha})$, where $\bar{\alpha} \doteq (\gamma - d)/2\gamma$. Whenever a control u appears, the associated controlled SPDE is of the form (11.14), and its solution is denoted by $\bar{X}^\varepsilon_x$. Our first result shows that $\mathscr{L}^p$ bounds hold for controlled SDEs, uniformly when the initial condition and controls lie in compact sets and $\varepsilon \in [0,1)$.

Lemma 11.27 *If K is any compact subset of $\mathbb{B}_\alpha$ and $M < \infty$, then for every $p \in [1, \infty)$,*

$$\sup_{u \in \bar{\mathscr{A}}_{b,M}} \sup_{x \in K} \sup_{\varepsilon \in [0,1)} \sup_{(t,r) \in [0,T] \times O} E \|\bar{X}_x^\varepsilon(t,r)\|^p < \infty.$$

Proof By Hölder's inequality, it suffices to establish the claim for all sufficiently large p. Using the standard bound for the pth power of a sum in terms of the pth powers of the summands and Doob's inequality (D.2) for the stochastic integral, there exists $c_1 \in (0, \infty)$ such that

$$\begin{aligned} E\|\bar{X}_x^\varepsilon(t,r)\|^p \le\ & c_1 \left\| \int_O G(t,0,r,q)x(q)dq \right\|^p \\ & + c_1 E \left\| \int_0^t \int_O G(t,s,r,q) R\left(s,q,\bar{X}_x^\varepsilon(s,q)\right) dqds \right\|^p \\ & + c_1 E \left[\int_0^t \int_O |G(t,s,r,q)|^2 \left| A\left(s,q,\bar{X}_x^\varepsilon(s,q)\right)\right|^2 dqds \right]^{\frac{p}{2}} \\ & + c_1 E \left[\int_0^t \int_O |G(t,s,r,q)|\, |A\left(s,q,\bar{X}_x^\varepsilon(s,q)\right)|\, |u(s,q)| dqds \right]^p. \end{aligned}$$

Using (11.19) and the Cauchy-Schwarz inequality, the entire sum on the right-hand side above can be bounded by

$$c_2 \left[1 + E \left[\int_0^t \int_O |G(t,s,r,q)|^2 \|\bar{X}_x^\varepsilon(s,q)\|^2 dq\, ds \right]^{\frac{p}{2}} \right].$$

If $p > 2$, then Hölder's inequality yields

$$\Lambda_p(t) \le c_2 \left[1 + \left(\int_0^t \int_O |G(t,s,r,q)|^{2\tilde{p}} dq\, ds \right)^{\frac{p-2}{2}} \int_0^t \Lambda_p(s) ds \right],$$

where

$$\Lambda_p(t) \doteq \sup_{u \in \bar{\mathscr{A}}_{b,M}} \sup_{x \in K} \sup_{\varepsilon \in [0,1)} \sup_{r \in O} E \|\bar{X}_x^\varepsilon(t,r)\|^p$$

and $\tilde{p} \doteq p/(p-2)$. Recall that $\bar{\alpha} < 1/2$ (see Remark 11.19). Using (11.16) and (11.17), we obtain

$$\int_O |G(t,s,r,q)|^{2\tilde{p}} dq \le (K(T))^{2\tilde{p}} (t-s)^{-\frac{d}{\gamma}(2\tilde{p}-1)}.$$

Suppose p_0 is large enough that $(\frac{2p_0}{p_0-2} - 1)(1 - 2\bar{\alpha}) < 1$. Noting that $\bar{\alpha} \doteq (\gamma - d)/2\gamma$ implies $d/\gamma = 1 - 2\bar{\alpha}$, we have that for all $p \ge p_0$ and $t \in [0, T]$,

$$\left[\int_0^t \int_O |G(t,s,r,q)|^{2\tilde{p}} dq ds\right]^{\frac{p-2}{2}} \leq c_3 T^{[1-(2\tilde{p}-1)(1-2\tilde{\alpha})]\frac{p-2}{2}}.$$

Thus for every $p \geq p_0$ there exists a constant c_4 such that

$$\Lambda_p(t) \leq c_4 \left[1 + \int_0^t \Lambda_p(s) ds\right].$$

The result now follows from Gronwall's lemma. □

The following lemma will be instrumental in proving tightness and weak convergence in Banach spaces such as $\mathbb{B}_\alpha$ and $\mathbb{B}_\alpha^T$. Recall that ρ denotes the Euclidean distance in $[0,T] \times O \subset \mathbb{R}^{d+1}$.

Lemma 11.28 *Let $\mathscr{V}$ be a collection of $\mathbb{R}^d$-valued predictable processes such that for all $p \in [2,\infty)$,*

$$\sup_{f\in\mathscr{V}} \sup_{(t,r)\in[0,T]\times O} \mathbb{E}\|f(t,r)\|^p < \infty. \tag{11.24}$$

Also, let $\mathscr{U} \subset \bar{\mathscr{A}}_{b,M}$ for some $M < \infty$. For $f \in \mathscr{V}$ and $u \in \mathscr{U}$, define

$$\Psi_1(t,r) \doteq \int_0^t \int_O G(t,s,r,q) f(s,q) B(dq \times ds),$$
$$\Psi_2(t,r) \doteq \int_0^t \int_O G(t,s,r,q) f(s,q) u(s,q) dq ds,$$

where the dependence on f and u is not made explicit in the notation. Then for all $\alpha < \bar{\alpha}$ and $i = 1,2$,

$$\sup_{f\in\mathscr{V},u\in\mathscr{U}} E\left[\sup_{\rho((t,r),(s,q))<1} \frac{\|\Psi_i(t,r) - \Psi_i(s,q)\|}{\rho((t,r),(s,q))^\alpha}\right] < \infty.$$

Proof We will prove the result for $i = 1$; the proof for $i = 2$ is identical (except for an additional application of the Cauchy-Schwarz inequality), and thus omitted. Henceforth we write, for simplicity, Ψ_1 as Ψ. We apply Theorem 6 of [157], according to which it suffices to show that there are $p \in (2,\infty)$, $c_p \in (0,\infty)$, and a function $\hat{\omega} : [0,\infty) \to [0,\infty)$ satisfying

$$\int_0^1 \frac{\hat{\omega}(u)}{u^{1+\alpha+(d+1)/p}} du < \infty \tag{11.25}$$

such that for all $0 \leq t_1 < t_2 \leq T$ and $r_1, r_2 \in O$, one has

$$\sup_{f\in\mathscr{V},u\in\mathscr{U}} E\|\Psi(t_2,r_2) - \Psi(t_1,r_1)\|^p \leq c_p \left(\hat{\omega}(\rho((t_1,r_1),(t_2,r_2)))\right)^p. \tag{11.26}$$

We will show that (11.26) holds with $\hat{\omega}(u) = u^{\alpha_0}$ for some $\alpha_0 \in (\alpha, \bar{\alpha})$ and all p sufficiently large. With such choices (and p large enough), the integrand in (11.25) will be of the form u^{β} with $\beta \in (-1, 0)$. This will establish the result.

Fix α_1 such that $\alpha < \alpha_1 < \bar{\alpha}$ and let $t_1 < t_2, r_1, r_2 \in O$ and $p > 2$. We will need p to be sufficiently large, and the choice of p will be fixed in the course of the proof. By the Burkholder–Davis–Gundy inequality [Appendix D, (D.3)], there exists a constant c_1 such that

$$E \|\Psi(t_2, r_2) - \Psi(t_1, r_1)\|^p \leq c_1 E \left[\int_0^T \int_O |G(t_2, s, r_2, q) - G(t_1, s, r_1, q)|^2 \|f(s, q)\|^2 dq\, ds \right]^{\frac{p}{2}}. \quad (11.27)$$

Let $\tilde{p} = p/(p-2)$ and $\delta = 4/p$. Note that $(2-\delta)\tilde{p} = \delta p/2 = 2$. Hölder's inequality (with parameters $p/(p-2)$ and $p/2$) and (11.24) give that the right-hand side of (11.27) is bounded above by

$$\begin{aligned} & c_1 \left[\int_0^T \int_O |G(t_2, s, r_2, q) - G(t_1, s, r_1, q)|^{(2-\delta)\tilde{p}}\, dq\, ds \right]^{\frac{p-2}{2}} \\ & \quad \times \left[\int_0^T \int_O |G(t_2, s, r_2, q) - G(t_1, s, r_1, q)|^{\delta p/2} E\|f(s, q)\|^p dq\, ds \right] \\ & \leq c_2 \left[\int_0^T \int_O |G(t_2, s, r_2, q) - G(t_1, s, r_1, q)|^2\, dq\, ds \right]^{\frac{p}{2}} \end{aligned} \quad (11.28)$$

for a suitable constant c_2 that is independent of f. From part (d) of Remark 11.19, the expression in (11.28) can be bounded (for p large enough) by

$$c_3 \rho\left((t_1, r_1), (t_2, r_2)\right)^{\alpha_1 p}.$$

The result follows. □

The next lemma will be used to prove that the stochastic integral appearing in $\bar{X}_x^{\varepsilon}$ converges to 0 in $\mathscr{C}([0, T] \times O)$, a result that will be strengthened shortly.

Lemma 11.29 *Let $\mathscr{V}$ and Ψ_1 be as in Lemma 11.28, and for $f \in \mathscr{V}$, let $Z_f^{\varepsilon} \doteq \sqrt{\varepsilon}\, \Psi_1$. Then for every sequence $\{f_\varepsilon\} \subset \mathscr{V}$, $Z_{f_\varepsilon}^{\varepsilon} \to 0$ in $\mathscr{C}([0, T] \times O)$ and in probability as $\varepsilon \to 0$.*

Proof Note that for $t \in [0, T]$ and $r \in O$,

$$\begin{aligned} \sup_{f \in \mathscr{V}} E|\Psi_1(t, r)|^2 &= \sup_{f \in \mathscr{V}} \int_0^t \int_O |G(t, s, r, q)|^2 E\|f(s, q)\|^2\, dq ds \\ &\leq c_1 \int_0^t \int_O |G(t, s, r, q)|^2 dq ds \\ &< \infty, \end{aligned}$$

where the last inequality is from (11.21). This shows that for such (t, r), $Z^{\varepsilon}_{f_\varepsilon}(t, r) \to 0$ in $\mathscr{L}^2$ and hence in probability. For $\delta \in (0, 1)$ and $x \in \mathscr{C}([0, T] \times O)$, define

$$\omega(x, \delta) \doteq \sup\left[\|x(t, r) - x(t', r')\| : \rho\left((t, r), (t', r')\right) \le \delta\right].$$

Then $\omega(Z^{\varepsilon}_{f_\varepsilon}, \delta) \le \sqrt{\varepsilon}\delta^{\alpha} M^{\varepsilon}_{f_\varepsilon}$, where

$$M^{\varepsilon}_{f_\varepsilon} \doteq \sup_{0<\rho((t,r),(s,q))<1} \frac{\|\Psi_1(t, r) - \Psi_1(s, q)\|}{\rho\left((t, r), (s, q)\right)^{\alpha}}.$$

Since $\alpha < \bar{\alpha}$, it follows by Lemma 11.28 that

$$\lim_{\delta\to 0} \sup_{\varepsilon\in(0,1)} E\omega(Z^{\varepsilon}_{f_\varepsilon}, \delta) = 0.$$

This establishes a form of uniform equicontinuity, and the result now follows from Theorem 14.5 of [167]. □

We now establish the main convergence result.

Proof (*of Theorem* 11.25) Consider sequences $\{x^{\varepsilon}\}$ and $\{u^{\varepsilon}\}$ as in the statement of Theorem 11.25. Letting $\bar{X}^{\varepsilon}_{x^\varepsilon}$ denote the corresponding controlled process, define

$$\begin{aligned}
Z^{\varepsilon}_1(t, r) &\doteq \int_O G(t, 0, r, q)x^{\varepsilon}(q)dq,\\
Z^{\varepsilon}_2(t, r) &\doteq \int_0^t \int_O G(t, s, r, q)R(s, q, \bar{X}^{\varepsilon}_{x^\varepsilon}(s, q))dqds,\\
Z^{\varepsilon}_3(t, r) &\doteq \sqrt{\varepsilon}\int_0^t \int_O G(t, s, r, q)A(s, q, \bar{X}^{\varepsilon}_{x^\varepsilon}(s, q))B(dq \times ds),\\
Z^{\varepsilon}_4(t, r) &\doteq \int_0^t \int_O G(t, s, r, q)A(s, q, \bar{X}^{\varepsilon}_{x^\varepsilon}(s, q))u^{\varepsilon}(s, q)dqds.
\end{aligned}$$

We first show that $\{Z^{\varepsilon}_i\}$ is tight in $\mathscr{C}([0, T] : \mathbb{B}_\alpha)$, for $i = 1, 2, 3, 4$. For $i = 1$, this follows from part (b) of Condition 11.18. Recall that the norm on $\mathbb{B}^T_\alpha$ is

$$\|\psi\|_{\alpha,T} \doteq \|\psi\|_{0,T} + \sup_{s,t\in[0,T],r,q\in O,s\neq t,r\neq q} \frac{|\psi(t, r) - \psi(s, q)|}{\rho\left((t, r), (s, q)\right)^{\alpha}},$$

with $\|\psi\|_{0,T} \doteq \sup_{t\in[0,T],q\in O} |\psi(t, r)|$. Since $\mathbb{B}^T_{\alpha^*}$ is compactly embedded in $\mathbb{B}^T_\alpha$ for $\bar{\alpha} > \alpha^* > \alpha$ (cf. [147, Lemma 6.33]), it suffices to show that for some $\alpha^* \in (\alpha, \bar{\alpha})$,

$$\sup_{\varepsilon\in(0,1)} P\left\{\|Z^{\varepsilon}_i\|_{\alpha^*,T} > K\right\} \to 0 \text{ as } K \to \infty \text{ for } i = 2, 3, 4. \tag{11.29}$$

For $i = 2, 4$, (11.29) is an immediate consequence of

$$\sup_{\varepsilon\in(0,1)} E\|Z_i^\varepsilon\|_{\alpha^*,T} < \infty,$$

which follows from Lemma 11.28, the linear growth condition (11.19), and Lemma 11.27. For $i = 3$, in view of Lemma 11.29, it suffices to establish

$$\sup_{\varepsilon\in(0,1)} E[Z_3^\varepsilon]_{\alpha^*,T} < \infty, \tag{11.30}$$

where for $z \in \mathbb{B}_\alpha^T$, $[z]_{\alpha^*,T} = \|z\|_{\alpha^*,T} - \|z\|_{0,T}$. From the linear growth condition (11.19) and Lemma 11.27, it follows that

$$\sup_{\varepsilon\in[0,1)} \sup_{(t,r)\in[0,T]\times O} E|A(t,r,\bar{X}^\varepsilon_{x^\varepsilon}(t,r))|^p < \infty.$$

The bound in (11.30) now follows on using Lemma 11.28, with $d = 1$ and

$$\mathscr{V} \doteq \{(t,r) \mapsto A(t,r,\bar{X}^\varepsilon_{x^\varepsilon}(t,r)),\ \varepsilon \in (0,1)\}.$$

Having shown tightness of Z_i^ε for $i = 1,2,3,4$, we can extract a subsequence along which each of these processes and also $\bar{X}^\varepsilon_{x^\varepsilon}$ jointly converge in distribution, with $\bar{X}^\varepsilon_{x^\varepsilon}$ taking values in $\mathscr{C}([0,T] : \mathbb{B}_\alpha)$. Let Z_i and $\bar{X}_x$ denote the respective limits. We will show that

$$\begin{aligned}
Z_1(t,r) &= \int_O G(t,0,r,q)x(q)dq,\\
Z_2(t,r) &= \int_0^t \int_O G(t,s,r,q)R(s,q,\bar{X}_x(s,q))dqds,\\
Z_3(t,r) &= 0,\\
Z_4(t,r) &= \int_0^t \int_O G(t,s,r,q)A(s,q,\bar{X}_x(s,q))u(s,q)dqds.
\end{aligned} \tag{11.31}$$

The uniqueness result Theorem 11.24 will then complete the proof.

Convergence for $i = 1$ follows from part (b) of Condition 11.18. The case $i = 3$ follows from Lemmas 11.29, 11.27, and the linear growth condition. To deal with the cases $i = 2,4$, we invoke the Skorohod representation theorem, which allows us to assume with probability one convergence for the purposes of identifying the limits. We give the proof of convergence only for the harder case $i = 4$. Denote the right side of (11.31) by $\hat{Z}_4(t,r)$. We have the bound

$$\begin{aligned}
&\left|Z_4^\varepsilon(t,r) - \hat{Z}_4(t,r)\right|\\
&\quad\le \int_0^t \int_O |G(t,s,r,q)| \left|A(s,q,\bar{X}^\varepsilon_{x^\varepsilon}(s,q)) - A(s,q,\bar{X}_x(s,q))\right| |u^\varepsilon(s,q)|\, dqds
\end{aligned}$$

$$+\left|\int_0^t\int_O G(t,s,r,q)A(s,q,\bar{X}_x(s,q))\,(u^\varepsilon(s,q)-u(s,q))\,dqds\right|. \quad (11.32)$$

Using the Cauchy-Schwarz inequality and the uniform Lipschitz property of A, for a suitable constant $c \in (0,\infty)$ the first term on the right side of (11.32) can be bounded above by

$$\sqrt{M}\left[\int_0^t\int_O |G(t,s,r,q)|^2\left|A(s,q,\bar{X}^\varepsilon_{x^\varepsilon}(s,q))-A(s,q,\bar{X}_x(s,q))\right|^2 dqds\right]^{1/2}$$
$$\le c\left(\sup_{(s,q)\in[0,T]\times O}\left\|\bar{X}^\varepsilon_{x^\varepsilon}(s,q)-\bar{X}_x(s,q)\right\|\right),$$

and thus it converges to 0 as $\varepsilon \to 0$. The second term in (11.32) converges to 0 as well, since $u^\varepsilon \to u$ as elements of $\bar{\mathscr{I}}_{b,M}$ and by (11.20) and the linear growth assumption (11.19),

$$\int_0^t\int_O |G(t,s,r,q)|^2\left|A(s,q,\bar{X}_x(s,q))\right|^2 dqds < \infty.$$

By uniqueness of limits and noting that $\hat{Z}_4$ is a continuous random field, we see that $Z_4 = \hat{Z}_4$, and the proof is complete. □

11.4 Notes

Some general references for stochastic partial differential equations are [169, 175, 221, 243]. The material of this chapter is largely taken from [43]. The approach taken is different from that of [170, 235] and other early works on large deviations for SPDE [50, 52, 56, 60, 127, 139, 160, 209, 252, 261]. The arguments used in these papers, which build on the ideas of [7], proceed by approximating the original model by a suitable time and/or space discretization. First one establishes an LDP for the approximate system and then argues that an LDP continues to hold as one approaches the original infinite dimensional model. For the last step, one needs suitable exponential probability estimates. These are usually the most technical aspects of the proofs, and they often assume conditions stronger than those needed for the LDP. Examples of various models to which the approach has been applied can be found in the references listed at the beginning of Part III of this book.

An alternative approach, based on nonlinear semigroup theory and infinite dimensional Hamilton–Jacobi–Bellman (HJB) equations, has been developed in [131, 132]. This approach relies on a uniqueness result for the corresponding infinite dimensional nonlinear PDEs. The uniqueness requirement on the limit HJB equation is an extraneous artifact of the approach, and different models seem to require different methods for this, in general very hard, uniqueness problem. In contrast to the weak

convergence approach, it requires an analysis of the model that goes significantly beyond the unique solvability of the SPDE.

One of the main reasons for proving a sample path LDP for any given stochastic system is as a step to validating the corresponding Freidlin–Wentzell large-time theory [140]. A key distinction between the cases of finite and infinite dimensional state is that open neighborhoods of points, which have compact closure in the finite dimensional case, are merely bounded in the infinite dimensional case. This means, unfortunately, that one should prove that the large deviation estimates are uniform for initial conditions in bounded sets in the latter case. As was discussed in Chap. 1, it is usually easy to establish a Laplace principle that is uniform with respect to initial conditions in compact sets using a straightforward argument by contradiction, which then gives the corresponding uniform LDP (Proposition 1.14). A different approach is needed for the infinite dimensional problem if one wants an LDP that is uniform over bounded sets, and one way to deal with the issue within the Laplace principle formalism is presented in [227].

The paper [38] studies large deviations for reaction–diffusion SPDE driven by a Poisson noise using representations for Poisson random measures of the form presented in Chap. 8.

Chapter 12
Stochastic Flows of Diffeomorphisms and Image Matching

The previous chapter considered in detail an example driven by a Brownian sheet, namely a stochastic reaction–diffusion equation. In this chapter we consider an application of one of the other formulations of infinite dimensional Brownian motion, which is the infinite sequence of independent one-dimensional Brownian motions. Such a collection will be used to define a general class of Brownian flows of diffeomorphisms [178], which are a special case of the stochastic flows of diffeomorphisms studied in [16, 25, 124, 178]. We will consider small noise asymptotics, prove the corresponding LDP, and then use it to give a Bayesian interpretation of an estimator used for image matching.

Elementary examples of Brownian flows are those constructed by solving finite dimensional Itô stochastic differential equations. More precisely, suppose b, $f_i, i = 1, \ldots, m$, are functions from $\mathbb{R}^d \times [0, T]$ to $\mathbb{R}^d$ that are continuous in (x, t) and $(k + 1)$-times continuously differentiable (with uniformly bounded derivatives) in x. Let $(\Omega, \mathscr{F}, P)$ be a probability space with a filtration $\{\mathscr{F}_t\}$ and let $\beta_1, \ldots, \beta_m$ be independent standard $\{\mathscr{F}_t\}$-Brownian motions. Then for each $s \in [0, T]$ and $x \in \mathbb{R}^d$, there is a unique continuous $\{\mathscr{F}_t\}$-adapted, $\mathbb{R}^d$-valued process $\phi_{s,t}(x)$, $s \le t \le T$, satisfying

$$\phi_{s,t}(x) = x + \int_s^t b(\phi_{s,r}(x), r)dr + \sum_{i=1}^{m} \int_s^t f_i(\phi_{s,r}(x), r)d\beta_i(r). \tag{12.1}$$

By choosing a suitable modification, $\{\phi_{s,t}\}_{0\le s\le t\le T}$ defines a Brownian flow of $\mathscr{C}^k$-diffeomorphisms (see Sect. 12.1). In particular, letting $\mathscr{D}^k$ denote the topological group of $\mathscr{C}^k$-diffeomorphisms (see Sect. 12.2 for precise definitions of the topology and the metric on $\mathscr{D}^k$), one has that $\phi = \{\phi_{0,t}\}_{0\le t\le T}$ is a random variable with values

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_12

in the Polish space $\hat{\mathscr{W}}_k = \mathscr{C}([0,T] : \mathscr{D}^k)$.[1] For $\varepsilon \in (0,\infty)$, when f_i is replaced by $\sqrt{\varepsilon} f_i$ in (12.1), we write the corresponding flow as ϕ^ε. Large deviations for ϕ^ε in $\hat{\mathscr{W}}_k$, as $\varepsilon \to 0$, have been studied for the case $k = 0$ in [10, 202] and for general k in [5].

As is well known (cf. [16, 163, 178]), not all Brownian flows can be expressed as in (12.1), and in general one needs infinitely many Brownian motions to obtain a stochastic differential equation (SDE) representation for the flow. Indeed, correlation structures (in the spatial parameter) that one is likely to encounter in applications generically lead to a formulation with infinitely many Brownian motions. One such example is given in Sect. 12.4. Thus, following Kunita's notation [178] for stochastic integration with respect to semimartingales with a spatial parameter, the study of general Brownian flows of $\mathscr{C}^k$-diffeomorphisms leads to SDEs of the form

$$d\phi_{s,t}(x) = \Phi(\phi_{s,t}(x), dt), \;\; \phi_{s,s}(x) = x, \;\; 0 \le s \le t \le T, \;\; x \in \mathbb{R}^d, \tag{12.2}$$

where $\Phi(x,t)$ is a $\mathscr{C}^{k+1}$-Brownian motion (see Definition 12.3). Note that Φ can be regarded as a random variable with values in the Polish space $\mathscr{W}_k = \mathscr{C}([0,T] : \mathscr{C}^{k+1}(\mathbb{R}^d))$, where $\mathscr{C}^{k+1}(\mathbb{R}^d)$ is the space of $(k+1)$ times continuously differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}^d$. Representations of such Brownian motions in terms of infinitely many independent standard real Brownian motions is well known (see, e.g., Kunita [178, Exercise 3.2.10]). Indeed, one can represent Φ as

$$\Phi(x,t) \doteq \int_0^t b(x,r)dr + \sum_{i=1}^{\infty} \int_0^t f_i(x,r)d\beta_i(r), \;\; (x,t) \in \mathbb{R}^d \times [0,T], \tag{12.3}$$

where $\{\beta_i\}_{i=1}^{\infty}$ is an infinite sequence of iid real Brownian motions and b, f_i are suitable functions from $\mathbb{R}^d \times [0,T]$ to $\mathbb{R}^d$ (see below Definition 12.3 for details).

Letting $a(x,y,t) = \sum_{i=1}^{\infty} f_i(x,t) f_i^T(y,t)$ for $x, y \in \mathbb{R}^d$ and $t \in [0,T]$, the functions (a,b) are referred to as the **local characteristics** of the Brownian motion Φ. When Eq. (12.2) is driven by the Brownian motion Φ^ε with local characteristics $(\varepsilon a, b)$, we denote the corresponding solution by ϕ^ε. In this chapter we establish a large deviation principle for $(\phi^\varepsilon, \Phi^\varepsilon)$ in $\hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1}$. Note that the LDP is established in a larger space than the one in which $(\phi^\varepsilon, \Phi^\varepsilon)$ take values (namely, $\hat{\mathscr{W}}_k \times \mathscr{W}_k$). This is consistent with results in [5, 10, 202], which consider stochastic flows driven by only finitely many real Brownian motions. The main technical difficulty in establishing the LDP in $\hat{\mathscr{W}}_k \times \mathscr{W}_k$ is the proof of a result analogous to Proposition 12.18, which establishes tightness of certain controlled processes, when $k-1$ is replaced by k.

In Sect. 12.4 we study an application of these results to a problem in image analysis. Stochastic diffeomorphic flows have been suggested for modeling *prior* statistical distributions on the space of possible images/targets of interest in the study of nonlinear inverse problems in this field (see [99] and references therein). Along with

[1] Although elsewhere in the book $\mathscr{D}$ is used for a Skorohod space such as $\mathscr{D}([0,T] : \mathbb{R}^d)$, in this chapter only it is used for such spaces of diffeomorphisms.

a data model, noise-corrupted observations together with such a prior distribution can then be used to compute a *posterior* distribution on this space, from which it is possible to construct various estimators of the true image underlying the observations, such as an approximate maximum likelihood estimator. Motivated by such a Bayesian framework, a variational approach to the image-matching problem has been suggested and analyzed in [99]. Here we develop a rigorous asymptotic theory that relates a standard Bayesian formulation of the problem, in the small noise limit, to the deterministic variational approach taken in [99]. This is done in Theorem 12.20 of Sect. 12.4.

We now give an outline of the chapter. Section 12.1 contains notation needed for this chapter as well as the definitions of $\mathscr{C}^k$-Brownian motion and Brownian flow. Section 12.2 presents the main large deviation result. The weak convergence needed to prove this result, Theorem 12.8, is established in Sect. 12.3. We would like to highlight the fact that apart from the much greater level of detail required due to the complexity of the state space, the argument here is very much the same as that used to establish the finite dimensional results in Sect. 3.2.1. We also note that as in the reaction–diffusion example of Chap. 11, we do not have to start from scratch, and in fact, the analysis of the controlled systems appearing in the representation will borrow much from Kunita's corresponding qualitative analysis of the original system as presented in [178]. Finally, Sect. 12.4 introduces the image analysis problem and uses the results of Sect. 12.2 to obtain an asymptotic result relating the Bayesian formulation of the problem to the deterministic variational approach of [99].

12.1 Notation and Definitions

There is a great deal of specialized notation associated with the study of stochastic flows of diffeomorphisms. In order to minimize the notational conflict with standard references in the area such as [178], we adopt much of this notation for this chapter only. In particular, the following list gives notation that is specific to this chapter and that may differ from the notation used for the same objects elsewhere in the book. This notation generally follows [178].

- Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ be a multi-index of nonnegative integers and $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$. For an $|\alpha|$-times differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, set $\partial^\alpha f \doteq \partial_x^\alpha f = \frac{\partial^{|\alpha|} f}{(\partial x_1)^{\alpha_1} \cdots (\partial x_d)^{\alpha_d}}$. For such an f, we write $\frac{\partial f(x)}{\partial x_i}$ as $\partial_i f$. If $f = (f_1, f_2, \ldots, f_d)^T$ is an $|\alpha|$-times differentiable function from $\mathbb{R}^d$ to $\mathbb{R}^d$, we write $\partial^\alpha f \doteq (\partial^\alpha f_1, \partial^\alpha f_2, \ldots, \partial^\alpha f_d)^T$. By convention, $\partial^0 f = f$.
- For $m \in \mathbb{N}_0$, let $\mathscr{C}^m$ denote the space of m-times continuously differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$.
- For any subset $A \subset \mathbb{R}^d$, $m \in \mathbb{N}_0$, and $f \in \mathscr{C}^m$, let

$$\|f\|_{m;A} \doteq \sum_{0 \le |\alpha| \le m} \sup_{x \in A} |\partial^\alpha f(x)|.$$

The space $\mathscr{C}^m$ is a Fréchet space [226, Sect. 1.8] with the countable collection of seminorms $\|f\|_{m;A_n}$, $A_n \doteq \{x : \|x\| \le n\}$ [178]. In particular, it is a Polish space.

- For $0 < \delta \le 1$, let

$$\|f\|_{m,\delta;A} \doteq \|f\|_{m;A} + \sum_{|\alpha|=m} \sup_{x,y\in A; x\neq y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x-y\|^\delta}$$

and

$$\mathscr{C}^{m,\delta} \doteq \left\{ f \in \mathscr{C}^m : \|f\|_{m,\delta;A_n} < \infty \text{ for any } n \in \mathbb{N} \right\}.$$

The seminorms $\{\|\cdot\|_{m,\delta;A_n}, n \in \mathbb{N}\}$ make $\mathscr{C}^{m,\delta}$ a Fréchet space.

- For $m \in \mathbb{N}_0$ let $\widetilde{\mathscr{C}}^m$ denote the space of functions $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $g(x, y)$, $x, y \in \mathbb{R}^d$ is m-times continuously differentiable with respect to both x and y. Endowed with the seminorms

$$\|g\|^{\sim}_{m;A_n} \doteq \sum_{0\le|\alpha|\le m} \sup_{x,y\in A_n} |\partial_x^\alpha \partial_y^\alpha g(x, y)|,$$

where $n \in \mathbb{N}$, $\widetilde{\mathscr{C}}^m$ is a Fréchet space. Also, for $0 < \delta \le 1$, let

$$\|g\|^{\sim}_{m,\delta;A_n} \doteq \|g\|^{\sim}_{m;A_n} + \sum_{|\alpha|=m} \sup_{\substack{x\neq\bar{x}, y\neq\bar{y}\\ x,y,\bar{x},\bar{y}\in A_n}} \frac{|\Delta^\alpha_{x,\bar{x}} g(y) - \Delta^\alpha_{x,\bar{x}} g(\bar{y})|}{\|x-\bar{x}\|^\delta \, \|y-\bar{y}\|^\delta},$$

where $\Delta^\alpha_{x,\bar{x}} g(y) \doteq \hat{\partial}^\alpha_{x,y} g(x, y) - \hat{\partial}^\alpha_{\bar{x},y} g(\bar{x}, y)$, $\hat{\partial}^\alpha_{x,y} g(x, y) \doteq \partial_x^\alpha \partial_y^\alpha g(x, y)$. Then

$$\widetilde{\mathscr{C}}^{m,\delta} \doteq \left\{ g \in \widetilde{\mathscr{C}}^m; \|g\|^{\sim}_{m,\delta;A_n} < \infty, \text{ for every } n \in \mathbb{N} \right\}$$

is a Fréchet space with respect to the seminorms $\{\|\cdot\|^{\sim}_{m,\delta;A_n}, n \in \mathbb{N}\}$.

- We write $\|f\|_{m;\mathbb{R}^d}$ as $\|f\|_m$. The norms $\|\cdot\|_{m,\delta}$, $\|\cdot\|^{\sim}_m$, $\|\cdot\|^{\sim}_{m,\delta}$ are to be interpreted in a similar manner.
- Let $\mathscr{C}^m(\mathbb{R}^d) \doteq \left\{ f = (f_1, \ldots, f_d)^T : f_i \in \mathscr{C}^m, i = 1, \ldots, d \right\}$ and with some abuse of notation $\|f\|_{m,A} \doteq \sum_{i=1}^d \|f_i\|_{m,A}$. Spaces such as $\mathscr{C}^{m,\delta}(\mathbb{R}^d)$, $\widetilde{\mathscr{C}}^m(\mathbb{R}^{d\times d})$ and $\widetilde{\mathscr{C}}^{m,\delta}(\mathbb{R}^{d\times d})$ and their corresponding seminorms are defined similarly. Note that here the argument indicates the range space, and that in particular, $h \in \widetilde{\mathscr{C}}^{m,\delta}(\mathbb{R}^{d\times d})$ is a map from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathbb{R}^{d\times d}$.
- Let $\mathscr{C}^{m,\delta}_T(\mathbb{R}^d)$ and $\widetilde{\mathscr{C}}^{m,\delta}_T(\mathbb{R}^{d\times d})$ be the classes of measurable functions $b : [0, T] \to \mathscr{C}^{m,\delta}(\mathbb{R}^d)$ and $a : [0, T] \to \widetilde{\mathscr{C}}^{m,\delta}(\mathbb{R}^{d\times d})$, respectively, such that

$$\|b\|_{T,m,\delta} \doteq \sup_{0\le t\le T} \|b(t)\|_{m,\delta} < \infty \text{ and } \|a\|^{\sim}_{T,m,\delta} \doteq \sup_{0\le t\le T} \|a(t)\|^{\sim}_{m,\delta} < \infty.$$

For convenience, we also recall the following notation from Chap. 11. The Hilbert space l_2 is defined by

$$l_2 \doteq \left\{ (x_1, x_2, \ldots) : \ x_i \in \mathbb{R}, \ i \in \mathbb{N} \text{ and } \sum_{i=1}^{\infty} x_i^2 < \infty \right\},$$

where the inner product on l_2 is $\langle x, y \rangle_0 \doteq \sum_{i=1}^{\infty} x_i y_i$, $x, y \in l_2$. We denote the corresponding norm by $\| \cdot \|_0$. Given a probability space $(\Omega, \mathscr{F}, P)$ and a filtration $\{\mathscr{F}_t\}$ satisfying the usual conditions, we recall that

$$\bar{\mathscr{A}} \doteq \left\{ \phi = \{\phi_i\}_{i=1}^{\infty} \,\middle|\, \phi_i : [0,T] \to \mathbb{R} \text{ is } \{\mathscr{F}_t\}\text{-predictable for all } i \right.$$
$$\left. \text{and } P\left\{ \int_0^T \|\phi(s)\|_0^2 ds < \infty \right\} = 1 \right\}.$$

Recall also the definition

$$S_N \doteq \left\{ \phi = \{\phi_i\}_{i=1}^{\infty} \in \mathscr{L}^2([0,T] : l^2) \text{ such that } \int_0^T \|\phi(s)\|_0^2 ds \le N \right\},$$

and that when equipped with the weak topology for the Hilbert space $\mathscr{L}^2([0,T] : l^2)$, S_N is a compact Polish space. Lastly, recall

$$\bar{\mathscr{A}}_{b,N} \doteq \left\{ u \in \bar{\mathscr{A}} : u(\omega) \in S_N, \ P\text{-a.s.} \right\} \text{ and } \bar{\mathscr{A}}_b \doteq \cup_{N \in \mathbb{N}} \bar{\mathscr{A}}_{b,N}. \tag{12.4}$$

When referring to convergence in distribution of S_N-valued random variables, we always consider S_N with the weak topology.

We next give definitions that are particular to this chapter. Let $\circ$ denote the composition of maps and let ι denote the identity map on $\mathbb{R}^d$.

Definition 12.1 A collection $\left\{ \phi_{s,t}(x) : 0 \le s \le t \le T, x \in \mathbb{R}^d \right\}$ of $\mathbb{R}^d$-valued random variables on some $(\Omega, \mathscr{F}, P, \{\mathscr{F}_t\})$ is called a **forward stochastic flow of homeomorphisms** if there exists $N \in \mathscr{F}$, with $P(N) = 0$, such that for all $\omega \in N^c$:

(a) $(s,t,x) \mapsto \phi_{s,t}(x,\omega)$ is continuous;

(b) $\phi_{s,u}(\omega) = \phi_{t,u}(\omega) \circ \phi_{s,t}(\omega)$ holds for all s, t, u, $0 \le s \le t \le u \le T$;

(c) $\phi_{s,s}(\omega) = \iota$ for all s, $0 \le s \le T$;

(d) the map $\phi_{s,t}(\omega) : \mathbb{R}^d \to \mathbb{R}^d$ is an onto homeomorphism for all s, t, $0 \le s \le t \le T$.

If in addition $\phi_{s,t}(x,\omega)$ is k-times differentiable with respect to x for all $s \le t$ and the derivatives are continuous in (s,t,x), it is called a **stochastic flow of $\mathscr{C}^k$-diffeomorphisms**.

We now introduce a Brownian motion with a spatial parameter, with local characteristics (a, b). Fix $k \in \mathbb{N}$ and $\delta \in (0, 1]$. Throughout this chapter, we will assume the following.

Condition 12.2 *The coefficients satisfy* $(a, b) \in \widetilde{\mathscr{C}}_T^{k,\delta}(\mathbb{R}^{d\times d}) \times \mathscr{C}_T^{k,\delta}(\mathbb{R}^d)$.

Fix ν such that $0 < \nu < \delta$.

Definition 12.3 A continuous stochastic process $\{\Phi(t)\}_{t\geq 0}$ on $(\Omega, \mathscr{F}, P, \{\mathscr{F}_t\})$ with values in $\mathscr{C}^{k,\nu}(\mathbb{R}^d)$ is said to be a $\mathscr{C}^{k,\nu}$**-Brownian motion with local characteristics** (a, b) if $\Phi(0)$, $\Phi(t_{i+1}) - \Phi(t_i)$, $i = 0, 1, \ldots, n-1$, are independent $\mathscr{C}^{k,\nu}(\mathbb{R}^d)$-valued random variables whenever $0 \leq t_0 < t_1 < \cdots < t_n \leq T$, and if for each $x \in \mathbb{R}^d$, $M(x,t) \doteq \Phi(x,t) - \int_0^t b(x,r)dr$ is a continuous (d-dimensional) martingale such that the matrix quadratic variation process (see Sect. D.1) satisfies $\langle\langle M(x,\cdot), M(y,\cdot)\rangle\rangle_t = \int_0^t a(x,y,r)dr$ for all $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$.

The existence of a $\mathscr{C}^{k,\nu}$-Brownian motion with local characteristics (a, b) follows from [178] (see, e.g., Theorem 3.1.2 and Exercise 3.2.10). Indeed, for every $\gamma < \delta$, one can represent Φ as in (12.3), where $f_i : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ are such that for each $t \in [0,T]$, $f_i(\cdot,t) \in \mathscr{C}^{k,\gamma}(\mathbb{R}^d)$,

$$a(x,y,t) = \sum_{i=1}^{\infty} f_i(x,t) f_i^T(y,t), \text{ a.e. } t, \tag{12.5}$$

and

$$\int_0^T \sum_{i=1}^{\infty} |f_i(x,r)|^2 dr \leq T\|a\|_{T,k,\delta}^{\sim} < \infty.$$

In particular, note that if Φ is a $\mathscr{C}^{k,\nu}$-valued Brownian motion, its finite dimensional restriction $(\Phi(x_1,\cdot), \Phi(x_2,\cdot), \ldots, \Phi(x_n,\cdot))^T$ is an nd-dimensional Gaussian process with independent increments for all $(x_1, \ldots, x_n) \in \mathbb{R}^{nd}$. If Φ is as defined by (12.3) and $\{\phi_t\}_{0\leq t\leq 1}$ is a continuous $\mathbb{R}^d$-valued $\{\mathscr{F}_t\}$-adapted stochastic process, then the stochastic integral $\int_0^t \Phi(\phi_r, dr)$ is a well-defined d-dimensional continuous $\{\mathscr{F}_t\}$-adapted stochastic process (see Chap. 3, Sect. 2, pages 71–86 of [178]).

Definition 12.4 Let Φ be as in Definition 12.3. Then for each $s \in [0,T]$ and $x \in \mathbb{R}^d$, there is a unique continuous $\mathscr{F}_t$-adapted, $\mathbb{R}^d$-valued process $\phi_{s,t}(x)$, $s \leq t \leq T$ satisfying $\phi_{s,t}(x) = x + \int_s^t \Phi(\phi_{s,r}(x), dr)$, $t \in [s,T]$. This stochastic process is called the solution of Itô's stochastic differential equation based on the Brownian motion Φ.

From Theorem D.5 it follows that $\{\phi_{s,t}\}_{0\leq s\leq t\leq T}$, as introduced in Definition 12.4, has a modification that is a forward stochastic flow of $\mathscr{C}^k$-diffeomorphisms.

12.2 Statement of the LDP

Given $\varepsilon > 0$, let Φ^ε be a $\mathscr{C}^{k,\nu}$-Brownian motion on $(\Omega, \mathscr{F}, P, \{\mathscr{F}_t\})$, with local characteristics $(\varepsilon a, b)$, where (a, b) satisfy Condition 12.2. Without loss of generality we assume that Φ^ε is represented as

$$\Phi^{\varepsilon}(x,t) \doteq \int_0^t b(x,r)dr + \sqrt{\varepsilon}\sum_{l=1}^{\infty}\int_0^t f_l(x,r)d\beta_l(r),\ (x,t) \in \mathbb{R}^d \times [0,T], \tag{12.6}$$

where $(\beta_l, f_l)_{l\in\mathbb{N}}$ are as in (12.3) and (12.5). Define the martingale $M(x,t)$ by

$$\sqrt{\varepsilon}M(x,t) \doteq \Phi^{\varepsilon}(x,t) - \int_0^t b(x,r)dr. \tag{12.7}$$

With an abuse of notation, when $\varepsilon = \varepsilon_n$, we write Φ^{ε} as Φ^n. Observe that

$$\langle\!\langle M(x,\cdot), \beta_l(\cdot)\rangle\!\rangle_t = \int_0^t f_l(x,r)dr \text{ for all } t \in [0,T],\ \text{a.s.}$$

Let $\phi^{\varepsilon} = \{\phi^{\varepsilon}_{s,t}(x)\}$ be the forward stochastic flow of $\mathscr{C}^k$-diffeomorphisms based on Φ^{ε}. With another abuse of notation, we write $\phi^{\varepsilon}_{0,t}$ as ϕ^{ε}_t and sometimes use ϕ^{ε} to denote $\{\phi^{\varepsilon}_t(x)\}_{0\le t\le T, x\in\mathbb{R}^d}$.

In this chapter we show that $\{(\phi^{\varepsilon}, \Phi^{\varepsilon})\}_{\varepsilon>0}$ satisfies a large deviation principle on two suitable function spaces. The appropriate spaces are defined as follows. For $m \in \mathbb{N}$, let $\mathscr{D}^m$ be the group of $\mathscr{C}^m$-diffeomorphisms on $\mathbb{R}^d$; $\mathscr{D}^m$ is endowed with the metric

$$d_m(\phi,\psi) \doteq \lambda_m(\phi,\psi) + \lambda_m(\phi^{-1},\psi^{-1}), \tag{12.8}$$

where

$$\lambda_m(\phi,\psi) \doteq \sum_{|\alpha|\le m} \rho(\partial^{\alpha}\phi, \partial^{\alpha}\psi), \tag{12.9}$$

$$\rho(\phi,\psi) \doteq \sum_{N=1}^{\infty} \frac{1}{2^N} \frac{\sup_{\|x\|\le N} |\phi(x) - \psi(x)|}{1 + \sup_{\|x\|\le N} |\phi(x) - \psi(x)|}.$$

Under this metric, $\mathscr{D}^m$ is a Polish space (see [178, Chap. 4]). Let $\hat{\mathscr{W}}_m \doteq \mathscr{C}\big([0,T] : \mathscr{D}^m\big)$ be the set of all continuous maps from $[0,T]$ to $\mathscr{D}^m$ and let $\mathscr{W}_m \doteq \mathscr{C}\big([0,T] : \mathscr{C}^m(\mathbb{R}^d)\big)$ be the set of all continuous maps from $[0,T]$ to $\mathscr{C}^m(\mathbb{R}^d)$. Hence being an element of $\mathscr{W}_m$ implies a smoothness in x, while being in $\hat{\mathscr{W}}_m$ says that the function is a diffeomorphism. The space $\hat{\mathscr{W}}_m$ endowed with the metric $\hat{d}_m(\phi,\psi) = \sup_{0\le t\le T} d_m(\phi(t),\psi(t))$ and the space $\mathscr{W}_m$ endowed with the metric $\bar{d}_m(\phi,\psi) = \sup_{0\le t\le T} \lambda_m(\phi(t),\psi(t))$ are both Polish spaces. Note that $(\phi^{\varepsilon}, \Phi^{\varepsilon})$ belongs to $\hat{\mathscr{W}}_k \times \mathscr{W}_k \subset \hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1} \subset \mathscr{W}_{k-1} \times \mathscr{W}_{k-1}$. We will show that the pair $\{(\phi^{\varepsilon}, \Phi^{\varepsilon})\}_{\varepsilon>0}$ satisfies LDPs in both of the spaces $\hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1}$ and $\mathscr{W}_{k-1} \times \mathscr{W}_{k-1}$, with a rate function I that is introduced below.

Let $u = \{u_l\}_{l=1}^{\infty} \in \bar{\mathscr{A}}_b$, and recall that this implies that there is $M < \infty$ such that w.p.1 $\sum_{l=1}^{\infty}\int_0^T |u_l(s)|^2\,ds \le M$. Given any such control, we want to construct a corresponding controlled flow in the form of a perturbed analogue of (12.6). To state the LDP, we need only consider the case $u \in \mathscr{L}^2([0,T] : l_2)$, but for later use, it is

convenient to consider the more general case in which u is allowed to be random. We first need to interpret this control as a drift. Observe that $Z_t \doteq \sum_{l=1}^{\infty} \int_0^t u_l(s) d\beta_l(s)$ is a continuous square-integrable martingale. For any $\gamma < \delta$ one can find $b_u : \mathbb{R}^d \times [0,T] \times \Omega \to \mathbb{R}^d$ such that $b_u(\cdot,t,\omega) \in \mathscr{C}^{k,\gamma}(\mathbb{R}^d)$ for a.e. (t,ω), such that for each $x \in \mathbb{R}^d$, $b_u(x,\cdot)$ is predictable, and such that $\int_0^t b_u(x,s)ds = \langle\langle Z, M(x,\cdot)\rangle\rangle_t$ for each $(x,t) \in \mathbb{R}^d \times [0,T]$. In particular, for each $x \in \mathbb{R}^d$, $b_u(x,t) \doteq \sum_{l=1}^{\infty} u_l(t) f_l(x,t)$ for almost every (t,ω). Furthermore, for some $c \in (0,\infty)$,

$$\|b_u(\cdot,t)\|^2_{k,\gamma} \le c\|a\|^{\sim}_{T,k,\delta} \sum_{l=1}^{\infty} |u_l(t)|^2, \ [dt \times P] - \text{a.e. in } (t,\omega). \tag{12.10}$$

The proofs of these statements follow along the lines of Exercise 3.2.10 and Lemma 3.2.3 of [178]. Next, define

$$\bar{\Phi}(x,t) \doteq \int_0^t b_u(x,s)ds + \int_0^t b(x,s)ds. \tag{12.11}$$

It follows that $\bar{\Phi}(\cdot,t)$ is a $\mathscr{C}^{k,\gamma}(\mathbb{R}^d)$-valued continuous adapted stochastic process. Let $\hat{b}_u \doteq b_u + b$, and for $(t_0,x) \in [0,T] \times \mathbb{R}^d$, let $\{\bar{\phi}_{t_0,t}(x)\}_{t_0 \le t \le T}$ be the unique solution of the equation

$$\bar{\phi}_{t_0,t}(x) \doteq x + \int_{t_0}^t \hat{b}_u\big(\bar{\phi}_{t_0,t}(x), r\big)dr, \ t \in [t_0,T]. \tag{12.12}$$

By Theorem D.5, $\{\bar{\phi}_{s,t}\}_{0 \le s \le t \le T}$ is a forward flow of $\mathscr{C}^k$-diffeomorphisms.

For $(\phi^0, \Phi^0) \in \hat{\mathscr{W}}_k \times \mathscr{W}_k$, define

$$I(\phi^0,\Phi^0) \doteq \inf_u \frac{1}{2} \int_0^T \|u(s)\|_0^2 ds, \tag{12.13}$$

where the infimum is taken over all u such that $(\phi^0,\Phi^0) = (\bar{\phi},\bar{\Phi})$, where $\bar{\phi}$ and $\bar{\Phi}$ are given by (12.12) and (12.11), respectively. If $(\phi^0,\Phi^0) \in (\mathscr{W}_{k-1} \times \mathscr{W}_{k-1}) \setminus (\hat{\mathscr{W}}_k \times \mathscr{W}_k)$, we set $I(\phi^0,\Phi^0) = \infty$. We denote the restriction of I to $\hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1}$ by the same symbol. The following is the main result of the section.

Theorem 12.5 *The family $(\phi^\varepsilon,\Phi^\varepsilon)_{\varepsilon>0}$ satisfies an LDP in the spaces $\hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1}$ and $\mathscr{W}_{k-1} \times \mathscr{W}_{k-1}$ with rate function I.*

Let $\{u^n\}_{n=1}^{\infty}$ (with $u^n = \{u_l^n\}_{l=1}^{\infty}$) be a sequence in $\bar{\mathscr{A}}_{b,N}$ for some fixed $N < \infty$. Let $\{\varepsilon_n\}_{n\in\mathbb{N}}$ be a sequence such that $\varepsilon_n \ge 0$ for each n and $\varepsilon_n \to 0$ as $n \to \infty$. Note that we allow $\varepsilon_n = 0$ for all n. Recall $M(x,t)$ from (12.7). Following our standard convention, controlled versions of processes are indicated by an overbar, without explicitly denoting the dependence on the control. Thus we define

$$\bar{\Phi}^n(x,t) \doteq \int_0^t \hat{b}_{u^n}(x,r)dr + \sqrt{\varepsilon_n}\, M(x,t) \tag{12.14}$$

and let $\bar{\phi}^n$ be the solution to

$$\bar{\phi}^n_t(x) = x + \int_0^t \hat{b}_{u^n}\big(\bar{\phi}^n_r(x), r\big)dr + \sqrt{\varepsilon_n}\int_0^t M\big(\bar{\phi}^n_r(x), dr\big). \tag{12.15}$$

Clearly $\bar{\Phi}^n \in \mathscr{W}_k$, and from Theorem D.5, Eq. (12.15) has a unique solution $\bar{\phi}^n$ that lies in $\hat{\mathscr{W}}_k$ a.s. We next introduce some basic weak convergence definitions.

Definition 12.6 Let $\{u^n\}_{n\in\mathbb{N}} \subset \bar{\mathscr{A}}_{b,N}$, $u \in \bar{\mathscr{A}}_{b,N}$, and let $(\bar{\phi}^n, \bar{\Phi}^n)$ and $(\bar{\phi}, \bar{\Phi})$ be defined by (12.14), (12.15) and (12.11), (12.12), respectively. Let $\hat{P}^n_{k-1}$, $\hat{P}_{k-1}$ be the measures induced by $(\bar{\phi}^n, \bar{\Phi}^n)$, $(\bar{\phi}, \bar{\Phi})$, respectively, on $\hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1}$. Thus for $A \in \mathscr{B}(\hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1})$, we have

$$\hat{P}^n_{k-1}(A) = P\big((\bar{\phi}^n, \bar{\Phi}^n) \in A\big), \quad \hat{P}_{k-1}(A) = P\big((\bar{\phi}, \bar{\Phi}) \in A\big).$$

The sequence $\big\{(\bar{\phi}^n, \bar{\Phi}^n)\big\}_{n\in\mathbb{N}}$ is said to **converge weakly as $\mathscr{D}^{k-1}$-flows** to $(\bar{\phi}, \bar{\Phi})$ as $n \to \infty$ if $\hat{P}^n_{k-1}$ converges weakly to $\hat{P}_{k-1}$ as $n \to \infty$.

Definition 12.7 Let $\{u^n\}_{n\in\mathbb{N}} \subset \bar{\mathscr{A}}_{b,N}$, $u \in \bar{\mathscr{A}}_{b,N}$, and let P^n_{k-1}, P_{k-1} be the measures induced by $(\bar{\phi}^n, \bar{\Phi}^n)$, $(\bar{\phi}, \bar{\Phi})$ respectively on $\mathscr{W}_{k-1} \times \mathscr{W}_{k-1}$. The sequence $\big\{(\bar{\phi}^n, \bar{\Phi}^n)\big\}_{n\in\mathbb{N}}$ is said to **converge weakly as $\mathscr{C}^{k-1}$-flows** to $(\bar{\phi}, \bar{\Phi})$ as $n \to \infty$ if P^n_{k-1} converges weakly to P_{k-1} as $n \to \infty$.

As is the case throughout the book, the proofs of large deviation properties essentially reduce to weak convergence questions for controlled analogues of the original process. For the present problem, the following theorem gives the needed result. The proof is given in the next section.

Theorem 12.8 *Let $u^n, u \in \bar{\mathscr{A}}_{b,N}$ be such that u^n converges to u in distribution, and define $(\bar{\phi}^n, \bar{\Phi}^n)$, $(\bar{\phi}, \bar{\Phi})$ as in (12.14), (12.15) and (12.11), (12.12). Then the sequence $\{(\bar{\phi}^n, \bar{\Phi}^n)\}_{n\in\mathbb{N}}$ converges weakly as $\mathscr{C}^{k-1}$-flows and $\mathscr{D}^{k-1}$-flows to $(\bar{\phi}, \bar{\Phi})$ as $n \to \infty$.*

Proof of Theorem 12.5 We will show only that the sequence $(\phi^\varepsilon, \Phi^\varepsilon)$ satisfies an LDP in $\hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1}$ with rate function I defined as in (12.13). The LDP in $\mathscr{W}_{k-1} \times \mathscr{W}_{k-1}$ follows similarly. We apply Theorem 11.13. Let $\mathscr{G}^\varepsilon : \mathscr{C}([0,T] : \mathbb{R}^\infty) \to \hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1}$ be a measurable map such that $\mathscr{G}^\varepsilon(\sqrt{\varepsilon}\beta) = (\phi^\varepsilon, \Phi^\varepsilon)$ a.s., where Φ^ε is given by (12.6) and ϕ^ε is the associated flow based on Φ^ε. The existence of such a map follows through an adaptation of the classical Yamada–Watanabe proof [159, Chap. IV] for finite dimensional diffusions with unique pathwise solutions. Define $\mathscr{G}^0 : \mathscr{C}([0,T] : \mathbb{R}^\infty) \to \hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1}$ by $\mathscr{G}^0\big(\int_0^\cdot u(s)ds\big) = (\bar{\phi}, \bar{\Phi})$ if $u \in \mathscr{L}^2([0,T] : l_2)$ and with $\bar{\phi}, \bar{\Phi}$ as defined in (12.12) and (12.11), respectively. Set $\mathscr{G}^0(f) = 0$ for all other

$f \in \mathscr{C}([0,T]:\mathbb{R}^\infty)$. We now verify Condition 11.12 with $\mathscr{G}^\varepsilon$ and $\mathscr{G}^0$ as just defined. In the present setting there is only one initial condition of interest. Hence $\mathscr{Z}$ is a single point, and so we verify the modified form of Condition 11.12 in which $\mathscr{G}^\varepsilon$ and $\mathscr{G}^0$ are maps defined on $\mathscr{C}([0,T]:\mathbb{R}^\infty)$ rather than on $\mathscr{Z} \times \mathscr{C}([0,T]:\mathbb{R}^\infty)$.

Fix $N < \infty$ and consider $\Gamma_N \doteq \{\mathscr{G}^0(\int_0^\cdot u(s)ds), u \in S_N\}$. We now show that Γ_N is a compact subset of $\hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1}$. If $u \mapsto \mathscr{G}^0(\int_0^\cdot u(s)ds)$ is continuous on S_N, then as the continuous forward image of a compact set, Γ_N is also compact. Hence it suffices to show that if $u^n, u \in S_N$ are such that $u^n \to u$, then $\mathscr{G}^0(\int_0^\cdot u^n(s)ds) \to \mathscr{G}^0(\int_0^\cdot u(s)ds)$ in $\hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1}$. This is immediate from Theorem 12.8 on noting that $\mathscr{G}^0(\int_0^\cdot u^n(s)ds) = (\bar{\phi}^n, \bar{\Phi}^n)$, where $\bar{\phi}^n$, $\bar{\Phi}^n$ are as in (12.15) and (12.14) respectively with $\varepsilon_n = 0$; and $\mathscr{G}^0(\int_0^\cdot u(s)ds) = (\bar{\phi}, \bar{\Phi})$, where $\bar{\phi}$, $\bar{\Phi}$ are as in (12.12) and (12.11) respectively. This verifies part (a) of Condition 11.12.

Next let $\{u^n\} \subset \bar{\mathscr{A}}_{b,N}$ and $\varepsilon_n \in (0,\infty)$ be such that $\varepsilon_n \to 0$ and u^n converges in distribution to some u as $n \to \infty$. We now show that $\mathscr{G}^{\varepsilon_n}(\sqrt{\varepsilon_n}\beta + \int_0^\cdot u^n(s)ds) \Rightarrow \mathscr{G}^0(\int_0^\cdot u(s)ds)$ in $\hat{\mathscr{W}}_{k-1} \times \mathscr{W}_{k-1}$ as $n \to \infty$. An application of Girsanov's theorem [see below Theorem D.1] shows that $\mathscr{G}^{\varepsilon_n}(\sqrt{\varepsilon_n}\beta + \int_0^\cdot u^n(s)ds) = (\bar{\phi}^n, \bar{\Phi}^n)$, where $\bar{\phi}^n, \bar{\Phi}^n$ are defined as in (12.15) and (12.14), respectively [see, for example, Sect. 10.2.1, where a similar argument for finite dimensional jump-diffusions was used]. Also, $\mathscr{G}^0(\int_0^\cdot u(s)ds) = (\bar{\phi}, \bar{\Phi})$, where $\bar{\phi}$, $\bar{\Phi}$ are as in (12.12) and (12.11), respectively. The desired convergence now follows from Theorem 12.8. This verifies part (b) of Condition 11.12, and Theorem 12.5 now follows from Theorem 11.13. □

12.3 Weak Convergence for Controlled Flows

This section will present the proof of Theorem 12.8. It is worth recalling assumptions that will be in effect for this section, which are that $\{u^n\}$ is converging to u in distribution as an S_N-valued sequence of random variables, where we recall that S_N is a compact Polish space under the weak topology, and that by Condition 12.2, $(a,b) \in \tilde{\mathscr{C}}_T^{k,\delta}(\mathbb{R}^{d\times d}) \times \mathscr{C}_T^{k,\delta}(\mathbb{R}^d)$, for some $k \in \mathbb{N}$ and $\delta \in (0,1]$.

We begin by introducing the (m,p)-point motion of the flow and the related notion of "convergence as diffusions." Let $\mathbf{x} \doteq (x_1, x_2, \ldots, x_m)$ and $\mathbf{y} \doteq (y_1, y_2, \ldots, y_p)$ be arbitrary fixed points in $\mathbb{R}^{d\times m}$ and $\mathbb{R}^{d\times p}$, respectively. Set

$$\bar{\phi}_t^n(\mathbf{x}) \doteq \big(\bar{\phi}_t^n(x_1), \bar{\phi}_t^n(x_2), \ldots, \bar{\phi}_t^n(x_m)\big)$$

and

$$\bar{\Phi}^n(\mathbf{y},t) \doteq \big(\bar{\Phi}^n(y_1,t), \bar{\Phi}^n(y_2,t), \ldots, \bar{\Phi}^n(y_p,t)\big).$$

Then the pair $\{\bar{\phi}_t^n(\mathbf{x}), \bar{\Phi}^n(\mathbf{y},t)\}$ is a continuous stochastic process with values in $\mathbb{R}^{d\times m} \times \mathbb{R}^{d\times p}$ and is called an (m,p)-point motion of the flow. Let $V_m \doteq \mathscr{C}\big([0,T]:\mathbb{R}^{d\times m}\big)$ be the space of all continuous maps from $[0,T]$ to $\mathbb{R}^{d\times m}$, equipped with the usual uniform topology, and let $V_{m,p} = V_m \times V_p$ be the product space. Let $\bar{\phi}^n(\mathbf{x})$ and $\bar{\Phi}^n(\mathbf{y})$ denote $\bar{\phi}_\cdot^n(\mathbf{x})$ and $\bar{\Phi}^n(\mathbf{y},\cdot)$, respectively.

Definition 12.9 Let $P^n_{\mathbf{x},\mathbf{y}}$ and $P_{\mathbf{x},\mathbf{y}}$ be the measures induced by $(\bar{\phi}^n(\mathbf{x}), \bar{\Phi}^n(\mathbf{y}))$ and $(\bar{\phi}(\mathbf{x}), \bar{\Phi}(\mathbf{y}))$ on $V_{m,p}$, respectively. Thus for $A \in \mathscr{B}(V_{m,p})$,

$$P^n_{\mathbf{x},\mathbf{y}} = P\big((\bar{\phi}^n(\mathbf{x}), \bar{\Phi}^n(\mathbf{y})) \in A\big), \quad P_{\mathbf{x},\mathbf{y}} = P\big((\bar{\phi}(\mathbf{x}), \bar{\Phi}(\mathbf{y})) \in A\big).$$

The sequence $\{(\bar{\phi}^n, \bar{\Phi}^n)\}_{n\in\mathbb{N}}$ is said to **converge weakly as diffusions** to $(\bar{\phi}, \bar{\Phi})$ as $n \to \infty$ if $P^n_{\mathbf{x},\mathbf{y}}$ converges weakly to $P_{\mathbf{x},\mathbf{y}}$ as $n \to \infty$ for each $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{d\times m} \times \mathbb{R}^{d\times p}$ and $m, p \in \mathbb{N}$.

The following result [178, Theorem 5.1.1] is a key ingredient in the proof of Theorem 12.8. It is analogous to the criteria for weak convergence of finite dimensional stochastic processes in terms of tightness as processes and convergence of finite dimensional distributions.

Theorem 12.10 *The family of probability measures $\hat{P}^n_{k-1}$(respectively, P^n_{k-1}) converges weakly to probability measures $\hat{P}_{k-1}$(respectively, P_{k-1}) as $n \to \infty$ if and only if the following two conditions are satisfied:*

1. *The sequence $\{(\bar{\phi}^n, \bar{\Phi}^n)\}_{n\in\mathbb{N}}$ converges weakly as diffusions to $(\bar{\phi}, \bar{\Phi})$ as $n \to \infty$.*
2. *The sequence $\{\hat{P}^n_{k-1}\}$ (respectively, $\{P^n_{k-1}\}$) is tight.*

We will show first that under the condition of Theorem 12.8, the sequence $\{(\bar{\phi}^n, \bar{\Phi}^n)\}$ converges weakly as diffusions to $(\bar{\phi}, \bar{\Phi})$ as $n \to \infty$. We begin with the following lemma.

Lemma 12.11 *For each $x \in \mathbb{R}^d$,*

$$E \sup_{0\le t\le T} \left\| \sum_{l=1}^{\infty} \int_0^t f_l(x, s)\, d\beta_l(s) \right\|^2 < \infty, \tag{12.16}$$

$$\sup_{n\in\mathbb{N}} E \sup_{0\le t\le T} \left\| \sum_{l=1}^{\infty} \int_0^t f_l(\bar{\phi}^n_s(x), s)\, d\beta_l(s) \right\|^2 < \infty. \tag{12.17}$$

Proof We will prove only (12.17). The proof of (12.16) follows in a similar manner. From the Burkholder–Davis–Gundy inequality [see (D.3)], (12.5), and the definition of $\|a\|^{\sim}_{T,k,\delta}$, the left-hand side of (12.17) is bounded by

$$\begin{aligned} & c_1 E \left| \sum_{l=1}^{\infty} \int_0^T \mathrm{tr}\big(f_l(\bar{\phi}^n_s(x), s)\, f_l^T(\bar{\phi}^n_s(x), s)\big) ds \right| \\ & \quad = c_1 E \left| \int_0^T \mathrm{tr}\left(a(\bar{\phi}^n_s(x), \bar{\phi}^n_s(x), s)\right) ds \right| \\ & \quad \le c_2 \|a\|^{\sim}_{T,k,\delta}, \end{aligned}$$

where for a $d \times d$ matrix A, $\mathrm{tr}(A)$ denotes its trace. The last expression is finite, since a belongs to $\tilde{\mathscr{C}}_T^{k,\delta}(\mathbb{R}^{d\times d})$. □

An immediate consequence of Lemma 12.11 is the following corollary [recall (12.14), (12.15), and the definition of $M(x,t)$ in (12.6), (12.7)]. The continuity in t follows from the fact that $M(x,\cdot)$ is continuous for each $x \in \mathbb{R}^d$.

Corollary 12.12 *For each $x \in \mathbb{R}^d$ and $t \in [0,T]$, we have*

$$\bar{\Phi}^n(x,t) = \int_0^t \hat{b}_{u^n}(x,r)dr + R_n(x,t)$$

and

$$\bar{\phi}_t^n(x) = x + \int_0^t \hat{b}_{u^n}(\bar{\phi}_r^n(x),r)dr + T_n(x,t),$$

where $R_n(x,\cdot)$ and $T_n(x,\cdot)$ are continuous stochastic processes with values in $\mathbb{R}^d$, which satisfy $\sup_{0\le t\le T}\{\|R_n(x,t)\| + \|T_n(x,t)\|\} \to 0$ in probability as $n \to \infty$.

Since $\mathbf{x}$ and $\mathbf{y}$ are finite dimensional, the next lemma will imply the tightness of $P^n_{\mathbf{x},\mathbf{y}}$.

Lemma 12.13 *For each $x \in \mathbb{R}^d$, the sequence $\{(\bar{\phi}^n(x), \bar{\Phi}^n(x))\}_{n\in\mathbb{N}}$ is tight in $\mathscr{C}([0,T] : \mathbb{R}^d \times \mathbb{R}^d)$.*

Proof We will argue only the tightness of $\{\bar{\phi}^n(x)\}$, since tightness of $\{\bar{\Phi}^n(x)\}$ is proved similarly. Corollary 12.12 yields that $T_n(x,\cdot)$ is tight in $\mathscr{C}([0,T] : \mathbb{R}^d)$. Thus it suffices to show the tightness of $\{\int_0^\cdot \hat{b}_{u^n}(\bar{\phi}_r^n(x),r)dr\}$. Consider any $p \in (0,\infty)$ and recall that $\hat{b}_u \doteq b_u + b$. From the Cauchy-Schwarz inequality, (12.10), and recalling that $u^n \in \bar{\mathscr{A}}_{b,N}$, we see that $E\|\int_s^t \hat{b}_{u^n}(\bar{\phi}_r^n(x),r)dr\|^p$ is bounded by

$$E\left[\int_s^t \|\hat{b}_{u^n}(\bar{\phi}_r^n(x),r)\|^2 dr\right]^{p/2}(t-s)^{p/2} \le c_1(\|a\|_{T,k,\delta}^{\sim} + \|b\|_{T,k,\delta}^2)^{p/2}(t-s)^{p/2}$$
$$\le c_2(t-s)^{p/2}.$$

The result follows. □

Proposition 12.14 *The sequence $\{(\bar{\phi}^n, \bar{\Phi}^n)\}_{n\in\mathbb{N}}$ converges weakly as diffusions to $(\bar{\phi}, \bar{\Phi})$ as $n \to \infty$.*

Proof We recall that b_u was defined above (12.10) according to $\int_0^t b_u(x,s)ds = \langle\langle Z, M(x,\cdot)\rangle\rangle_t$ for $(x,t) \in \mathbb{R}^d \times [0,T]$, that it takes the form $b_u(x,t) \doteq \sum_{l=1}^\infty u_l(t) f_l(x,t)$ a.e. (t,ω), and that $\hat{b}_u \doteq b_u + b$. We claim that it suffices to show that for each $t \in [0,T]$, the map

$$(\xi, v) \mapsto \int_0^t \hat{b}_v(\xi_s, s)ds \tag{12.18}$$

from $\mathscr{C}([0,T]:\mathbb{R}^d)\times S_N$ to $\mathbb{R}^d$ is continuous. To prove the claim, note that in view of the tightness established in Lemma 12.13 and uniqueness of the solution of the second equation in (12.19), the proposition will follow if for every $x\in\mathbb{R}^d$ and each fixed $t\in[0,T]$, every weak limit point $(\bar\phi(x),\bar\Phi(x),\bar u)$ of $(\bar\phi^n(x),\bar\Phi^n(x),u^n)$ satisfies

$$\bar\Phi(x,t)=\int_0^t \hat b_{\bar u}(x,r)dr,\quad \bar\phi_t(x)=x+\int_0^t \hat b_{\bar u}(\bar\phi_r(x),r)dr,\ \text{a.s.} \tag{12.19}$$

Now fix a weakly convergent subsequence of $(\bar\phi^n(x),\bar\Phi^n(x),u^n)$ and $t\in[0,T]$. From (12.18) and $(\bar\phi^n(x),u^n(x))\Rightarrow(\bar\phi(x),\bar u(x))$, it follows that

$$\left(\bar\phi^n_t(x),\int_0^t \hat b_{u^n}(\bar\phi^n_r(x),r)dr\right)\Rightarrow\left(\bar\phi_t(x),\int_0^t \hat b_{\bar u}(\bar\phi_r(x),r)dr\right).$$

The second equality in (12.19) is now an immediate consequence of the second equality in Corollary 12.12. The first equality in (12.19) is proved similarly on noting that (12.18) in particular implies that for each $x\in\mathbb{R}^d$, the map $v\mapsto\int_0^t\hat b_v(x,s)ds$, from S_N to $\mathbb{R}^d$, is continuous.

We now prove the continuity of the mapping (12.18). Let $(\xi^n,v^n)\to(\xi,v)$ in $\mathscr{C}([0,T]:\mathbb{R}^d)\times S_N$. Then

$$\left\|\int_0^t\big(\hat b_{v^n}(\xi^n_s,s)-\hat b_v(\xi_s,s)\big)ds\right\|\le T^n_1+T^n_2, \tag{12.20}$$

where

$$T^n_1\doteq\left\|\int_0^t\big(\hat b_{v^n}(\xi^n_s,s)-\hat b_{v^n}(\xi_s,s)\big)ds\right\|,\quad T^n_2\doteq\left\|\int_0^t\big(\hat b_{v^n}(\xi_s,s)-\hat b_v(\xi_s,s)\big)ds\right\|.$$

Recall that $\hat b_u=b+b_u$ and $b_u(x,t)=\sum_{l=1}^\infty u_l(t)f_l(x,t)$. Since $v^n\to v$ weakly in $\mathscr{L}^2([0,T]:l_2)$ and

$$\sum_{l=1}^\infty\int_0^t\|f_l(x,s)\|^2ds\le T\|a\|^{\sim}_{T,k,\delta}<\infty,$$

we have for each $x\in\mathbb{R}^d$ that

$$\left\|\int_0^t\big(\hat b_{v^n}(x,s)-\hat b_v(x,s)\big)ds\right\|=\left\|\sum_{l=1}^\infty\int_0^t f_l(x,s)(v^n_l(s)-v_l(s))ds\right\|\to 0. \tag{12.21}$$

Furthermore, from (12.10) (recall $k\ge1$), we have that for some $c_1\in(0,\infty)$ and all $x,y\in\mathbb{R}^d$, $0\le t\le T$,

$$\left\| \int_0^t \left(\hat{b}_{v^n}(x,s) - \hat{b}_{v^n}(y,s) \right) ds \right\| \le \|x-y\| \int_0^t \left(\|b_{v^n}(s)\|_{k,\gamma} + \|b(s)\|_{k,\gamma} \right) ds$$
$$\le c_1 \|x-y\|. \tag{12.22}$$

The Ascoli–Arzelà theorem (in the spatial variable) and equations (12.21), (12.22) yield that the expression on the left side of (12.21) converges to 0 uniformly for x in compact subsets of $\mathbb{R}^d$. Thus $T_2^n \to 0$ as $n \to \infty$. Following similar arguments, T_1^n is bounded by $c_2 \sup_{0 \le s \le T} \|\xi_s^n - \xi_s\|$, which converges to 0 as $n \to \infty$. Hence (12.20) converges to 0 as $n \to \infty$, and the result follows. □

We next show the tightness of the family of probability measures $\{P_{k-1}^n\}$, where P_{k-1}^n is the measure induced by $(\bar{\phi}^n, \bar{\Phi}^n)$ on $\mathscr{W}_{k-1} \times \mathscr{W}_{k-1}$. Key ingredients in the proof are the following uniform $\mathscr{L}^p$ estimates on $\partial^\alpha \bar{\Phi}^n(x,t)$ and $\partial^\alpha \bar{\phi}_t^n(x)$.

Lemma 12.15 *For each $p \in [1,\infty)$, there exists $k_1 \in (0,\infty)$ such that for all $t, s \in [0,T]$, $x \in \mathbb{R}^d$, $n \in \mathbb{N}$, and $|\alpha| \le k$,*

$$E \left\| \partial^\alpha \bar{\Phi}^n(x,t) - \partial^\alpha \bar{\Phi}^n(x,s) \right\|^p \le k_1 \, |t-s|^{p/2}. \tag{12.23}$$

Proof We use that $u \in \bar{\mathscr{A}}_{b,N}$ [which was defined in (12.4)] and that $\gamma \in (0,\delta)$. Fix a multi-index α such that $|\alpha| \le k$ and $p \in [1,\infty)$. Using the Burkholder–Davis–Gundy inequality for the martingale $\partial^\alpha M(x,\cdot)$ and the fact that $a \in \widetilde{\mathscr{C}}_T^{k,\delta}(\mathbb{R}^d)$, we obtain that for some $c_1 \in (0,\infty)$ and all $x \in \mathbb{R}^d$, $t, s \in [0,T]$, $s \le t$,

$$E \left\| \partial^\alpha M(x,t) - \partial^\alpha M(x,s) \right\|^p \le c_1 \, |t-s|^{p/2}. \tag{12.24}$$

Recalling that $\hat{b}_{u^n}(\cdot,t) \in \mathscr{C}^{k,\gamma}(\mathbb{R}^d)$ for a.e. (t,ω) and using (12.10), we get

$$\int_0^t \sup_{x \in \mathbb{R}^d} \|\partial^\alpha \hat{b}_{u^n}(x,r)\| dr < \infty \text{ a.e.},$$

and thus $\partial^\alpha \int_0^t \hat{b}_{u^n}(x,r) dr = \int_0^t \partial^\alpha \hat{b}_{u^n}(x,r) dr$ a.e. An application of the Cauchy-Schwarz inequality and (12.10) now give, for some $c_2 \in (0,\infty)$, that

$$E \left\| \partial^\alpha \int_s^t \hat{b}_{u^n}(x,r) dr \right\|^p \le c_2 \, |t-s|^{p/2}. \tag{12.25}$$

Equation (12.23) is an immediate consequence of (12.24), (12.25), and the definition of $\bar{\Phi}^n$ in (12.14). □

For $g : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$, let $D_y g(y,r)$ denote the $d \times d$ matrix whose ijth entry is $\partial g_i(y,r)/\partial y_j$, and let ∂_i denote differentiation with respect to x_i. Differentiating in (12.15), we obtain

$$\begin{aligned}\partial_1\bar{\phi}^n_t(x) &= \partial_1 x + \int_0^t [D_y\hat{b}_{u^n}(\bar{\phi}^n_r(x), r)\partial_1\bar{\phi}^n_r(x)]dr \\ &\quad + \sqrt{\varepsilon_n}\int_0^t D_y M(\bar{\phi}^n_r(x), dr)\partial_1\bar{\phi}^n_r(x) \\ &= \partial_1 x + \int_0^t D_y\bar{\Phi}^n(\bar{\phi}^n_r(x), dr)\partial_1\bar{\phi}^n_r(x).\end{aligned}$$

By repeated differentiation, one obtains the following lemma, whose proof follows along the lines of Theorem 3.3.3 of [178]. Given $0 \le m \le k$, let Λ_m be the set of all multi-indices α satisfying $|\alpha| \le m$. For a multi-index γ, let $m(\gamma)$ denote the number of multi-indices γ_0 such that $|\gamma_0| \le |\gamma|$. Also, for a $|\gamma|$-times differentiable function $\Psi : \mathbb{R}^d \to \mathbb{R}$, denote by $\partial^{\le|\gamma|}\Psi(x)$ the $m(\gamma)$-dimensional vector with entries $\partial^{\gamma_0}\Psi(x)$, $|\gamma_0| \le |\gamma|$. If $\Psi = (\Psi_1, \Psi_2, \ldots, \Psi_d) : \mathbb{R}^d \to \mathbb{R}^d$ is such that each Ψ_i is $|\gamma|$-times continuously differentiable, then $\partial^{\le|\gamma|}\Psi(x) \doteq (\partial^{\le|\gamma|}\Psi_1(x), \ldots, \partial^{\le|\gamma|}\Psi_d(x))$. We will call a map $\zeta : \mathbb{R}^m \to \mathbb{R}^d$ a polynomial of degree at most $\wp$ if $\zeta(x) = (\zeta_1(x), \ldots, \zeta_d(x))^T$ and each $\zeta_i : \mathbb{R}^m \to \mathbb{R}$ is a polynomial of degree at most $\wp$. Also, for $u, v \in \mathbb{R}^l$, we define

$$u * v \doteq (u_1 v_1, \ldots, u_l v_l)^T.$$

Lemma 12.16 *Let α be a multi-index such that $|\alpha| \le k$. Then there exist subsets $\Lambda^1_\alpha \subset \Lambda_{|\alpha|}$, $\Lambda^2_\alpha \subset \Lambda_{|\alpha|-1}$ and polynomials $\zeta^\alpha_{\beta,\theta} : \mathbb{R}^{m(\theta)} \to \mathbb{R}^d$ of degree at most $|\alpha|$ such that $\partial^\alpha\bar{\phi}^n$ satisfies*

$$\begin{aligned}\partial^\alpha\bar{\phi}^n_t(x) &= \partial^\alpha x + \int_0^t G^n\big(\partial^\alpha\bar{\phi}^n_r(x), \bar{\phi}^n_r(x), dr\big) \\ &\quad + \sum_{(\beta,\theta)\in\Lambda^1_\alpha\times\Lambda^2_\alpha}\int_0^t G^{\alpha,n}_{\beta,\theta}\big(\partial^{\le|\theta|}\bar{\phi}^n_r(x), \bar{\phi}^n_r(x), dr\big),\end{aligned} \tag{12.26}$$

where for $x, y \in \mathbb{R}^d$, $G^n(x, y, r) \doteq D_y\bar{\Phi}^n(y, r)\cdot x$, and for $(x, y) \in \mathbb{R}^{m(\theta)} \times \mathbb{R}^d$,

$$G^{\alpha,n}_{\beta,\theta}(x, y, r) \doteq \zeta^\alpha_{\beta,\theta}(x) * \partial^\beta\bar{\Phi}^n(y, r).$$

Note in particular that in the third term on the right-hand side of (12.26), one finds only partial derivatives of $\bar{\phi}^n_r(x)$ of order strictly less than $|\alpha|$.

Lemma 12.17 *For each $p \in [1, \infty)$ and $L \in (0, \infty)$, there is $k_1 = k_1(k, p, L) \in (0, \infty)$ such that for every multi-index α, $|\alpha| \le k$, and $s, t \in [0, T]$,*

$$\sup_{n\in\mathbb{N}}\sup_{\|x\|\le L} E\sup_{0\le t\le T}\|\partial^\alpha\bar{\phi}^n_t(x)\|^p \le k_1 \tag{12.27}$$

$$\sup_{n\in\mathbb{N}}\sup_{\|x\|\le L} E\big\|\partial^\alpha\bar{\phi}^n_t(x) - \partial^\alpha\bar{\phi}^n_s(x)\big\|^p \le k_1\,|t-s|^{p/2}. \tag{12.28}$$

Proof We first prove inequality (12.27). It suffices to prove (12.27) for $\alpha = 0$ and establish that if for some $m < k$, it holds for $\partial^\alpha \bar{\phi}_t^n$ with $|\alpha| \le m$ and all $p \in [1, \infty)$, then it also holds for $\partial_i \partial^\alpha \bar{\phi}_t^n$ with all $p \in [1, \infty)$ (with a possibly larger constant k_1) and $i = 1, \ldots, d$. The desired result then follows by induction.

Consider first $\alpha = 0$. For this case, the bound in (12.27) follows immediately on using (12.10) and applying the Burkholder–Davis–Gundy inequality to the square-integrable martingale $N_t = \int_0^t M(\bar{\phi}_r^n(x), dr)$ [here we use that $\langle\langle N \rangle\rangle_t = \int_0^t a(\bar{\phi}_r^n(x), \bar{\phi}_r^n(x), r)dr$ and $a \in \widetilde{\mathscr{C}}_T^{k,\delta}(\mathbb{R}^{d\times d})$].

Next suppose that (12.27) holds for all multi-indices α with $|\alpha| \le m$, for some $m < k$. Fix α with $|\alpha| \le m$, and $i \in \{1, 2, \ldots, d\}$, and consider the multi-index $\tilde{\alpha} = \alpha + e_i$, where e_i is a d-dimensional vector with 1 in the ith entry and 0 elsewhere. From Lemma 12.16, one finds that $\partial^{\tilde{\alpha}} \bar{\phi}_t^n$ solves (12.26) for $\alpha = \tilde{\alpha}$. Note that for $\beta \in \Lambda_{\tilde{\alpha}}^1$,

$$\partial^\beta \bar{\Phi}^n(x,t) = \int_0^t \partial^\beta \hat{b}_{u^n}(x,s)ds + \sqrt{\varepsilon_n}\partial^\beta M(x,t).$$

From (12.10) and recalling that $(b, a) \in \mathscr{C}_T^{k,\delta}(\mathbb{R}^d) \times \widetilde{\mathscr{C}}_T^{k,\delta}(\mathbb{R}^{d\times d})$, we have that for some $c_1, c_2 \in (0, \infty)$,

$$\sup_{n\in\mathbb{N}} \sup_{0\le t\le T} \sup_{x\in\mathbb{R}^d} \left\| \int_0^t \partial^\beta \hat{b}_{u^n}(x,s)ds \right\| \le c_1 \text{ and } \sup_{0\le t\le T} \sup_{x\in\mathbb{R}^d} \left\| \langle\langle \partial^\beta M(x,\cdot) \rangle\rangle_t \right\| \le c_2.$$

This along with the assumption

$$\sup_{n\in\mathbb{N}} \sup_{\|x\|\le L} E \sup_{0\le t\le T} \|\partial^\nu \bar{\phi}_t^n(x)\|^p \le k_1 \text{ for all } \nu \text{ with } |\nu| \le |\alpha|$$

shows that for some $c_3 \in (0, \infty)$, for all $(\beta, \theta) \in \Lambda_{\tilde{\alpha}}^1 \times \Lambda_{\tilde{\alpha}}^2$,

$$\sup_{n\in\mathbb{N}} \sup_{\|x\|\le L} E \sup_{0\le t\le T} \left\| \int_0^t G_{\beta,\theta}^{\tilde{\alpha},n}\left(\partial^{\le|\theta|}\bar{\phi}_r^n(x), \bar{\phi}_r^n(x), dr\right) \right\|^p \le c_3.$$

Also, in a similar manner one has for some $c_4 \in (0, \infty)$,

$$E \sup_{0\le t\le T} \left\| \int_0^s G^n\left(\partial^{\tilde{\alpha}}\bar{\phi}_r^n(x), \bar{\phi}_r^n(x), dr\right) \right\|^p \le c_4 \int_0^T E \sup_{0\le r\le s} \|\partial^{\tilde{\alpha}}\bar{\phi}_r^n(x)\|^p ds.$$

Combining the last two inequalities with (12.26), we obtain

$$\sup_{n\in\mathbb{N}} \sup_{\|x\|\le L} E \sup_{0\le s\le t} \left\| \partial^{\tilde{\alpha}}\bar{\phi}_s^n(x) \right\|^p \le c_3 + c_4 \sup_{n\in\mathbb{N}} \sup_{\|x\|\le L} \int_0^t E\left(\sup_{0\le r\le s} \left\| \partial^{\tilde{\alpha}}\bar{\phi}_r^n(x) \right\|^p \right) ds,$$

where the constants c_3, c_4 depend on L. Now an application of Gronwall's inequality shows that for some $c_5 \in (0, \infty)$,

$$\sup_{n\in\mathbb{N}}\sup_{\|x\|\le L} E \sup_{0\le t\le T} \|\partial^{\tilde{\alpha}}\bar{\phi}_t^n(x)\|^p \le c_5.$$

This establishes (12.27) for all $\tilde{\alpha}$ with $|\tilde{\alpha}| \le |\alpha| + 1$. Finally, consider (12.28). For $t, s \in [0, T]$, $s \le t$, we have from (12.26) that

$$\partial^\alpha \bar{\phi}_t^n(x) - \partial^\alpha \bar{\phi}_s^n(x) = \int_s^t G^n\big(\partial^\alpha \bar{\phi}_r^n(x), \bar{\phi}_r^n(x), dr\big) + \sum_{(\beta,\theta)\in\Lambda_\alpha^1\times\Lambda_\alpha^2} \int_s^t G_{\beta,\theta}^{\alpha,n}\big(\partial^{\le|\theta|}\bar{\phi}_r^n(x), \bar{\phi}_r^n(x), dr\big). \tag{12.29}$$

Using (12.27) on the right-hand side of (12.29), we now have (12.28) via an application of Hölder's inequality and the Burkholder–Davis–Gundy inequality. □

The proof of Theorem 12.8 proceeds along the lines of Sect. 5.4 of [178]. We begin by introducing certain Sobolev spaces. Let $j \in \mathbb{N}$ and $p \in (1, \infty)$ be given. Let B_N denote the open ball in $\mathbb{R}^d$ with radius N. Let $h : \mathbb{R}^d \to \mathbb{R}^d$ be a function such that the distributional derivative $\partial^\alpha h$, when restricted to B_N, is in $\mathscr{L}^p(B_N)$ for all α such that $|\alpha| \le j$. For such h, define

$$\|h\|_{j,p:N} \doteq \left(\sum_{|\alpha|\le j}\int_{B_N} \big\|\partial^\alpha h(x)\big\|^p dx\right)^{1/p}.$$

The space $H_{j,p}^{loc} \doteq \{h : \mathbb{R}^d \to \mathbb{R}^d,\ \|h\|_{j,p:N} < \infty \text{ for all } N\}$ together with the seminorms defined above is a real separable semireflexive Fréchet space (see [226], [178, Sect. 5.4]). By Sobolev's embedding theorem, we have $H_{j+1,p}^{loc} \subset \mathscr{C}^j(\mathbb{R}^d) \subset H_{j,p}^{loc}$ if $p > d$. Furthermore, the embedding $\iota : H_{j+1,p}^{loc} \to \mathscr{C}^j(\mathbb{R}^d)$ is a compact operator by the Rellich–Kondrachov theorem (see [1]).

Proposition 12.18 *The sequence* $\{(\bar{\phi}^n, \bar{\Phi}^n)\}_{n\in\mathbb{N}}$ *is tight in* $\mathscr{W}_{k-1} \times \mathscr{W}_{k-1}$.

Proof It suffices to show that both $\{\bar{\phi}^n\}_{n\in\mathbb{N}}$ and $\{\bar{\Phi}^n\}_{n\in\mathbb{N}}$ are tight in $\mathscr{W}_{k-1}$. We will use Kolmogorov's tightness criterion [178, Theorem 1.4.7, p. 38]. From Lemmas 12.15 and 12.17, we have that for each $p \in (1, \infty)$, $N \in \mathbb{N}$, there exist $c_1, c_2 \in (0, \infty)$ such that for all $t, s \in [0, T]$,

$$\sup_{n\in\mathbb{N}} E\|\bar{\phi}_t^n - \bar{\phi}_s^n\|_{k,p:N}^p \le c_1\,|t-s|^{p/2},$$
$$\sup_{n\in\mathbb{N}} E\|\bar{\Phi}^n(t) - \bar{\Phi}^n(s)\|_{k,p:N}^p \le c_2\,|t-s|^{p/2}.$$

Furthermore, since $\bar{\Phi}^n(\cdot, 0) = 0$ and $\bar{\phi}_0^n(x) = x$, we get that for each $p \in (1, \infty)$, $N \in \mathbb{N}$, there exist $c_3, c_4 \in (0, \infty)$ such that for all $t \in [0, T]$,

$$\sup_{n\in\mathbb{N}} E\|\bar{\phi}_t^n\|_{k,p:N}^p \le c_3 \text{ and } \sup_{n\in\mathbb{N}} E\|\bar{\Phi}^n(t)\|_{k,p:N}^p \le c_4.$$

Applying Theorem 1.4.7 of [178] with $p > 2$ now gives tightness of $\{\bar{\phi}^n\}_{n\in\mathbb{N}}$ and $\{\bar{\Phi}^n\}_{n\in\mathbb{N}}$ in the semiweak topology (see [178, Sect. 1.4]) on $\mathscr{C}([0,T] : H^{loc}_{k,p})$. Since the embedding map $\iota : H^{loc}_{k,p} \to \mathscr{C}^{k-1}$ is compact, tightness in $\mathscr{W}_{k-1} \times \mathscr{W}_{k-1}$ follows, where the topology is that generated by $\bar{d}_{k-1}$, with $\bar{d}_{k-1}(\phi,\psi) = \sup_{0\le t\le T} \lambda_{k-1}(\phi(t), \psi(t))$ and λ_{k-1} as defined in (12.9) (see [178, pp. 246–247]). □

This gives tightness in the space of smooth maps, and it will give an LDP in that space. To extend to smooth invertible maps we will use the following lemma, whose proof can be found in Sect. 2.1 of [16]. Recall the definitions (12.8) and (12.9).

Lemma 12.19 *Let* $f_n, f \in \hat{\mathscr{W}}_{k-1}$ *be such that* $\sup_{0\le t\le T} \lambda_{k-1}(f_n(t), f(t)) \to 0$ *as* $n \to \infty$. *Then* $\sup_{0\le t\le T} d_{k-1}(f_n(t), f(t)) \to 0$.

Proof of Theorem 12.8 Convergence as $\mathscr{C}^{k-1}$-flows follows directly from Theorem 12.10, Propositions 12.14 and 12.18. Using Skorohod's representation theorem, one can find a sequence of pairs $\{(\tilde{\phi}^n, \tilde{\Phi}^n)\}_{n\in\mathbb{N}}$ that has the same distribution as $\{(\bar{\phi}^n, \bar{\Phi}^n)\}_{n\in\mathbb{N}}$ and $(\tilde{\phi}, \tilde{\Phi})$, which has the same distribution as $(\bar{\phi}, \bar{\Phi})$, and such that $\sup_{0\le t\le T}\left[\lambda_k(\tilde{\phi}^n_t, \tilde{\phi}_t) + \lambda_k(\tilde{\Phi}^n(t), \tilde{\Phi}(t))\right] \to 0$, a.s. Since $\bar{\phi}^n, \bar{\phi} \in \hat{\mathscr{W}}_k$ a.s., we also have $\tilde{\phi}^n, \tilde{\phi} \in \hat{\mathscr{W}}_k$ a.s. Thus from Lemma 12.19, $\sup_{0\le t\le T} d_{k-1}(\tilde{\phi}^n_t, \tilde{\phi}_t) \to 0$ a.s. Hence $(\bar{\phi}^n, \bar{\Phi}^n) \to (\bar{\phi}, \bar{\Phi})$ in distribution as $\mathscr{D}^{k-1}$-flows. □

12.4 Application to Image Analysis

A common approach to image-matching problems (see [99, 151, 201] and references therein) is to consider an $\mathbb{R}^p$-valued continuous and bounded function $\mathscr{T}(\cdot)$, referred to as the "template" function, that represents some canonical example of a structure of interest. Although one can consider other scenarios, for problems from medical imaging the template is defined on some bounded open set $O \subset \mathbb{R}^3$, an assumption we make in this section. By considering all possible smooth transformations $h : O \to O$, one can generate a rich library of targets (or images) given by the form $\mathscr{T}\big(h(\cdot)\big)$.

In typical situations we are given data generated by an a priori unknown function h, and the key issue of image matching is estimating h from the observed data. A Bayesian approach to this problem requires a prior distribution on the space of transformations and the formulation of a noise/data model. The "maximum" of the posterior distribution on the space of transformations given the data can then be used as an estimate $\hat{h}$ for the unknown underlying transformation h. In certain applications (e.g., medical diagnosis), the goal is to obtain numerical approximations for certain key structures present in the image, such as volumes of subregions, curvatures and surface areas. If the prior distribution on the transformations (and in particular the estimated transformation) is on the space of diffeomorphisms, then this type of information can be recovered from the template. Motivated by such a Bayesian approach, a variational problem on the space of $\mathscr{C}^m$-diffeomorphic flows was formulated and analyzed in [99].

Before giving the description of this variational problem, we note that although the chief motivation for the variational problem studied in [99] came from Bayesian considerations, no rigorous results on relationships between the two formulations (variational and Bayesian) were established. The goal of this section is to apply the asymptotic theory developed earlier in the chapter and connect a Bayesian formulation of the image-matching problem with the variational approach taken in [99]. The precise result we establish is Theorem 12.20, given at the end of this section. The result is an application of Theorem 12.5 for local characteristics $(a, b) = (a, 0)$ and $a \in \widetilde{\mathscr{C}}_T^{k,1/2}(\mathbb{R}^{3\times 3})$, with $k = m - 2$.

Let $\mathscr{C}_0^\infty(O)$ be the space of infinitely differentiable real-valued functions on O with compact support in O. The starting point of the variational formulation is a differential operator L on $[\mathscr{C}_0^\infty(O)]^3$, the exact form of which is determined from specific features of the problem under study. The formulation, particularly for problems from biology, often uses principles from physics and continuum mechanics as a guide in the selection of L. We refer the reader to [61, 62], where natural choices of L for shape models from anatomy are provided.

Define the norm $\|\cdot\|_L$ on $[\mathscr{C}_0^\infty(O)]^3$ by

$$\|g\|_L^2 \doteq \sum_{i=1}^3 \int_O |(Lg)_i(u)|^2 du,$$

where we write a function $g \in [\mathscr{C}_0^\infty(O)]^3$ as $(g_1, g_2, g_3)^T$. It is assumed that $\|\cdot\|_L$ generates an inner product on $[\mathscr{C}_0^\infty(O)]^3$, and that the Hilbert space $\mathscr{H}$ defined as the closure of $[\mathscr{C}_0^\infty(O)]^3$ with this inner product is separable. We will need the functions in $\mathscr{H}$ to have sufficient regularity and thus assume that the norm $\|\cdot\|_L$ dominates an appropriate Sobolev norm. More precisely, let $\mathscr{W}_0^{m+2,2}(O)$ be the closure of $\mathscr{C}_0^\infty(O)$ with respect to the norm

$$\|g\|_{\mathscr{W}_0^{m+2,2}(O)} \doteq \Big(\int_O \sum_{|\alpha|\le m+2} |\partial^\alpha g(u)|^2 du\Big)^{1/2}, \quad g \in \mathscr{C}_0^\infty(O),$$

where α denotes a multi-index and $m \ge 3$. Define $\mathscr{V}_m \doteq [\mathscr{W}_0^{m+2,2}(O)]^{\otimes 3}$, where $\otimes$ is used to denote the usual tensor product of Hilbert spaces. We denote the norm on $\mathscr{V}_m$ by $\|\cdot\|_{\mathscr{V}_m}$. The main regularity condition on L is the following domination requirement on the $\|\cdot\|_L$ norm. There exists a constant $c \in (0, \infty)$ such that

$$\|f\|_L \ge c\|f\|_{\mathscr{V}_m} \text{ for all } f \in [\mathscr{C}_0^\infty(O)]^3.$$

This condition ensures that $\mathscr{H} \subset \mathscr{C}^{m,1/2}(\bar{O})$ (see [1, Theorem 4.12, parts II and III, p. 85]).

To simplify notation we will take $T = 1$, though in general, this parameter affects the tradeoff between the distributions used to model the prior and data noise. We

denote the Hilbert space $\mathscr{L}^2([0,1]:\mathscr{H})$ by $\mathscr{M}$. For a fixed $\vartheta \in \mathscr{M}$, let $\{\eta_{s,t}(x)\}_{s\le t\le 1}$ be the unique solution of the ordinary differential equation

$$\frac{\partial \eta_{s,t}(x)}{\partial t} \doteq \vartheta\big(\eta_{s,t}(x), t\big), \quad \eta_{s,s}(x) = x, \ 0 \le s \le t \le 1. \tag{12.30}$$

Then it follows that $\{\eta_{s,t}\}_{0\le s\le t\le 1}$ is a forward flow of $\mathscr{C}^m$-diffeomorphisms on O (see Theorem D.5). Since $\vartheta(\cdot, t)$ has compact support in $\mathscr{O}$, one can extend $\eta_{s,t}$ to all of $\mathbb{R}^3$ by setting $\eta_{s,t}(x) = x$ for $x \in O^c$. Extended in this way, $\eta_{s,t}$ can be considered an element of $\mathscr{D}^m$, as defined in Sect. 12.2. Denoting $\eta_{0,1}$ by h_ϑ, we can now generate a family of smooth transformations (diffeomorphisms) on O by varying $\vartheta \in \mathscr{M}$. Specifically, the library of transformations that is used in the variational formulation of the image-matching problem is $\{h_\vartheta : \vartheta \in \mathscr{M}\}$.

We next describe the data that is used in selecting the transformation h_{ϑ^*} for which the image $\mathscr{T}\big(h_{\vartheta^*}(\cdot)\big)$ best matches the data. Let $\mathscr{I}$ be a finite index set and $\{O_i\}_{i\in\mathscr{I}}$ a collection of disjoint subsets of O such that $\cup_{i\in\mathscr{I}} O_i = O$. The data $\{d_i\}_{i\in\mathscr{I}}$ represent the integrated responses over each of the subsets O_i, $i \in \mathscr{I}$. More precisely, if $\mathscr{T}\big(h(\cdot)\big)$ were the true underlying image and if the data were error-free and noiseless, then we would have $d_i = \int_{O_i} \mathscr{T}(h(\sigma))d\sigma/\mathrm{vol}(O_i)$, $i \in \mathscr{I}$, where vol denotes volume. Let $d = (d_1, d_2, \ldots, d_n)^T$, where $n = |\mathscr{I}|$. Defining $Y_d(x) = d_i$, $x \in O_i$, $i \in \mathscr{I}$, the expression

$$\frac{1}{2}\int_O \big\|\mathscr{T}(h_\vartheta(x)) - Y_d(x)\big\|^2 dx$$

is a measure of discrepancy between a candidate target image $\mathscr{T}(h_\vartheta(\cdot))$ and the observations. This suggests a natural variational criterion for selecting the "best" transformation matching the data. The objective function that is minimized in the variational formulation of the image-matching problem is a sum of two terms, the first reflecting the "likelihood" of the transformation or change-of-variable h_ϑ, and the second measuring the conformity of the transformed template with the observed data. More precisely, for $\vartheta \in \mathscr{M}$ define

$$J_d(\vartheta) \doteq \frac{1}{2}\Big(\|\vartheta\|^2_{\mathscr{M}} + \int_O \big\|\mathscr{T}(h_\vartheta(x)) - Y_d(x)\big\|^2 dx\Big). \tag{12.31}$$

Then $\vartheta^* \in \mathrm{argmin}_{\vartheta\in\mathscr{M}} J_d(\vartheta)$ represents the "optimal" velocity field that matches the data d and for which the h_{ϑ^*}, obtained by solving (12.30), gives the "optimal" transformation. This transformation then yields an estimate of the target image as $\mathscr{T}(h_{\vartheta^*}(\cdot))$. Suppose that for each $h \in \mathscr{D}^0$, we define

$$\hat{J}_d(h) \doteq \inf_{\vartheta\in\Psi_h} J_d(\vartheta), \tag{12.32}$$

where $\Psi_h \doteq \{\vartheta \in \mathscr{M} : h = h_\vartheta\}$. Then an equivalent characterizations is $h^* = h_{\vartheta^*} \in \mathrm{argmin}_h \hat{J}_d(h)$.

Up to a relabeling of the time variable, the variational formulation just described (in particular the cost function in (12.31)) was motivated in [99] by Bayesian considerations, but no rigorous justification was provided. [In [99], the orientation of time is consistent with the change-of-variable evolving toward the identity mapping at the terminal time. To relate the variational problem to stochastic flows, it is more convenient to have the identity mapping at time zero.] We next introduce a stochastic Bayesian formulation of the image-matching problem and describe the precise asymptotic result that we will establish.

Let $\{f_i\}$ be a complete orthonormal system in $\mathscr{H}$ and let $\beta = (\beta_i)_{i=1}^{\infty}$ be, as in Sect. 12.1, a sequence of independent standard real-valued Brownian motions on $(\Omega, \mathscr{F}, P, \{\mathscr{F}_t\})$. Consider the stochastic flow

$$d\phi_{s,t}(x) = \sqrt{\varepsilon} \sum_{i=1}^{\infty} f_i\big(\phi_{s,t}(x)\big) d\beta_i(t), \quad \phi_{s,s}(x) = x, \ x \in O, \ 0 \le s \le t \le 1, \tag{12.33}$$

where $\varepsilon \in (0, \infty)$ is fixed. From Maurin's theorem (see [1, Theorem 6.61, p. 202]), it follows that the embedding map $\mathscr{H} \to \mathscr{V}_{m-2} \doteq [\mathscr{W}_0^{m,2}(O)]^{\otimes 3}$ is Hilbert–Schmidt. Also, $\mathscr{V}_{m-2}$ is continuously embedded in $\mathscr{C}^{m-2,1/2}(\bar{O})$. Thus for some $k_1, k_2 \in (0, \infty)$ and all $u, x, y \in O$,

$$\sum_{i=1}^{\infty} \|f_i(u)\|^2 \le k_1 \sum_{i=1}^{\infty} \|f_i\|_{\mathscr{V}_{m-2}}^2 < \infty,$$

$$\sum_{i=1}^{\infty} \|f_i(x) - f_i(y)\|^2 \le k_1 \|x - y\|^2 \sum_{i=1}^{\infty} \|f_i\|_{\mathscr{V}_{m-2}}^2 = k_2 \|x - y\|^2.$$

One also has that if f_l is extended to all of $\mathbb{R}^3$ by setting $f_l(u) = 0$ for all $x \in O^c$, then $a(x, y) = \sum_{l=1}^{\infty} f_l(x) f_l^T(y)$ is in $\mathscr{C}_T^{m-2,1/2}(\mathbb{R}^{3\times 3})$. Thus it follows (see [178, pages 80 and 106]) that

$$\Phi(x, t) = \sum_{l=1}^{\infty} \int_0^t f_l(x) d\beta_l(r)$$

is a $\mathscr{C}^{m-2,\nu}$-Brownian motion, $0 < \nu < 1/2$, with local characteristics $(a, 0)$. Also, (12.33) admits a unique solution $\{\phi_{s,t}^{\varepsilon}(x)\}_{0\le s\le t\le 1}$ for each $x \in O$, and $\{\phi_{s,t}^{\varepsilon}\}_{0\le s\le t\le 1}$ is a forward flow of $\mathscr{C}^k$-diffeomorphisms, with $k = m - 2$ (see Theorem D.5). In particular, $X^{\varepsilon} \doteq \phi_{0,1}^{\varepsilon}$ is a random variable in the space of $\mathscr{C}^k$-diffeomorphisms on O. The law of X^{ε} (for a fixed $\varepsilon > 0$) on $\mathscr{D}^k$ will be used as the prior distribution on the transformation space $\mathscr{D}^k$. Note that $\mathscr{T}\big(X^{\varepsilon}(\cdot)\big)$ induces a measure on the space of target images.

We next consider the data model. Let $\mathscr{I}$ and n be as introduced below (12.30). We suppose that the data is given through an additive Gaussian noise model:

$$D_i = \int_{O_i} \mathscr{T}\big(X^\varepsilon(x)\big)dx + \sqrt{\varepsilon}\xi_i,$$

where $\{\xi_i\}_{i\in\mathscr{I}}$ is a family of independent p-dimensional standard normal random variables.

In a Bayesian approach to the image-matching problem one considers the posterior distribution of X^ε given the data D and uses the "mode" of this distribution as an estimate for the underlying true transformation. More precisely, let $\{\Gamma^\varepsilon\}_{\varepsilon>0}$ be a family of measurable maps from $\mathbb{R}^{np}$ to $\mathscr{P}(\mathscr{D}^k)$ (the space of probability measures on $\mathscr{D}^k$) such that

$$\Gamma^\varepsilon(A \mid D) = P\{X^\varepsilon \in A \mid D\} \quad \text{a.s. for all } A \in \mathscr{B}(\mathscr{D}^k).$$

We refer to $\Gamma^\varepsilon(\cdot \mid d)$ as a regular conditional probability distribution (r.c.p.d.) of X^ε given $D = d$. In Theorem 12.20 below, we show that there is an r.c.p.d. $\{\Gamma^\varepsilon(\cdot \mid d),\ d \in \mathbb{R}^{np}\}_{\varepsilon>0}$ such that for each $d \in \mathbb{R}^{np}$, the family $\{\Gamma^\varepsilon(\cdot \mid d)\}_{\varepsilon>0}$, regarded as elements of $\mathscr{P}(\mathscr{D}^{k-1}) \supset \mathscr{P}(\mathscr{D}^k)$, satisfies an LDP with rate function

$$I_d(h) \doteq \hat{J}_d(h) - \lambda_d, \text{ where } \lambda_d \doteq \inf_{h\in\mathscr{D}^{k-1}} \hat{J}_d(h) = \inf_{\vartheta\in\mathscr{M}} J_d(\vartheta).$$

Formally writing $\Gamma^\varepsilon(A \mid d) \approx \int_A e^{-\frac{I_d(h)}{\varepsilon}} dh$, one sees that for small ε, the "mode" of the posterior distribution given $D = d$, which represents the optimal transformation in the Bayesian formulation, can be formally interpreted as $\operatorname{argmin}_h I_d(h)$. Note that $\hat{J}_d(h) = \infty$ if $h \notin \mathscr{D}^m$ (recall $m = k + 2$). Theorem 12.20 in particular says that $h \in \mathscr{D}^m$ is a δ-minimizer for $I_d(h)$ (i.e., an element within $\delta > 0$ of the infimum) if and only if it is also a δ-minimizer for $\hat{J}_d(h)$. Thus Theorem 12.20 makes precise the asymptotic relationship between the variational and the Bayesian formulations of the image-matching problem described previously.

Theorem 12.20 *There exists an r.c.p.d. Γ^ε such that for each $d \in \mathbb{R}^n$, the family of probability measures $\{\Gamma^\varepsilon(d)\}_{\varepsilon>0}$ on $\mathscr{D}^{k-1}$ satisfies a large deviation principle (as $\varepsilon \to 0$) with rate function*

$$I_d(h) \doteq \hat{J}_d(h) - \lambda_d. \tag{12.34}$$

We begin with the following proposition. Let $\tilde{I} : \mathscr{D}^{k-1} \to [0, \infty]$ be defined by

$$\tilde{I}(h) \doteq \inf_{\vartheta\in\Psi_h} \left[\frac{1}{2}\|\vartheta\|^2_{\mathscr{M}}\right],$$

where $\Psi_h \doteq \{\vartheta \in \mathscr{M} : h = h_\vartheta\}$.

Proposition 12.21 *The family $\{X^\varepsilon\}_{\varepsilon>0}$ satisfies an LDP in $\mathscr{D}^{k-1}$ with rate function $\tilde{I}$.*

Remark 12.22 Proposition 12.21 is consistent with results in Sect. 12.2 in that although the local characteristics are in $\mathscr{C}^k$ and $X^\varepsilon \in \mathscr{D}^k$, the LDP is established

in the larger space $\mathscr{D}^{k-1}$. This is due to the tightness issues described at the beginning of the chapter. Furthermore, as noted below (12.30), if $\|\vartheta\|_{\mathscr{M}} < \infty$, then ϑ induces a flow of $\mathscr{C}^m$-diffeomorphisms on O. Thus if $h \in \mathscr{D}^{k-1} \setminus \mathscr{D}^m$, then Ψ_h is empty, and consequently $\tilde{I}(h) = \infty$. Hence there is a further widening of the "gap" between the regularity needed for the rate function to be finite and the regularity associated with the space in which the LDP is set. This is due to the fact that the variational problem is formulated essentially in terms of $\mathscr{L}^2$ norms of derivatives, while in the theory of stochastic flows as developed in [178], assumptions are phrased in terms of $\mathscr{L}^\infty$ norms.

Proof of Proposition 12.21 From Theorem 12.5 and an application of the contraction principle we have that $\{X^\varepsilon\}_{\varepsilon>0}$ satisfies an LDP in $\mathscr{D}^{k-1}$ with rate function

$$I^*(h) \doteq \inf_{u \in \mathscr{A}^*(h)} \left[\frac{1}{2} \int_0^T \|u(s)\|_0^2 ds \right],$$

where $\mathscr{A}^*(h) = \left\{ u \in \mathscr{L}^2([0,1] : l_2) : h = \bar{\phi}(1) \right\}$, and where $\bar{\phi}$ is defined via (12.12). Note that there is a one-to-one correspondence between $u \in \mathscr{L}^2([[0,1] : l_2)$ and $\vartheta \in \mathscr{M}$ given by $\vartheta(t,x) = \sum_{l=1}^\infty u_l(t) f_l(x)$ and $\int_0^T \|u(s)\|_0^2 ds = \|\vartheta\|_{\mathscr{M}}^2$. In particular, $u \in \mathscr{A}^*(h)$ if and only if $\vartheta \in \Psi_h$. Thus $I^*(h) = \tilde{I}(h)$, and the result follows. □

Proposition 12.23 *For each $d \in \mathbb{R}^n$, I_d defined in (12.34) is a rate function on $\mathscr{D}^{k-1}$.*

Proof From (12.34) and the definition of $\tilde{I}$, we have for $h \in \mathscr{D}^{k-1}$ that

$$\begin{aligned} I_d(h) = \tilde{I}(h) + \frac{1}{2} \int_O \|\mathscr{T}(h(x)) - Y_d(x)\|^2 dx \\ - \inf_{h \in \mathscr{D}^{k-1}} \left[\tilde{I}(h) + \frac{1}{2} \int_O \|\mathscr{T}(h(x)) - Y_d(x)\|^2 dx \right]. \end{aligned}$$

From Proposition 12.21, $\tilde{I}$ is a rate function and therefore has compact level sets. Additionally, $\mathscr{T}$ is a continuous and bounded function on O. The result follows. □

Proof of Theorem 12.20 We begin by noting that $\Gamma^\varepsilon(\cdot \mid d)$ defined by

$$\Gamma^\varepsilon(A \mid d) \doteq \frac{\displaystyle\int_A e^{-\frac{1}{2\varepsilon}\sum_{i=1}^n \left\| d_i - \int_{O_i} \mathscr{T}(h(y))dy \right\|^2} \mu^\varepsilon(dh)}{\displaystyle\int_{\mathscr{D}^{k-1}} e^{-\frac{1}{2\varepsilon}\sum_{i=1}^n \left\| d_i - \int_{O_i} \mathscr{T}(h(y))dy \right\|^2} \mu^\varepsilon(dh)},$$

where $\mu^\varepsilon = P \circ (X^\varepsilon)^{-1} \in \mathscr{P}(\mathscr{D}^{k-1})$, is an r.c.p.d. of X^ε given $D = d$. It suffices to show that for all d and all continuous and bounded real functions f on $\mathscr{D}^{k-1}$,

$$-\varepsilon \log \int_{\mathscr{D}^{k-1}} e^{-\frac{1}{\varepsilon} f(v)} \Gamma^\varepsilon(dv \mid d) \tag{12.35}$$

converges to $\inf_{h\in\mathscr{D}^{k-1}}[f(h)+I_d(h)]$ as $\varepsilon\to 0$. Note that (12.35) can be expressed as

$$-\varepsilon\log\int_{\mathscr{D}^{k-1}} e^{-\frac{1}{\varepsilon}\left(f(h)+\frac{1}{2}\sum_{i=1}^n\left\|d_i-\int_{O_i}\mathscr{T}(h(y))dy\right\|^2\right)}\mu^\varepsilon(dh)$$
$$+\varepsilon\log\int_{\mathscr{D}^{k-1}} e^{-\frac{1}{\varepsilon}\left(\frac{1}{2}\sum_{i=1}^n\left\|d_i-\int_{O_i}\mathscr{T}(h(y))dy\right\|^2\right)}\mu^\varepsilon(dh). \tag{12.36}$$

From Proposition 12.21, we see that the first term converges to

$$\inf_{h\in\mathscr{D}^{k-1}}\left[\tilde{I}(h)+f(h)+\frac{1}{2}\sum_{i=1}^n\left\|d_i-\int_{O_i}\mathscr{T}(h(y))dy\right\|^2\right\}$$
$$=\inf_{h\in\mathscr{D}^{k-1}}\inf_{\vartheta\in\Psi_h}\left[f(h)+\frac{1}{2}\|\vartheta\|^2_{\mathscr{M}}+\frac{1}{2}\int_O\|\mathscr{T}(h(y))-Y_d(y)\|^2dy\right]$$
$$=\inf_{h\in\mathscr{D}^{k-1}}[f(h)+\hat{J}_d(h)],$$

where the last equality is a consequence of (12.31) and (12.32). Taking $f=0$ in the last display, we see that the second term in (12.36) converges to $-\lambda_d$. This proves the result. □

12.5 Notes

General references for stochastic flows are [17, 125, 164, 178]. The properties of Sobolev spaces used in this chapter can be found in [1].

Variational mappings on path space for problems of image analysis are now a widely used tool. The first paper to formulate and rigorously analyze such variational problems is [99], although, as noted in this chapter, [99] does not give a precise Bayesian interpretation to the variational problem. This is carried out in [44], on which this chapter is based. Large deviations for finite dimensional stochastic flows are used for an asymptotic analysis of small noise finite dimensional anticipative SDEs in [202], and of finite dimensional diffusions generated by εL_1+L_2, where L_1, L_2 are two second-order differential operators, in [5]. Analogous problems for infinite dimensional models can be treated using the large deviation principle established here.

Chapter 13
Models with Special Features

13.1 Introduction

Chapters 8 through 12 considered representations in continuous time and their application to large and moderate deviation analyses of finite and infinite dimensional systems described by stochastic differential equations. In this chapter we complete our study of continuous time processes by considering additional problems with features that benefit from a somewhat different use of the representations and/or weak convergence arguments.

The first section gives an example from the important class of problems with "discontinuous statistics." The large deviation analysis of queueing networks and related systems generally fall into this category. Though in some instances such problems can be handled using the Contraction Principle (Theorem 1.16) and mappings on path space, the example presented here (the *weighted serve-the-longest* queueing system), cannot be treated in such a simple way, and a detailed weak convergence analysis of what happens in a neighborhood of the places where discontinuities occur is required.

The second section gives large and moderate deviation analyses for continuous time pure jump processes. Processes of this sort can in principle be represented as solutions of stochastic differential equations driven by Poisson random measures. However, when this is done the driving measures are simply "thinned" to produce the desired (possibly state dependent) jump rate for a given jump type. In particular, the points of each Poisson random measure play exactly the same role, which is just to indicate that a jump occurs. This is in contrast to the general SDE model of Chap. 10, where points that correspond to different locations of the spatial variable [i.e., x in $N(dt \times dx)$] in general correspond to different types of jumps. Owing to the homogeneous role played by the jumps corresponding to different locations of the spatial variable, a more efficient representation and analysis are possible, which is the main point of the section. Since many of the arguments of this section parallel others already given, in some places we present only the main steps, and point to these analogous proofs for more details.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_13

13.2 A Model with Discontinuous Statistics-Weighted Serve-the-Longest Queue

The aim of this section is to indicate some of the subtleties encountered when using weak convergence and representations to study the large deviation properties of processes with what we call "discontinuous statistics." In the Markovian setting these are processes for which the generator, even when applied to a smooth function, is not continuous in the state variable. While there are in general significant differences in the analysis of both upper and lower bounds when compared to more regular processes, we focus here on a model for which the main differences occur in the large deviation lower bound. This model will also be used in later chapters as an example for which accelerated Monte Carlo methods can be developed.

Consider a single server that must serve multiple queues of customers from different classes (Fig. 13.1). A common service discipline in this situation is the serve-the-longest queue policy, in which the longest queue is given priority. Here we consider a natural generalization of this discipline, namely, the weighted serve-the-longest queuing (WSLQ) policy. Under WSLQ, each queue length is multiplied by a constant to determine a "score" for that queue, and the queue with the largest score is granted priority. Such service policies are more appropriate than serve-the-longest policy when the different arrival queues or customer classes have different requirements or statistical properties. For example, if there is a finite queueing capacity to be split among the different classes, one may want to choose the partition and the weighting constants in order to optimize a certain performance measure.

Because WSLQ is a frequently proposed discipline for queueing models in communication problems, a large deviations analysis of this protocol can be useful [260]. However, service policies such as WSLQ are not smooth functions of the system state, and thus lead to multidimensional stochastic processes with discontinuous statistics. It is worthwhile to first explain why the large deviation analysis of these systems is difficult. As suggested by the results of [98], a large deviation upper bound can often be established using the same basic arguments as in either Chaps. 3, 4 or 10, so long as one defines the local rate function $L(x, \beta)$ to be the lower semicontinuous regularization of the local rate function as it would usually be defined on a pointwise (in x) basis, e.g., as in (4.5). However, this upper bound is generally *not* tight, even for the very simple situation of two regions of constant statistical behavior separated by a hyperplane of codimension one [95, Chap. 7].

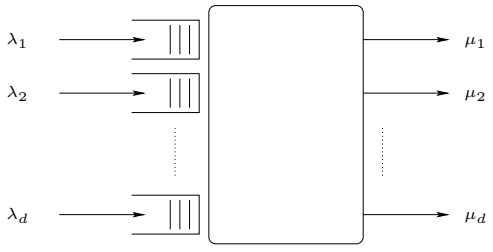

Fig. 13.1 WSLQ system

The reason for this gap is most easily identified by considering the corresponding lower bound. When proving a large deviation lower bound for a model with discontinuous statistics, it is necessary to analyze the probability that the process closely follows or tracks a constant velocity trajectory that lies on the interface of two or more regions of smooth statistical behavior. For this one has to consider all controls in these different regions that lead to the desired tracking behavior. The thorny issue is how to characterize such controls. In the case of two regions [95, Chap. 7], this can be done in a satisfactory fashion and it turns out that the large deviation rate function is a modified version of the upper bound in [98]. The modification is made to explicitly include certain "stability about the interface" conditions, and part of the reason that everything works out nicely in the setup of [95] is that these stability conditions can be easily characterized.

A key observation that makes possible a large deviations analysis for WSLQ is that for this model, the required stability conditions are *implicitly* built into the upper bound rate function obtained by lower semicontinuous regularization. More precisely, it can be shown that in the lower bound analysis one can restrict, a priori, to a class of controls for which the stability conditions are automatically implied (see Sect. 13.2.5.3 for additional discussion). Thus while it is true that the upper bound of [98] is not tight in general, it is so in this case due to the structural properties of WSLQ policy.

The study of the large deviation properties of WSLQ is partly motivated by the problem of estimating buffer overflow probabilities for stable WSLQ systems using either importance sampling or particle splitting (here we mean stable in a standard sense, such as positive recurrence, rather than stability about the interface). It turns out that the simple form of the large deviation local rate function as exhibited in (13.2) and (13.3) is helpful in constructing simple and asymptotically optimal importance sampling schemes, a topic to be discussed on the last part of this book. A WSLQ example is given in Sect. 17.4.

This section is organized as follows. In Sect. 13.2.1, we introduce the single server system with WSLQ policy. In Sect. 13.2.2, we state the main result, whose proof is presented in Sect. 13.2.3 (upper bound) and Sect. 13.2.5 (lower bound).

13.2.1 Problem Formulation

Consider a server with d customer classes, where customers of class i arrive according to a Poisson process with rate λ_i and are buffered at queue i for $i = 1, \ldots, d$. The service time for a customer of class i is exponentially distributed with rate μ_i.

The service policy is determined according to the WSLQ discipline, which can be described as follows. Let c_i be the weight associated with class i. If the size of queue i is q_i, then the "score" of queue i is defined as $c_i q_i$, and service priority will be given to the queue with the maximal score. When there are multiple queues with the maximal score, the assignment of priority among these queues can be arbitrary—the choice

is unimportant and will lead to the same rate function. We adopt the convention that when there are ties, the priority will be given to the queue with the largest index.

The system state at time t is the vector of queue lengths and is denoted by $Q(t) \doteq (Q_1(t), \ldots, Q_d(t))$. Then Q is a continuous time pure jump Markov process whose possible jumps belong to the set

$$\mathscr{V} \doteq \{\pm e_1, \pm e_2, \ldots, \pm e_d\}.$$

For $v \in \mathscr{V}$, let $r(x; v)$ denote the jump intensity of process Q from state x to state $x + v$. Under the WSLQ discipline, these jump intensities are as follows. For $x = (x_1, \ldots, x_d) \in \mathbb{R}^d_+$ and $x \neq 0$, let $\pi(x)$ denote the indices of queues that have the maximal score, that is,

$$\pi(x) \doteq \left\{1 \leq i \leq d : c_i x_i = \max_j c_j x_j\right\}.$$

Then

$$r(x; v) = \begin{cases} \lambda_i, & \text{if } v = e_i \text{ and } i = 1, \ldots, d, \\ \mu_i, & \text{if } v = -e_i \text{ where } i = \max \pi(x), \\ 0, & \text{otherwise.} \end{cases}$$

For $x = 0$, there is no service and the jump intensities are

$$r(0; v) = \begin{cases} \lambda_i, & \text{if } v = e_i \text{ and } i = 1, \ldots, d, \\ 0, & \text{otherwise.} \end{cases}$$

We also set

$$\pi(0) \doteq \{0, 1, 2, \ldots, d\}.$$

An illustrative figure for the case of two queues in Fig. 13.2.

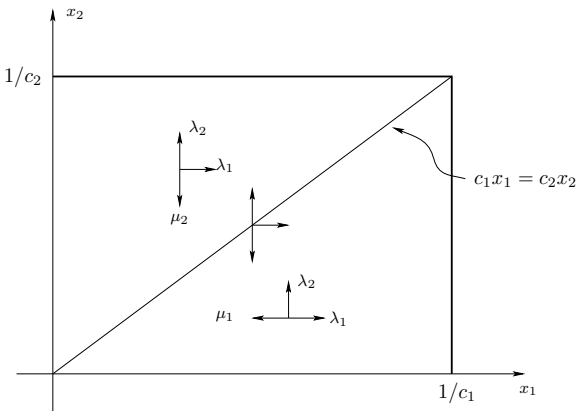

Fig. 13.2 System dynamics for $d = 2$

Remark 13.1 It is not difficult to see that jump distributions and jump intensities stay constant in the regions where $\pi(\cdot)$ is constant. Discontinuities in jump distributions occur when $\pi(\cdot)$ changes, and every x with $|\pi(x)| \geq 2$ (i.e., when there is a tie) is in fact a discontinuous point. Therefore, we have various discontinuity interfaces with different dimensions. For example, for every subset $A \subset \{1, 2, \ldots, d\}$ with $|A| \geq 2$ or $A = \{0, 1, 2, \ldots, d\}$, the set $\{x \in \mathbb{R}^d_+ : \pi(x) = A\}$ defines an interface with dimension $d - |A| + 1$.

Remark 13.2 The definition of $\pi(0)$ is introduced to cope with the discontinuous dynamics at the origin. Note that with this definition, $\pi(x)$ can only be $\{0, 1, 2, \ldots, d\}$ if $x = 0$ or a subset of $\{1, 2, \ldots, d\}$ if $x \neq 0$.

Remark 13.3 A useful observation is that π is *upper semicontinuous* as a set-valued function. That is, for any $x \in \mathbb{R}^d_+$, $\pi(y) \subset \pi(x)$ for all y in a small neighborhood of x.

13.2.2 Form of the Rate Function and Statement of the Laplace Principle

In order to state a large deviation principle, we fix $T \in (0, \infty)$, and for each $n \in \mathbb{N}$ let $\{X^n(t)\}_{t\in[0,T]}$ be the scaled process defined by

$$X^n(t) \doteq \frac{1}{n} Q(nt).$$

Then X^n is a continuous time Markov process with generator

$$\mathscr{L}^n f(x) \doteq n \sum_{v\in\mathscr{V}} r(x; v) \left[f\left(x + v/n\right) - f(x) \right].$$

The processes $\{X^n\}$ take values in the space of right-continuous functions with left limits $\mathscr{D}([0, T] : \mathbb{R}^d)$, which is endowed with the usual Skorohod metric and is a Polish space [24, Chapter 3].

For each $i = 1, \ldots, d$, let $H^{(i)}$ be the convex function given by

$$H^{(i)}(\alpha) \doteq \mu_i(e^{-\alpha_i} - 1) + \sum_{j=1}^{d} \lambda_j(e^{\alpha_j} - 1),$$

for all $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^d$. We also define

$$H^{(0)}(\alpha) \doteq \sum_{j=1}^{d} \lambda_j(e^{\alpha_j} - 1)$$

for $\alpha \in \mathbb{R}^d$.

For each nonempty subset $A \subset \{1, \ldots, d\}$ or $A = \{0, 1, \ldots, d\}$, let L^A be the Legendre transform of $H^A \doteq \max_{i \in A} H^{(i)}$, that is,

$$L^A(\beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \left[\langle \alpha, \beta \rangle - \max_{i \in A} H^{(i)}(\alpha) \right]$$

for each $\beta \in \mathbb{R}^d$. Clearly, L^A is convex and nonnegative. When A is a singleton $\{i\}$, we write L^A as $L^{(i)}$. Recall the function $\ell(x) \doteq x \log x - x + 1$ for $x \geq 0$. We next state some variational representations that are based on convex duality.

Lemma 13.4 *Given $\beta \in \mathbb{R}^d$, the following representation for $L^{(i)}(\beta)$ holds. For each $i = 1, 2, \ldots, d$*

$$L^{(i)}(\beta) = \inf \left[\sum_{j=1}^{d} \lambda_j \ell(\bar{\lambda}_j / \lambda_j) + \mu_i \ell(c) : \sum_{j=1}^{d} \bar{\lambda}_j / e_j - \mu_i c e_i = \beta \right], \tag{13.1}$$

and for $i = 0$

$$L^{(0)}(\beta) = \inf \left[\sum_{j=1}^{d} \lambda_j \ell\left(\bar{\lambda}_j / \lambda_j\right) : \sum_{j=1}^{d} \bar{\lambda}_j e_j = \beta \right].$$

In every case the infimum is attained.

Proof For $\lambda > 0$ and $v \in \mathbb{R}^d$, straightforward calculation shows that the Legendre transform for the convex function

$$h(\alpha) \doteq \lambda \left(e^{\langle \alpha, v \rangle} - 1 \right), \quad \alpha \in \mathbb{R}^d$$

is

$$h^*(\beta) \doteq \begin{cases} \lambda \ell(\bar{\lambda}/\lambda), & \text{if } \beta = \bar{\lambda} v \text{ for some } \bar{\lambda} \in \mathbb{R}_+ \\ \infty, & \text{otherwise} \end{cases}, \quad \beta \in \mathbb{R}^d.$$

The representation for $L^{(i)}$ then follows directly from [97, Corollary D.4.2]. The fact that the infimum is attained follows easily from the growth properties and lower semicontinuity of ℓ, and the proof is omitted. □

Note that we can also interpret $L^{(i)}$ as the local rate function for the jump Markov process with generator

$$n\mu_i \left[f\left(x - e_i/n\right) - f(x) \right] + n \sum_{i=1}^{d} \lambda_i \left[f\left(x + e_i/n\right) - f(x) \right],$$

see (10.4). A useful representation of L^A [97, Corollary D.4.3] is

$$L^A(\beta) = \inf \left[\sum_{i \in A} \rho^i L^{(i)}(\beta^i) \right], \tag{13.2}$$

where the infimum is taken over all $\{(\rho^i, \beta^i) : i \in A\}$ such that

$$\rho^i \geq 0, \quad \sum_{i \in A} \rho^i = 1, \quad \sum_{i \in A} \rho^i \beta^i = \beta. \tag{13.3}$$

An alternative representation for $L^A(\beta)$ can be given in terms of the relevant possible jump types.

Lemma 13.5 *Given $\beta \in \mathbb{R}^d$, we have the following representation for $L^A(\beta)$. If $A \subset \{1, 2, \ldots, d\}$ is nonempty then*

$$L^A(\beta) = \inf \left[\sum_{i \in A} \rho^i \mu_i \ell\left(\bar{\mu}_i / \mu_i\right) + \sum_{j=1}^{d} \lambda_j \ell\left(\bar{\lambda}_j / \lambda_j\right) \right],$$

where the infimum is taken over all collections of $(\rho^i, \bar{\mu}_i, \bar{\lambda}_j)$ such that

$$\rho^i \geq 0, \quad \sum_{i \in A} \rho^i = 1, \quad -\sum_{i \in A} \rho^i \bar{\mu}_i e_i + \sum_{j=1}^{d} \bar{\lambda}_j e_j = \beta. \tag{13.4}$$

If $A = \{0, 1, 2, \ldots, d\}$ then

$$L^A(\beta) = \inf \left[\sum_{i=1}^{d} \rho^i \mu_i \ell\left(\bar{\mu}_i / \mu_i\right) + \sum_{j=1}^{d} \lambda_j \ell\left(\bar{\lambda}_j / \lambda_j\right) \right],$$

where the infimum is taken over all collections of $(\rho^i, \bar{\mu}_i, \bar{\lambda}_j)$ such that

$$\rho^i \geq 0, \quad \sum_{i=0}^{d} \rho^i = 1, \quad -\sum_{i=1}^{d} \rho^i \bar{\mu}_i e_i + \sum_{j=1}^{d} \bar{\lambda}_j e_j = \beta.$$

Proof We only present the proof for the first claim, since the second is similar. Using Lemma 13.4 and Eqs. (13.2)–(13.3), we have

$$L^A(\beta) = \inf \left[\sum_{i \in A} \rho^i \left\{ \mu_i \ell\left(\bar{\mu}_i / \mu_i\right) + \sum_{j=1}^{d} \lambda_j \ell\left(\bar{\lambda}_j^i / \lambda_j\right) \right\} \right],$$

where the infimum is taken over all $(\rho^i, \bar{\mu}_i, \bar{\lambda}^i_j)$ such that

$$\rho^i \geq 0, \quad \sum_{i\in A} \rho^i = 1, \quad \sum_{i\in A} \rho^i \left[-\bar{\mu}_i e_i + \sum_{j=1}^{d} \bar{\lambda}^i_j e_j \right] = \beta. \tag{13.5}$$

Abusing notation, let $\bar{\lambda}_j \doteq \sum_{i\in A} \rho^i \bar{\lambda}^i_j$ for $j = 1, 2, \ldots, d$. Using (13.5), the collection $(\rho^i, \bar{\mu}_i, \bar{\lambda}_j)$ satisfies the constraints (13.4). Observe that by the convexity of ℓ

$$\sum_{i\in A} \rho^i \sum_{j=1}^{d} \lambda_j \ell\left(\bar{\lambda}^i_j/\lambda_j\right) = \sum_{j=1}^{d} \lambda_j \sum_{i\in A} \rho^i \ell\left(\bar{\lambda}^i_j/\lambda_j\right) \geq \sum_{j=1}^{d} \lambda_j \ell\left(\bar{\lambda}_j/\lambda_j\right),$$

with equality if $\bar{\lambda}^i_j = \bar{\lambda}^{i'}_j$ for every $i, i' \in A$ and $j = 1, \ldots, d$. The first claim of Lemma 13.5 then follows. □

Remark 13.6 The representation of L^A in Lemma 13.5 remains valid if we further constrain ρ^i, $\bar{\mu}_i$, and $\bar{\lambda}_i$ to be strictly positive for every $i \in A$. This is an easy consequence of the fact that ℓ is finite and continuous on the interval $[0, \infty)$.

Remark 13.7 Given any nonempty strict subset $A \subset \{1, 2, \ldots, d\}$ and any $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$, $L^A(\beta)$ is finite if and only if $\beta_j \geq 0$ for all $j \notin A$. In particular, for $A = \{1, 2, \ldots, d\}$ or $\{0, 1, 2, \ldots, d\}$, $L^A(\beta)$ is finite for every $\beta \in \mathbb{R}^d$.

For $x \in \mathbb{R}^d_+$ let $L(x, \beta) \doteq L^{\pi(x)}(\beta)$. Note that if $A \subset B$ then $L^A(\beta) \geq L^B(\beta)$ for all $\beta \in \mathbb{R}^d$, and therefore upper semicontinuity of π implies $L(\cdot, \cdot)$ is lower semicontinuous. We can now define the process level rate function. For $x \in \mathbb{R}^d_+$, define $I_x : \mathscr{D}([0, T] : \mathbb{R}^d) \to [0, \infty]$ by

$$I_x(\psi) \doteq \int_0^T L(\psi(t), \dot{\psi}(t))dt \tag{13.6}$$

if $\psi \in \mathscr{AC}_x([0, T] : \mathbb{R}^d_+)$, and otherwise set $I_x(\psi) \doteq \infty$.

Let E_{x_n} denote the expectation conditioned on $X^n(0) = x_n$. Note that the only initial conditions that are meaningful for this model must have components of the form j/n with $j \in \mathbb{N}_0$. Hence varying initial conditions are needed in the statement of any sort of Laplace principle. Recall the definition of a uniform LDP from Definition 1.11. In the current setting where only certain forms of initial conditions are meaningful, this definition needs to be slightly modified in that we replace K therein by $K_n \doteq K \cap (\mathbb{N}_0/n)^d$. With this modified definition, the following result states a uniform Laplace principle on compacts.

Theorem 13.8 *The sequence of processes $\{X^n\}_{n\in\mathbb{N}}$ satisfies the Laplace principle with rate function I_x uniformly on compacts. Thus for any sequence $\{x_n\} \subset (\mathbb{N}_0/n)^d$ such that $x_n \to x$ and any bounded continuous function $F : \mathscr{D}([0, T] : \mathbb{R}^d) \to \mathbb{R}$*

$$\lim_{n\to\infty} -\frac{1}{n}\log E_{x_n}\left\{\exp\left[-nF(X^n)\right]\right\} = \inf_{\psi\in\mathscr{D}([0,T]:\mathbb{R}^d)}\left[I_x(\psi)+F(\psi)\right],$$

and in addition I_x has compact level sets on compacts.

In the following sections, we follow our usual convention and give the proof for the case $T = 1$. The general case involves only notational differences. The proof of Theorem 13.8 is given in the next three subsections, which present the upper bound, properties of the rate function, and the lower bound, respectively.

13.2.3 Laplace Upper Bound

The proof of the Laplace upper bound, which is a lower bound on the variational representations, is very similar to analogous proofs given previously, and in particular to corresponding proofs in Chaps. 3, 4 and 10. As a consequence the proof of this bound is only sketched. As noted previously, for some problems with discontinuous statistics the analysis of the upper bound is much more involved than the one needed for the WSLQ model. The reason will be made more precise when we get to the lower bound.

The starting point for the analysis is a stochastic differential equation formulation of the process model X^n. Recall that $\pi(x)$ selects the queue that is given service when $X^n(t) = x$. Let N^n be a Poisson random measure with $\mathscr{X} \doteq \{\pm1, \pm2, \ldots, \pm d\}$, and with intensity measure $n\nu$, where $\nu(\{i\}) = \lambda_i, \nu(\{-i\}) = \mu_i\,, i = 1, \ldots, d$. Then the SDE formulation for X^n is given by

$$X^n(t) = x_n + \frac{1}{n}\sum_{i=1}^{d}\left(e_i N^n([0,t]\times\{i\}) - \int_{[0,t]} 1_{\{i=\max\pi(X^n(s-)),X^n(s-)\neq 0\}} e_i N^n(ds\times\{-i\})\right).$$

To simplify the notation we write $N_i^n(ds)$ and $N_{-i}^n(ds)$ for $N^n(ds\times\{i\})$ and $N^n(ds\times\{-i\})$, and similarly for the controlled versions. Likewise, the controls $\varphi(t,i)$ and $\varphi(t,-i)$ are also denoted by $\varphi_i(t)$ and $\varphi_{-i}(t)$. Recall that $\mathscr{X}_1 = \mathscr{X}\times[0,1]$ and $\Sigma(\mathscr{X}_1)$ denotes the space of all locally finite measures on $(\mathscr{X}_1, \mathscr{B}(\mathscr{X}_1))$. It is straightforward to show there is a measurable mapping $\mathscr{G}^n \in \mathscr{M}_b(\Sigma(\mathscr{X}_1))$ such that $X^n = \mathscr{G}^n(N^n)$, and hence by Theorem 8.12 and using the same argument as in Sect. 10.2.1 for the last equality, for any $F \in \mathscr{C}_b(\mathscr{D}([0,1]:\mathbb{R}^d))$

$$\begin{aligned} &-\frac{1}{n}\log E_{x_n}\exp\left\{-nF(X^n)\right\} \\ &\quad = -\frac{1}{n}\log E_{x_n}\exp\left\{-nF(\mathscr{G}^n(N^n))\right\} \end{aligned} \tag{13.7}$$

$$= \inf_{\varphi\in\mathscr{A}_b} E_{x_n}\left[\int_{\mathscr{X}_1} \ell\left(\varphi(t,z)\right)\nu_1(dt\times dz) + F(\mathscr{G}^n(N^{n\varphi}))\right]$$
$$= \inf_{\varphi\in\mathscr{A}_b} E_{x_n}\left[\int_{\mathscr{X}_1} \ell\left(\varphi(t,z)\right)\nu_1(dt\times dz) + F(\bar{X}^n)\right],$$

where $\nu_1(dt\times\{\pm i\}) = \nu(\{\pm i\})dt$, the space $\mathscr{A}_b$ is defined as in Sect. 8.2.1, $N^{n\varphi}$ is a controlled PRM with intensity $n\varphi\nu_1$,

$$\bar{X}^n(t) = x_n + \frac{1}{n}\sum_{i=1}^{d}\left(\int_{[0,t]} e_i N_i^{n\varphi}(ds) - \int_{[0,t]} 1_{\{i=\max\pi(\bar{X}^n(s-)),\bar{X}^n(s-)\neq 0\}} e_i N_{-i}^{n\varphi}(ds)\right),$$

and

$$\int_{\mathscr{X}_1} \ell\left(\varphi(t,z)\right)\nu_1(dt\times dz) = \sum_{i=1}^{d}\int_{[0,1]}\left[\lambda_i\ell\left(\varphi_i(t)\right) + \mu_i\ell\left(\varphi_{-i}(t)\right)\right]dt.$$

Thus the intensity of each jump type is perturbed by a corresponding control $\varphi_{\pm i}(t)$, with the usual cost in terms of ℓ.

To prove the variational lower bound we need to establish

$$\liminf_{n\to\infty}\inf_{\varphi\in\mathscr{A}_b} E_{x_n}\left[\int_{\mathscr{X}_1} \ell\left(\varphi(t,z)\right)\nu_1(dt\times dz) + F(\bar{X}^n)\right] \geq \inf_{\psi\in\mathscr{D}([0,1]:\mathbb{R}^d)}\left[I_x(\psi) + F(\psi)\right], \tag{13.8}$$

with I_x given by (13.6). The proof follows the lines of those given previously. Since F is bounded we can restrict to a class of controls which are tight and for which the costs are uniformly bounded in n. Let $\{\varphi^n\}$ be such a sequence of controls. It is enough to establish (13.8) along any convergent subsequence, again denoted by n. Decomposing $\bar{X}^n$ into a bounded variation part plus a martingale, we have $\bar{X}^n(t) = B^n(t) + M^n(t)$, where

$$B^n(t) = x_n + \sum_{i=1}^{d}\left(\int_{[0,t]}\lambda_i e_i\varphi_i^n(s)ds - \int_{[0,t]} 1_{\{i=\max\pi(\bar{X}^n(s-)),\bar{X}^n(s-)\neq 0\}}\mu_i e_i\varphi_{-i}^n(s)ds\right)$$

and where $M^n\to 0$ in $\mathscr{D}([0,1]:\mathbb{R}^d)$ in probability as $n\to\infty$. Using the definition $L(x,\beta)\doteq L^{\pi(x)}(\beta)$, (13.1) and (13.2), we have

$$\sum_{i=1}^{d}\left[\lambda_i\ell\left(\varphi_i^n(t)\right) + \mu_i\ell\left(\varphi_{-i}^n(t)\right)\right]$$

$$\geq \sum_{i=1}^{d} \left[\lambda_i \ell\left(\varphi_i^n(t)\right) + 1_{\{i=\max \pi(\bar{X}^n(t)), \bar{X}^n(t)\neq 0\}} \mu_i \ell(\varphi_{-i}^n(t))\right]$$
$$\geq L(\bar{X}^n(t), \dot{B}^n(t)). \tag{13.9}$$

Define random probability measures $\bar{\mu}^n$ on $\mathbb{R}^d \times [0, 1]$ according to

$$\bar{\mu}^n(A \times C) \doteq \int_C 1_A(\dot{B}^n(t))dt.$$

The superlinearity of ℓ implies tightness and uniform integrability of $\{\bar{\mu}^n\}$, and without loss of generality we assume that $\bar{\mu}^n$ converges weakly to $\bar{\mu}$ along the subsequence. The same argument as that used for Lemma 4.12 shows that if $(\bar{\mu}, \bar{X}, B)$ is the weak limit of $\{(\bar{\mu}^n, \bar{X}^n, B^n)\}$, then $\bar{X} = B$ w.p.1 and

$$\bar{X}(t) = x + \int_{\mathbb{R}^d \times [0,t]} y\bar{\mu}(dy \times ds) = x + \int_{\mathbb{R}^d \times [0,t]} y\bar{\mu}(dy|s)ds.$$

Recalling that $L(\cdot, \cdot)$ is lower semicontinuous, we obtain

$$\begin{aligned}
&\liminf_{n\to\infty} E_{x_n}\left[\int_{\mathscr{X}_1} \ell\left(\varphi(t, z)\right) \nu_1(dt \times dz) + F(\bar{X}^n)\right] \\
&\geq \liminf_{n\to\infty} E_{x_n}\left[\int_0^1 L(\bar{X}^n(t), \dot{S}^n(t))dt + F(\bar{X}^n)\right] \\
&= \liminf_{n\to\infty} E_{x_n}\left[\int_0^1 \int_{\mathbb{R}^d} L(\bar{X}^n(t), y)\bar{\mu}^n(dy|t)dt + F(\bar{X}^n)\right] \\
&\geq E_x\left[\int_0^1 \int_{\mathbb{R}^d} L(\bar{X}(t), y)\bar{\mu}(dy|t)dt + F(\bar{X})\right] \\
&\geq E_x\left[\int_0^1 L(\bar{X}(t), \dot{\bar{X}}(t))dt + F(\bar{X})\right] \\
&\geq \inf_{\psi\in\mathscr{D}([0,1]:\mathbb{R}^d)} \left[I_x(\psi) + F(\psi)\right],
\end{aligned}$$

where the first inequality is due to (13.9), the equality uses the definition of $\bar{\mu}^n$, Fatou's lemma and the lower semicontinuity of $L(x, \beta)$ imply the second inequality, Jensen's inequality gives the third, and the last follows since the corresponding bound without the expectation holds w.p.1. This completes the proof of the uniform Laplace upper bound. □

13.2.4 Properties of the Rate Function

The following lemma says that the rate function I_x has compact level sets on compacts.

Lemma 13.9 *For any compact set $C \subset \mathbb{R}^d$ and $M \in (0, \infty)$,*

$$S \doteq \cup_{x \in C} \{\psi : I_x(\psi) \leq M\}$$

is compact in $\mathscr{D}([0, 1] : \mathbb{R}^d)$.

Proof The argument is almost exactly the same as that of Theorem 4.13. For each $\alpha \in \mathbb{R}^d$ let

$$\bar{H}(\alpha) \doteq \max_{i \in \pi(0)} H^{(i)}(\alpha) < \infty.$$

Owing to the last inequality, $\bar{L}(\beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \left[\langle \alpha, \beta \rangle - \bar{H}(\alpha)\right]$ is superlinear in the sense that $\lim_{c \to \infty} \inf_{\{\beta : \|\beta\| \geq c\}} \bar{L}(\beta)/\|\beta\| = \infty$ [see the proof of part (c) of Lemma 4.14].

Let $\{\psi_j\}_{j \in \mathbb{N}} \subset S$ be given with the property that $I_{x_j}(\psi_j) \leq M$ for all $j \in \mathbb{N}$, with $x_j = \psi_j(0) \in C$. It suffices to show that $\{\psi_j\}$ is precompact, and that if ψ is the limit along any subsequence and $x = \psi(0)$, then $I_x(\psi) \leq M$. Define a probability measure γ^j on $\mathbb{R}^d \times [0, 1]$ by

$$\gamma^j(A \times B) = \int_B \delta_{\dot{\psi}_j(t)}(A) dt, \quad A \in \mathscr{B}(\mathbb{R}^d), B \in \mathscr{B}([0, 1]),$$

so that

$$\psi_j(t) = x_j + \int_{\mathbb{R}^d \times [0,t]} \beta \gamma^j(d\beta \times ds).$$

We claim that $\{\gamma^j\}$ is tight and uniformly integrable. For $c \in (0, \infty)$ let $g(c) \doteq \inf_{\{\beta : \|\beta\| \geq c\}} \bar{L}(\beta)/\|\beta\|$. Since $\bar{H}(\alpha) \geq H^A(\alpha)$ for all A and $\alpha \in \mathbb{R}^d$, $L(x, \beta) \geq \bar{L}(\beta)$ for all $(x, \beta) \in \mathbb{R}^d_+ \times \mathbb{R}^d$. It follows that for $c \in (0, \infty)$

$$\begin{aligned} M &\geq I_{x_j}(\psi_j) \\ &\geq \int_{\mathbb{R}^d \times [0,1]} \bar{L}(\beta) \gamma^j(d\beta \times ds) \\ &\geq g(c) \int_{\mathbb{R}^d \times [0,1]} \|\beta\| \, 1_{\{\|\beta\| \geq c\}} \gamma^j(d\beta \times ds). \end{aligned}$$

Since $g(c) \to \infty$ as $c \to \infty$

$$\lim_{c \to \infty} \sup_{j \in \mathbb{N}} \int_{\mathbb{R}^d \times [0,1]} \|\beta\| \, 1_{\{\|\beta\| \geq c\}} \gamma^j(d\beta \times ds) = 0,$$

and therefore $\{\gamma^j\}$ is tight and uniformly integrable.

If $0 \le s \le t \le 1$ then

$$\begin{aligned}\left\|\psi_j(t) - \psi_j(s)\right\| &\le \int_{\mathbb{R}^d \times [s,t]} \|\beta\| \, 1_{\{\|\beta\| \ge c\}} \gamma^j(d\beta \times dr) + c(t-s) \\ &\le \frac{M}{g(c)} + c(t-s),\end{aligned}$$

which together with precompactness of $\{\psi^j(0)\}$ shows that $\{\psi^j\}$ is precompact. Since $\{\gamma^j\}$ is tight there is a subsequence (again denoted by j) such that γ^j converges weakly to $\gamma \in \mathscr{P}(\mathbb{R}^d \times [0,1])$ and ψ^j converges uniformly to $\psi \in \mathscr{C}([0,1] : \mathbb{R}^d_+)$ as $j \to \infty$. Since the second marginal γ^j is Lebesgue measure the same is true of γ, and therefore $\gamma(d\beta \times dt)$ can be decomposed in the form $\gamma(d\beta|t)dt$, where $\gamma(d\beta|t)$ is a stochastic kernel on $\mathbb{R}^d$ given $[0,1]$. Using the uniform integrability we obtain

$$\psi(t) = x + \int_{\mathbb{R}^d \times [0,t]} \beta \gamma(d\beta|s)ds$$

by passing to the limit where $x \in C$. Hence along this subsequence

$$\begin{aligned}M &\ge \liminf_{j\to\infty} \int_0^1 L(\psi_j(t), \dot{\psi}_j(t))dt \\ &= \liminf_{j\to\infty} \int_{\mathbb{R}^d \times [0,1]} L(\psi_j(t), \beta)\gamma^j(d\beta \times dt) \\ &\ge \int_{\mathbb{R}^d \times [0,1]} L(\psi(t), \beta)\gamma(d\beta|t)dt \\ &\ge \int_{[0,1]} L(\psi(t), \dot{\psi}(t))dt,\end{aligned}$$

where the second inequality uses Fatou's lemma and the lower semicontinuity of $L(\cdot,\cdot)$, and the last inequality uses Jensen's inequality. □

13.2.5 *Laplace Lower Bound*

In this section we prove that if $x_n \to x$ then

$$\limsup_{n\to\infty} -\frac{1}{n} \log E_{x_n}\left\{\exp\left[-nF(X^n)\right]\right\} \le \inf_{\psi \in \mathscr{D}([0,1]:\mathbb{R}^d)} [I_x(\psi) + F(\psi)]. \quad (13.10)$$

An important step in the proof of (13.10) uses the following approximation lemma. Let $\mathscr{N}$ be the collection of functions $\psi^* \in \mathscr{AC}([0,1] : \mathbb{R}^d_+)$ that satisfy the following conditions:

(a) $\dot{\psi}^*$ is piecewise constant with finitely many jumps on (0,1).

(b) If (t, s) is an interval on which $\dot{\psi}^*$ is constant, then $\pi(\psi^*(\cdot))$ remains the same on the interval (t, s).

Lemma 13.10 *Given any $\psi \in \mathscr{D}([0,1] : \mathbb{R}^d)$ such that $I_x(\psi) < \infty$ and any $\delta > 0$, there exists $\psi^* \in \mathscr{N}$ such that $\|\psi - \psi^*\|_\infty < \delta$ and $I_x(\psi^*) \le I_x(\psi) + \delta$.*

The proof of Lemma 13.10 is lengthy but relatively straightforward, and for a proof we refer to [104]. The main idea can be described as follows. Suppose we consider the analogous question of approximation for a problem where the rate function is of the same form $I_x(\psi) = \int_0^1 L(\psi, \dot{\psi})ds$, but with L defined on $\mathbb{R}^d \times \mathbb{R}^d$ and constant on each of the sets $\{x : x_1 = 0\}$, $\{x : x_1 > 0\}$ and $\{x : x_1 < 0\}$. We also assume, as is true here, that L is jointly lower semicontinuous and convex in β for each x. Then given an absolutely continuous trajectory ψ, we can decompose $[0, 1]$ according to the closed set $B \doteq \{t : \psi_1(t) = 0\}$, where $\psi_1(t)$ is the first component of $\psi(t)$. If $t \in B^c$ and if there is a sufficiently "large" interval $(a, b) \subset B^c$ with $t \in (a, b)$ we can replace ψ by its linear interpolation on $[a_t, b_t]$, where $a_t = \inf\{s \in [0, t] : s \in B^c\}$ and $b_t = \sup\{s \in [t, 1] : s \in B^c\}$. If the distance between ψ and its linear interpolation is too large we break the interval up into (a finite number of) smaller intervals to control this distance. After removing a finite collection of such "large" intervals, all remaining open (relative to $[0, 1]$) intervals are small, and have the property that their endpoints are either in B or of the form $\{0, 1\}$. We then again replace ψ by its piecewise linear interpolation on these intervals, with a controllably small distance between the original and its replacement. If one of the endpoints were from $\{0, 1\}$ we are done. If both are from B, then we use that, owing to the lower semicontinuity, $L(x, \cdot)$ is smaller if $x_1 = 0$ than if $x_1 \neq 0$. Since the trajectory starts on and ends in B we can use Jensen's inequality and the lower semicontinuity to argue that the cost of a trajectory that stays entirely in $\{x : x_1 = 0\}$ is smaller than the original, which completes the argument. The general case is more complicated since there the discontinuities occur across sets of different dimensions, but the main idea is the same, and one starts by approximating parts of the trajectories near sets of the lowest dimension and works up to those parts in sets of higher dimension.

Since F is continuous, by Lemma 13.10 it suffices to show that for any $x_n \to x$ and $\psi^* \in \mathscr{N}$,

$$\limsup_{n\to\infty} -\frac{1}{n} \log E_{x_n} \left\{ \exp\left[-nF(X^n)\right] \right\} \le I_x(\psi^*) + F(\psi^*). \tag{13.11}$$

Using the control representation (13.7), the Laplace principle lower bound (13.11) for $\psi^* \in \mathscr{N}$ follows if one can, for an arbitrarily fixed $\varepsilon > 0$, construct a sequence of controls $\{\varphi^n_{\pm i}\}$ such that

$$\begin{aligned} \limsup_{n\to\infty} E_{x_n} \left[\int_0^1 \sum_{i=1}^d \left[\lambda_i \ell\left(\varphi_i^n(t)\right) + \mu_i \ell\left(\varphi_{-i}^n(t)\right) \right] dt + F(\bar{X}^n) \right] \\ \le I_x(\psi^*) + F(\psi^*) + \varepsilon. \end{aligned} \tag{13.12}$$

The details of the construction will be carried out in the next section.

13.2.5.1 The Construction of Controls and the Cost

Fix $\varepsilon > 0$. We will use ψ^* to construct a control $\{\varphi^n_{\pm i}\}$ based on the representation of the local rate function as in Lemma 13.5. Since $\psi^* \in \mathscr{N}$, there exist $0 = t_0 < t_1 < \cdots < t_K = 1$ such that for every k there are β_k and A_k with $\dot{\psi}^*(t) = \beta_k$ and $\pi(\psi^*(t)) = A_k$ for all $t \in (t_k, t_{k+1})$. We start by defining a suitable collection of $\{(\rho^i_k, \bar{\mu}_{i,k}, \bar{\lambda}_{j,k}) : 0 \le i \le d, 1 \le j \le d\}$. We consider the following two cases.

Case 1. Suppose $A_k = \{0, 1, 2, \ldots, d\}$. Lemma 13.5 and Remark 13.6 imply the existence of a collection $\{(\rho^i_k, \bar{\mu}_{i,k}, \bar{\lambda}_{j,k}) : 0 \le i \le d, 1 \le j \le d\}$ such that $\bar{\mu}_{i,k} > 0$, $\bar{\lambda}_{i,k} > 0$ for all i and

$$\rho^i_k > 0, \quad \sum_{i=0}^{d} \rho^i_k = 1, \quad -\sum_{i=1}^{d} \rho^i_k \bar{\mu}_{i,k} e_i + \sum_{j=1}^{d} \bar{\lambda}_{j,k} e_j = \beta_k, \tag{13.13}$$

$$\sum_{i=1}^{d} \rho^i_k \mu_i \ell\left(\bar{\mu}_{i,k}/\mu_{i,k}\right) + \sum_{j=1}^{d} \lambda_j \ell\left(\bar{\lambda}_{j,k}/\lambda_{j,k}\right) \le L^{A_k}(\beta_k) + \varepsilon.$$

Case 2. Suppose $A_k \subset \{1, 2, \ldots, d\}$. According to Lemma 13.5 and Remark 13.6, for each k there exists a collection $\{(\rho^i_k, \bar{\mu}_{i,k}, \bar{\lambda}_{j,k}) : i \in A_k, 1 \le j \le d\}$ such that $\bar{\mu}_{i,k} > 0$, $\bar{\lambda}_{i,k} > 0$ for all $i \in A_k$ and

$$\rho^i_k > 0, \quad \sum_{i \in A_k} \rho^i_k = 1, \quad -\sum_{i \in A_k} \rho^i_k \bar{\mu}_{i,k} e_i + \sum_{j=1}^{d} \bar{\lambda}_{j,k} e_j = \beta_k, \tag{13.14}$$

$$\sum_{i \in A_k} \rho^i_k \mu_i \ell\left(\bar{\mu}_{i,k}/\mu_{i,k}\right) + \sum_{j=1}^{d} \lambda_j \ell\left(\bar{\lambda}_{j,k}/\lambda_{j,k}\right) \le L^{A_k}(\beta_k) + \varepsilon. \tag{13.15}$$

We extend the definition by letting $\rho^i_k \doteq 0$, $\bar{\mu}_{i,k} \doteq \mu_i$ for $i \notin A_k$ and $i \ne 0$, and letting $\rho^0_k \doteq 0$.

A feedback control $\bar{r}$ is defined as follows. For $t \in [t_k, t_{k+1})$, let

$$\bar{r}(x, t; v) = \begin{cases} \bar{\lambda}_{j,k}, & \text{if } v = e_j \text{ and } j = 1, \ldots, d, \\ \bar{\mu}_{j,k}, & \text{if } v = -e_j \text{ where } j = \max \pi(x) \text{ and } x \ne 0, \\ 0\ , & \text{otherwise.} \end{cases}$$

In other words, on time interval $[t_k, t_{k+1})$, the system has arrival rates $\{\bar{\lambda}_{1,k}, \ldots, \bar{\lambda}_{d,k}\}$ and service rates $\{\bar{\mu}_{1,k}, \ldots, \bar{\mu}_{d,k}\}$ under the control $\bar{r}$. If $\bar{X}^n(t)$ is the corresponding

controlled process with initial value $x_n = \lfloor nx \rfloor / n$, then the predictable controls $\{\varphi^n_{\pm i}\}$ are defined for $i \in \{1, \ldots, d\}$ by

$$\varphi^n_i(t) = \frac{\bar{r}(\bar{X}^n(t), t; e_i)}{r(\bar{X}^n(t); e_i)}, \quad \varphi^n_{-i}(t) = \frac{\bar{r}(\bar{X}^n(t), t; -e_i)}{r(\bar{X}^n(t); -e_i)}$$

where $0/0$ is taken to be 1. The corresponding running cost for $t \in [t_k, t_{k+1})$ is

$$\sum_{j=1}^{d} \mu_j \ell\left(\bar{\mu}_{j,k}/\mu_j\right) 1_{\{\max \pi(\bar{X}^n(t))=j, \bar{X}^n(t)\neq 0\}} + \sum_{j=1}^{d} \lambda_j \ell\left(\bar{\lambda}_{j,k}/\lambda_j\right) \tag{13.16}$$

$$= \sum_{i\in A_k, i\neq 0} \mu_i \ell\left(\bar{\mu}_{i,k}/\mu_i\right) 1_{\{\max \pi(\bar{X}^n(t))=i, \bar{X}^n(t)\neq 0\}} + \sum_{j=1}^{d} \lambda_j \ell\left(\bar{\lambda}_{j,k}/\lambda_j\right),$$

here the equality holds since $\bar{\mu}_{i,k} = \mu_i$ for $i \notin A_k$.

For future use, for each $i = 1, \ldots, d$ we define

$$\beta^i_k \doteq -\bar{\mu}_{i,k} e_i + \sum_{j=1}^{d} \bar{\lambda}_{j,k} e_j, \tag{13.17}$$

which is the law of large number limit of the velocity of the controlled process if queue of class i is served. Analogously, we also define (when none of the queues are being served)

$$\beta^0_k \doteq \sum_{j=1}^{d} \bar{\lambda}_{j,k} e_j. \tag{13.18}$$

13.2.5.2 Weak Convergence Analysis

In this section we characterize the limit processes. For the lower bound more information is needed regarding the weak limits. For each j, define random sub-probability measures $\{\gamma^n_j\}$ on $[0, 1]$ by

$$\gamma^n_j(B) \doteq \int_B 1_{\{\max \pi(\bar{X}^n(t))=j, \bar{X}^n(t)\neq 0\}} dt, \quad j = 1, 2, \ldots, d,$$
$$\gamma^n_0(B) \doteq \int_B 1_{\{\bar{X}^n(t)=0\}} dt,$$

for Borel subsets $B \subset [0, 1]$, and denote $\gamma^n \doteq (\gamma^n_0, \gamma^n_1, \ldots, \gamma^n_d)$. We also define the stochastic processes

$$B^n(t) \doteq x_n + \sum_{j=0}^{d} \left[\sum_{k=0}^{\kappa(t)-1} \beta_k^j \gamma_j^n([t_k, t_{k+1})) + \beta_{\kappa(t)}^j \gamma_j^n([t_{\kappa(t)}, t)) \right],$$

where $\kappa(t) \doteq \max\{0 \le k \le K : t_k \le t\}$ and $x_n = \lfloor nx \rfloor / n$.

Proposition 13.11 *Given any subsequence of $(\gamma^n, B^n, \bar{X}^n)$, there exist a subsubsequence, a collection of random measures $\gamma \doteq (\gamma_0, \gamma_1, \ldots, \gamma_d)$ on $[0, 1]$, and a continuous process $\bar{X}$ such that the following hold.*

(a) *The subsubsequence converges in distribution to $(\gamma, \bar{X}, \bar{X})$.*

(b) *With probability one, γ_j is absolutely continuous with respect to Lebesgue measure on $[0, 1]$, and the set of densities $\{h_j\}_{j=0}^d$ satisfy*

$$\sum_{j=0}^{d} h_j(t) = \sum_{j \in \pi(\bar{X}(t))} h_j(t) = 1$$

for almost every t.

(c) *With probability one, the process $\bar{X}$ satisfies*

$$\bar{X}(t) = x + \sum_{j=0}^{d} \left[\sum_{k=0}^{\kappa(t)-1} \beta_k^j \gamma_j([t_k, t_{k+1})) + \beta_{\kappa(t)}^j \gamma_j([t_{\kappa(t)}, t)) \right]$$

for every t. Therefore, $\bar{X}$ is absolutely continuous with derivative

$$\frac{d\bar{X}(t)}{dt} = \sum_{j=0}^{d} \beta_{\kappa(t)}^j h_j(t).$$

Proof The family of random measures $\{\gamma_j^n\}$ is contained in the compact set of sub-probability measures on $[0, 1]$ and is therefore tight. Furthermore, since $\{B^n\}$ is uniformly Lipschitz continuous, it takes values in a compact subset of $\mathscr{C}([0, 1] : \mathbb{R}^d)$, and therefore is also tight. We also observe that for every $\varepsilon > 0$

$$\lim_{n \to \infty} P(\|\bar{X}^n - B^n\|_\infty > \varepsilon) = 0, \tag{13.19}$$

which in turn implies that $\{\bar{X}^n\}$ is tight. Equation (13.19) holds since $\bar{X}^n - B^n$ is a martingale, and thus by Doob's maximal inequality [see (D.2)] and the uniform boundedness of the jump intensity $\bar{r}$,

$$P\left(\sup_{0 \le t \le 1} \|\bar{X}^n(t) - B^n(t)\| > \varepsilon \right) \le \frac{4}{\varepsilon^2} E\left[\|\bar{X}^n(1) - B^n(1)\|^2 \right] \to 0.$$

It follows that there exists a subsubsequence that converges weakly to say (γ, B, B), with $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_d)$. By the Skorohod representation theorem, we assume

without loss of generality that the convergence is almost sure convergence, and all random variables are defined on a probability space $(\bar{\Omega}, \bar{\mathscr{F}}, \bar{P})$.

Since the $\{\gamma_j^n\}$ are absolutely continuous with respect to Lebesgue measure on $[0, 1]$ with uniformly bounded densities in both n and t, the limit γ_j is also absolutely continuous. Furthermore, if we define the process $\bar{X}$ as in (c), this implies that, for every $t \in [0, 1]$, $B^n(t)$ converges to $\bar{X}(t)$ almost surely. Therefore, with probability one $B(t) = \bar{X}(t)$ for all rational $t \in [0, 1]$, and since both B and $\bar{X}$ are continuous, $B = \bar{X}$ with probability one.

It remains to show the two equalities of (b). Since $\sum_{j=0}^d \gamma_j^n$ equals Lebesgue measure for every n, we have $\sum_{j=0}^d h_j(t) = 1$ for almost every t. The proof of the second equality is similar to that of [97, Theorem 7.4.4(c)]. Consider an $\omega \in \bar{\Omega}$ such that $\bar{X}(t, \omega)$ is a continuous function of $t \in [0, 1]$, $\gamma^n(\omega) \Rightarrow \gamma(\omega)$, and such that $\bar{X}^n(\cdot, \omega)$ converges to $\bar{X}(\cdot, \omega)$ in the Skorohod metric and hence also in the supremum norm since $\bar{X}(\cdot, \omega)$ is continuous [97, Theorem A.6.5]. By the upper semicontinuity of $\pi(\cdot)$ [Remark 13.3], it follows that for any $t \in (0, 1)$ and $A \subset \{0, 1, \ldots, d\}$ such that $\pi(\bar{X}(t, \omega)) \subset A$, there exist an open interval (a, b) containing t and $N \in \mathbb{N}$ such that $\pi(\bar{X}^n(s, \omega)) \subset A$ for all $n \geq \mathbb{N}$ and $s \in (a, b)$. Therefore $\sum_{j \notin A} \gamma_j^n(\omega)((a, b)) = 0$ for all $n \geq \mathbb{N}$. Taking the limit as $n \to \infty$ we have $\sum_{j \notin A} \gamma_j(\omega)((a, b)) = 0$, which in turn implies that

$$\sum_{j \notin A} \gamma_j(\omega)(\{t \in (0, 1) : \pi(\bar{X}(t, \omega)) \subset A\}) = 0,$$

or equivalently,

$$\int_0^1 \sum_{j \notin A} h_j(t, \omega) 1_{\{\pi(\bar{X}(t,\omega)) \subset A\}} \, dt = 0.$$

We claim that this implies the desired equality for any such $\omega \in \bar{\Omega}$. Otherwise, there exists a subset $D \subset (0, 1)$ with positive Lebesgue measure such that for every $t \in D$,

$$\sum_{j \notin \pi(\bar{X}(t,\omega))} h_j(t, \omega) > 0.$$

Since $\pi(\bar{X}(t, \omega))$ can only take finitely many possible values, there exists a subset $\bar{A} \subset \{0, 1, \ldots, d\}$ such that the set

$$\bar{D} \doteq \{t \in D : \pi(\bar{X}(t, \omega)) = \bar{A}\}$$

has positive Lebesgue measure. It follows that

$$\int_0^1 \sum_{j \notin \bar{A}} h_j(t, \omega) 1_{\{\pi(\bar{X}(t,\omega)) \subset \bar{A}\}} \, dt \geq \int_{\bar{D}} \sum_{j \notin \bar{A}} h_j(t, \omega) 1_{\{\pi(\bar{X}(t,\omega)) \subset \bar{A}\}} \, dt$$

$$= \int_{\bar{D}} \sum_{j \notin \pi(\bar{X}(t,\omega))} h_j(t,\omega)\, dt$$
$$> 0,$$

a contradiction. This completes the proof of the claim and thus of the proposition. □

13.2.5.3 Stability Analysis

In this section we prove a key lemma in the analysis that identifies the weak limit $\bar{X}$ as ψ^*. The proof uses the implied "stability about the interface" in a crucial way.

We discuss the main idea behind this stability property before giving the detailed proof. For the large deviation analysis, it is important to analyze the probability that the process tracks a segment of trajectory that lies on an interface, say $\{x : \pi(x) = A\}$, with a constant velocity β. To this end, it is natural to use the change of measure induced by β through the local rate function L as described in Lemma 13.5. However, for general systems, this very natural construction does *not* guarantee that $\bar{X}$ will follow or track ψ. When this happens L is not the true local rate (although it is an upper bound local rate), and additional conditions must be added for this tracking to occur.

For the WSLQ system, a stability condition is not explicitly needed since it is implicitly and automatically built into the upper bound local rate function L. To simplify the discussion, we can assume that $\psi(t)$ lies in $\{x : \pi(x) = A\}$ with $\dot{\psi}(t) = \beta$ for $t \in [0, 1]$. Then $K = 1$, and we drop k from the notation. Consider a set of arrival and service rates $\bar{\lambda}_i$ and $\bar{\mu}_i$ and fractions ρ^i associated with β through (13.14). For β to keep ψ in the set it must be the case that $\{\rho^i\}$ is the (strictly positive) solution to the system of equations

$$c_i(\bar{\lambda}_i - \rho^i \bar{\mu}_i) = c_j(\bar{\lambda}_j - \rho^j \bar{\mu}_j), \quad i, j \in A, \text{ and } \sum_{i \in A} \rho^i = 1. \tag{13.20}$$

Recall that $\{h_i(t)\}$ is the limit of the fraction of time the process $\bar{X}^n$ spends in the region $\{x : \max \pi(x) = i\}$, and that according to Proposition 13.11 these proportions satisfy

$$\sum_{i \in \pi(\bar{X}(t))} h_i(t) = 1.$$

Therefore, to show $\bar{X}(t) = \psi(t)$ it is enough to show that $\pi(\bar{X}(t)) \equiv A$ and $h_i(t) \equiv \rho^i$ for $i \in A$. This can be verified, and the argument is based on the fact that for any non-empty subset $B \subset \{1, 2, \ldots, d\}$ and any b, the solution $\{x_i : i \in B\}$ to the system of equations

$$c_i(\bar{\lambda}_i - x_i \bar{\mu}_i) = c_j(\bar{\lambda}_j - x_j \bar{\mu}_j), \quad i, j \in B, \text{ and } \sum_{i \in B} x_i = \varsigma, \tag{13.21}$$

is unique and component-wise strictly increasing with respect to ς. Indeed, it is not hard to check that one solution is

$$x_i = \frac{\bar{\lambda}_i}{\bar{\mu}_i} - \frac{1}{c_i \bar{\mu}_i} \cdot \left(\sum_{j \in B} \frac{1}{c_j \bar{\mu}_j} \right)^{-1} \cdot \left(-\varsigma + \sum_{j \in B} \frac{\bar{\lambda}_j}{\bar{\mu}_j} \right).$$

If $\bar{x}_i$ is another solution and $\Delta x_i = \bar{x}_i - x_i$, then $c_i \bar{\mu}_i \Delta x_i = c_j \bar{\mu}_j \Delta x_j$ implies that all Δx_i have the same sign, and then $\sum_{i \in B} \Delta x_i = 0$ implies they are all equal to zero.

Now suppose $\pi(\bar{X}(t)) = B$ on some time interval (a, b), where B is a strict subset of A. It is not difficult to see that $\{h_i(t) : i \in B\}$ is a solution to Eq. (13.21) because, by Proposition 13.11,

$$\frac{d}{dt}(\bar{X}(t))_i = \bar{\lambda}_i - h_i(t)\bar{\mu}_i, \quad \text{for } i \in B.$$

Thus both $\{h_i(t), i \in B\}$ and $\{\rho^i, i \in B\}$ are solutions. Since

$$\sum_{i \in B} \rho^i = 1 - \sum_{i \in A \setminus B} \rho^i < 1,$$

the monotonicity of x_i in ς implies $h_i(t) > \rho^i$ for all $i \in B$. We also have $h_j(t) < \rho^j$ for all $j \in A \setminus B$. Thus by comparing with $0 = \frac{d}{dt}\left[c_i(\psi(t))_i - c_j(\psi(t))_j\right]$, we find that for $i \in B$ and $j \in A \setminus B$,

$$\frac{d}{dt}\left[c_i(\bar{X}(t))_i - c_j(\bar{X}(t))_j\right] < 0. \tag{13.22}$$

Thus the differences between weighted queue lengths grow smaller, and the state is "pushed" towards the interface A. This can be used to prove by contradiction that $A \subset \pi(\bar{X}(t))$. The other direction $\pi(\bar{X}(t)) \subset A$ can be shown similarly. Once $\pi(\bar{X}(t)) = A$ is shown, $h_i(t) = \rho^i$ follows immediately from the uniqueness of the solution to Eq. (13.21).

Lemma 13.12 *Let $(\gamma, \bar{X}, \bar{X})$ be a limit of any weakly converging subsubsequence $(\gamma^n, B^n, \bar{X}^n)$ as in Proposition 13.11. Then with probability one, $\bar{X}(t) = \psi^*(t)$ for every $t \in [0, 1]$, and for each j,*

$$h_j(t) = \sum_{k=0}^{K-1} \rho_k^j 1_{(t_k, t_{k+1})}(t)$$

for almost every $t \in [0, 1]$.

Proof The proof is by induction. Since $x_n \to x$ we have $\bar{X}(0) = x = \psi^*(0)$. Assume that $\bar{X}(t) = \psi^*(t)$ for all $t \in [0, t_k]$. The goal is to show that $\bar{X}(t) = \psi^*(t)$ for all

$t \in [0, t_{k+1}]$. Define $A_k = \pi(\psi^*(t))$, $t \in (t_k, t_{k+1})$ and $A = \pi(\psi^*(t_k))$. Note that $A_k \subset A$ due to the continuity of ψ^* and the upper semicontinuity of π. Define the random time

$$\tau_k \doteq \inf \left\{ t > t_k : \pi(\bar{X}(t)) \not\subset A \right\}.$$

Since $\pi(\bar{X}(t_k)) = \pi(\psi^*(t_k)) = A$ and $\bar{X}$ is continuous, the upper semicontinuity of π implies that $\tau_k > t_k$ and $\pi(\bar{X}(\tau_k)) \not\subset A$. We claim that it suffices to show $\bar{X}(t) = \psi^*(t)$ and $h_j(t) = \rho_k^j$ for all $t \in (t_k, t_{k+1} \wedge \tau_k)$. Indeed, if this is the case, we must have $\tau_k \geq t_{k+1}$ with probability one, since otherwise by continuity $\bar{X}(\tau_k) = \psi^*(\tau_k)$ and thus $\pi(\bar{X}(\tau_k)) = A_k \subset A$, a contradiction.

The proof of $\bar{X}(t) = \psi^*(t)$ and $h_j(t) = \rho_k^j$ for all $t \in I_k \doteq (t_k, t_{k+1} \wedge \tau_k)$ proceeds as follows. By Proposition 13.11, we can assume that $\sum_{i \in A} h_i(t) = 1$ for all $t \in I_k$. It follows from (13.17), (13.18), and Proposition 13.11 that

$$\frac{d}{dt}\bar{X}(t) = \sum_{i \in \pi(\bar{X}(t))} \beta_k^i h_i(t) = \sum_{j=1}^{d} \bar{\lambda}_{j,k} e_j - \sum_{i \in \pi(\bar{X}(t)),\ i \neq 0} \bar{\mu}_{i,k} h_i(t) e_i. \tag{13.23}$$

We will assume that A_k is a strict subset of A and $0 \notin A$, and note that the cases when they are equal or $0 \in A$ are similar. Also, it is straightforward to show that $(\bar{X}(t))_i = (\psi^*(t))_i$ for $i \notin A$ and $t \in I_k$, and so we only have to establish $(\bar{X}(t))_i = (\psi^*(t))_i$ and $h_i(t) = \rho_k^i$ for $i \in A$ and $t \in I_k$. It follows from the definitions of π, A, and A_k that for every $i \in A_k$ and $j \in A \setminus A_k$

$$c_i \left(\psi^*(t_k)\right)_i - c_j \left(\psi^*(t_k)\right)_j = 0,$$

and for $t \in (t_k, t_{k+1})$

$$c_i \left(\psi^*(t)\right)_i - c_j \left(\psi^*(t)\right)_j > 0.$$

Since $\dot{\psi}^*(t) \equiv \beta_k$ for $t \in (t_k, t_{k+1})$, the last display and (13.14) yield

$$0 < c_i(\beta_k)_i - c_j(\beta_k)_j = c_i(\bar{\lambda}_{i,k} - \rho_k^i \bar{\mu}_{i,k}) - c_j \bar{\lambda}_{j,k}. \tag{13.24}$$

We first claim that if $i \in A_k$ and $j \in A \setminus A_k$ then $c_j(\bar{X}(t))_j \leq c_i(\bar{X}(t))_i$ for all $t \in I_k$. Indeed, using (13.24) for the last inequality and $h_j(t) \geq 0$, $\rho^i \geq 0$ for the first inequality, for any $t \in I_k$ at which $c_j(\bar{X}(t))_j > c_i(\bar{X}(t))_i$ holds we have

$$\begin{aligned}
&\frac{d}{dt}\left[c_j(\bar{X}(t))_j - c_i(\bar{X}(t))_i\right] \\
&\quad = c_j(\bar{\lambda}_{j,k} - h_j(t)\mu_{j,k}) - c_i \bar{\lambda}_{i,k} \\
&\quad = c_j \bar{\lambda}_{j,k} - c_i \bar{\lambda}_{i,k} + c_i \rho_k^i \mu_{i,k} - c_j h_j(t) \mu_{j,k} - c_i \rho_k^i \mu_{i,k} \\
&\quad < c_j \bar{\lambda}_{j,k} - c_i \bar{\lambda}_{i,k} + c_i \rho_k^i \mu_{i,k} \\
&\quad < 0.
\end{aligned}$$

Thus $c_j(\bar{X}(t))_j \le c_i(\bar{X}(t))_i$ follows.

If $c_i(\bar{X}(t))_i$ are not all the same for $i \in A_k$, then $\pi(\bar{X}(t)) \subset A_k$ but $\pi(\bar{X}(t)) \neq A_k$. If they are all the same then $i \in A_k$ implies $i \in \pi(\bar{X}(t))$, and it is possible that $j \in \pi(\bar{X}(t))$ for some $j \in A \setminus A_k$. Hence there are only two possibilities, which are $\pi(\bar{X}(t)) \subset A_k$ but $\pi(\bar{X}(t)) \neq A_k$, or $A_k \subset \pi(\bar{X}(t))$. If the second case occurs, then we use the fact that for any nonempty B the solution to (13.21) is unique and component-wise strictly increasing. Note that there are only finitely many possible choices for $\pi(\bar{X}(t))$. Consider the set $T_k(B)$ of t such that $\pi(\bar{X}(t)) = B$. Recall that for an absolutely continuous function $g : [0, 1] \to \mathbb{R}$ the Lebesgue measure of $\{t : g(t) = 0, \dot{g}(t) \neq 0\}$ is zero. It follows that within $T_k(B)$ it is enough to consider only t such that

$$c_j(\bar{\lambda}_{j,k} - h_j(t)\mu_{j,k}) = c_i(\bar{\lambda}_{i,k} - h_i(t)\mu_{i,k}) \text{ for all } i, j \in B. \tag{13.25}$$

If $B = A_k$ then $h_i(t) = \rho_k^i$ and there is nothing to prove. If B is strictly larger than A_k then $\sum_{i \in A_k} h_i(t) < 1$, and by the uniqueness of solutions to (13.21) $h_i(t) < \rho_k^i$ for $i \in A_k$. However in this case, for any $i \in A_k$ and $j \in B \setminus A_k$ Eq. (13.24) implies

$$\begin{aligned}
&\frac{d}{dt}\left[c_j(\bar{X}(t))_j - c_i(\bar{X}(t))_i\right] \\
&\quad = c_j(\bar{\lambda}_{j,k} - h_j(t)\mu_{j,k}) - c_i(\bar{\lambda}_{i,k} - h_i(t)\mu_{i,k}) \\
&\quad = c_j\bar{\lambda}_{j,k} - c_i\bar{\lambda}_{i,k} + c_i\rho_k^i\mu_{i,k} - c_j h_j(t)\mu_{j,k} - c_i[\rho_k^i - h_i(t)]\mu_{i,k} \\
&\quad < c_j\bar{\lambda}_{j,k} - c_i\bar{\lambda}_{i,k} + c_i\rho_k^i\mu_{i,k} \\
&\quad < 0,
\end{aligned}$$

a contradiction to (13.25). Finally there is the case where $\pi(\bar{X}(t)) \subset A_k$ and the inclusion is strict. Again we partition according to $\pi(\bar{X}(t))$. Suppose that $i \in B$ and $j \in A_k \setminus B$. Again using the properties of solutions to (13.21) we have $h_i(t) > \rho_k^i$ for $i \in B$. Using that ρ_k^i satisfy (13.20)

$$\begin{aligned}
&\frac{d}{dt}\left[c_j(\bar{X}(t))_j - c_i(\bar{X}(t))_i\right] \\
&\quad = c_j\bar{\lambda}_{j,k} - c_i\bar{\lambda}_{i,k} + c_i h_i(t)\mu_{i,k} \\
&\quad = c_j\bar{\lambda}_{j,k} - c_j\rho_k^j\mu_{j,k} - c_i\bar{\lambda}_{i,k} + c_i\rho_k^i\mu_{i,k} + c_i[h_i(t) - \rho_k^i]\mu_{i,k} + c_j\rho_k^j\mu_{j,k} \\
&\quad = c_i[h_i(t) - \rho_k^i]\mu_{i,k} + c_j\rho_k^j\mu_{j,k} \\
&\quad > 0,
\end{aligned}$$

again a contradiction to (13.25). Hence the only possibility is $\pi(\bar{X}(t)) = A_k$ with $h_i(t) = \rho_k^i$. □

13.2.5.4 Analysis of the Cost

In this section, we prove the Laplace lower bound, namely inequality (13.12). Recall the definition of γ_i^n from Sect. 13.2.5.2. Using (13.16),

$$\begin{aligned}
&\lim_{n\to\infty} E_{x_n}\left[\int_{t_k}^{t_{k+1}} \sum_{i=1}^{d} \left[\lambda_i \ell\left(\varphi_i^n(t)\right) + \mu_i \ell\left(\varphi_{-i}^n(t)\right)\right] dt\right] \\
&= \lim_{n\to\infty} E_{x_n}\left[\int_{t_k}^{t_{k+1}} \sum_{i\in A_k, i\neq 0} \mu_i \ell\left(\frac{\bar{\mu}_{i,k}}{\mu_i}\right) \gamma_i^n(dt) + \sum_{j=1}^{d} \lambda_j \ell\left(\frac{\bar{\lambda}_{j,k}}{\lambda_j}\right) dt\right] \\
&= \int_{t_k}^{t_{k+1}} \left[\sum_{i\in A_k, i\neq 0} \mu_i \ell\left(\frac{\bar{\mu}_{i,k}}{\mu_i}\right) h_i(t) + \sum_{j=1}^{d} \lambda_j \ell\left(\frac{\bar{\lambda}_{j,k}}{\lambda_j}\right)\right] dt \\
&= (t_{k+1} - t_k)\left[\sum_{i\in A_k, i\neq 0} \mu_i \ell\left(\frac{\bar{\mu}_{i,k}}{\mu_i}\right) \rho_k^i + \sum_{j=1}^{d} \lambda_j \ell\left(\frac{\bar{\lambda}_{j,k}}{\lambda_j}\right)\right] \\
&\leq (t_{k+1} - t_k) \cdot [L^{A_k}(\beta_k) + \varepsilon] \\
&= \int_{t_k}^{t_{k+1}} L(\psi^*(t), \dot{\psi}^*(t))dt + (t_{k+1} - t_k)\varepsilon,
\end{aligned}$$

where the second equality is from Proposition 13.11, the third uses Lemma 13.12, the inequality on the fourth line is from (13.15) and the last line uses the fact that for $t \in (t_k, t_{k+1})$, $\dot{\psi}^*(t) = \beta_k$ and $\pi(\psi^*(t)) = A_k$. Summing over k and using Proposition 13.11 and Lemma 13.12 to establish

$$\lim_{n\to\infty} E_{x_n} F(\bar{X}^n) = E_x F(\bar{X}) = F(\psi^*),$$

we complete the proof of the inequality (13.12) and hence also the proof of Theorem 13.8. □

13.3 A Class of Pure Jump Markov Processes

In this section we consider the class of $\mathbb{R}^d$-valued pure jump Markov processes $\{X^\varepsilon\}_{\varepsilon\in(0,1)}$ with infinitesimal generator of the form

$$\mathscr{L}^\varepsilon f(x) = \frac{1}{\varepsilon} \sum_{k=1}^{K} \lambda_k(x)[f(x + \varepsilon v_k(x)) - f(x)], \quad x \in \mathbb{R}^d, \tag{13.26}$$

where f is a real bounded measurable map on $\mathbb{R}^d$ and for $k = 1, \ldots, K$, $\lambda_k : \mathbb{R}^d \to [0, \infty)$ and $v_k : \mathbb{R}^d \to \mathbb{R}^d$ are Lipschitz continuous functions. Such processes are ubiquitous in applications. However, they are not usually written in the SDE form of Chap. 10, and even when represented as the solution of an SDE they do not satisfy the conditions of Chap. 10, since G as in (10.1) cannot in general be smooth. However, unlike the last section the lack of smoothness is not a obstacle, and the SDE formulation provides a variational representation that allows an easy derivation of large and moderate deviation theorems. We will prove a large deviation principle for such processes in Sect. 13.3.1 and a moderate deviation principle in Sect. 13.3.2.

13.3.1 Large Deviation Principle

The main result of this section is a large deviation principle for $\{X^\varepsilon\}$ in $\mathscr{D}([0, T] : \mathbb{R}^d)$ when $X^\varepsilon(0) = x^\varepsilon$ for some $x^\varepsilon \in \mathbb{R}^d$ and $x^\varepsilon \to x$ as $\varepsilon \to 0$. We begin by introducing the associated rate function. For $\xi \in \mathscr{C}([0, T] : \mathbb{R}^d)$ with $\xi(0) = x$ define $\mathscr{U}(\xi)$ to be the collection of all $\varphi = \{\varphi_i\}_{i=1}^K$, such that each $\varphi_i : [0, T] \times \mathbb{R}_+ \to \mathbb{R}_+$ is a measurable map and the equation

$$\xi(t) = x + \sum_{k=1}^{K} \int_{[0,t]\times\mathbb{R}_+} v_k(\xi(s)) 1_{[0,\lambda_k(\xi(s))]}(y) \varphi_k(s, y) dy ds, \quad t \in [0, T]$$

holds. Define $I : \mathscr{D}([0, T] : \mathbb{R}^d) \to [0, \infty]$ by

$$I(\xi) \doteq \inf_{\varphi \in \mathscr{U}(\xi)} \left[\sum_{k=1}^{K} \int_{[0,T]\times\mathbb{R}_+} \ell(\varphi_k(s, y)) dy ds \right]$$

if $\xi \in \mathscr{C}([0, T] : \mathbb{R}^d)$ (where as usual the infimum is taken to be ∞ if the set is empty), and $I(\xi) = \infty$ if $\xi \in \mathscr{D}([0, T] : \mathbb{R}^d) \setminus \mathscr{C}([0, T] : \mathbb{R}^d)$ or if $\xi(0) \neq x$. We will impose the following condition on $\{\lambda_k\}$.

Condition 13.13 $\lambda_k : \mathbb{R}^d \to [0, \infty)$ and $v_k : \mathbb{R}^d \to \mathbb{R}^d$ are Lipschitz continuous functions for all $k = 1, \ldots, K$, and there is $c \in (0, \infty)$ such that $|\log \lambda_k(x)| \leq c$ for all $k = 1, \ldots, K$ and $x \in \mathbb{R}^d$.

Assume Condition 13.13 is satisfied and take as given $\xi \in \mathscr{C}([0, T] : \mathbb{R}^d)$ and $\varphi \in \mathscr{U}(\xi)$. We can assume without loss that $\varphi_k(s, y) = 1$ if $y > \lambda_k(\xi(s))$. Let

$$\bar{\varphi}_k(s) = \frac{1}{\lambda_k(\xi(s))} \int_0^{\lambda_k(\xi(s))} \varphi_k(s, y) dy.$$

Then of course

$$\xi(t) = x + \sum_{k=1}^{K} \int_0^t v_k(\xi(s))\lambda_k(\xi(s))\bar{\varphi}_k(s)ds, \quad t \in [0, T] \tag{13.27}$$

holds, and by Jensen's inequality and since $\ell \geq 0$

$$\sum_{k=1}^{K} \int_{[0,T]\times\mathbb{R}_+} \ell(\varphi_k(s, y))dyds \geq \sum_{k=1}^{K} \int_0^T \lambda_k(\xi(s))\ell(\bar{\varphi}_k(s))ds.$$

Hence if $\bar{\mathscr{U}}(\xi)$ is the collection of $\bar{\varphi} = \{\bar{\varphi}_i\}_{i=1}^K$, $\bar{\varphi}_i : [0, T] \to \mathbb{R}_+$ that satisfy (13.27), then an alternative and simpler expression for the rate function is

$$I(\xi) = \inf_{\bar{\varphi}\in\bar{\mathscr{U}}(\xi)} \left[\sum_{k=1}^{K} \int_0^T \lambda_k(\xi(s))\ell(\bar{\varphi}_k(s))ds \right].$$

Theorem 13.14 *Suppose that Condition 13.13 is satisfied. Let X^ε be the Markov process with infinitesimal generator $\mathscr{L}^\varepsilon$ in (13.26) and initial value $X^\varepsilon(0) = x^\varepsilon$ for some $x^\varepsilon \in \mathbb{R}^d$. Suppose $x^\varepsilon \to x$ as $\varepsilon \to 0$. Then $\{X^\varepsilon\}$ satisfies a large deviation principle on $\mathscr{D}([0, T] : \mathbb{R}^d)$ with rate function I.*

As usual, we give the proof only for the case $T = 1$. Let $F : \mathscr{D}([0, 1] : \mathbb{R}^d) \to \mathbb{R}$ be a bounded and continuous function. In order to prove the theorem it suffices to show that I has compact level sets and that for any such F

$$\lim_{\varepsilon\to 0} -\varepsilon \log E \exp\left\{-\varepsilon^{-1}F(X^\varepsilon)\right\} = \inf_{\xi\in\mathscr{D}([0,1]:\mathbb{R}^d)} [I(\xi) + F(\xi)].$$

The upper bound is shown in Sect. 13.3.1.1 and the lower bound in Sect. 13.3.1.2. The proof of compactness of level sets is similar to that of the Laplace upper bound and therefore omitted. Since the statement allows general convergent initial conditions, Proposition 1.12 implies that the LDP is uniform with respect to initial conditions in compact sets.

13.3.1.1 Laplace Upper Bound

In this section we will prove that for every $F \in \mathscr{C}_b(\mathscr{D}([0, 1] : \mathbb{R}^d))$

$$\limsup_{\varepsilon\to 0} \varepsilon \log E \exp\left\{-\varepsilon^{-1}F(X^\varepsilon)\right\} \leq -\inf_{\xi\in\mathscr{D}([0,1]:\mathbb{R}^d)} [I(\xi) + F(\xi)].$$

Let $(\Omega, \mathscr{F}, P)$ be a probability space equipped with a filtration $\{\mathscr{F}_t\}_{0\leq t\leq 1}$ that satisfies the usual conditions (see Appendix D). Let $\bar{N}_i$, for $i = 1, \ldots, K$, be mutually

independent $\mathscr{F}_t$-Poisson random measures on $[0,1]\times\mathbb{R}_+\times\mathbb{R}_+$ with intensity measure $\bar{\nu}_1(ds\times dy\times dz)=\bar{\nu}(dy\times dz)ds$, where $\bar{\nu}$ is Lebesgue measure on $\mathbb{R}^2_+$ (see Definition 8.11). Following the notation in Sect. 8.2, for $M\in(0,\infty)$ let $\bar{\mathscr{A}}_{b,M}$ be the collection of all $\mathscr{F}_t$-predictable $\varphi:[0,1]\times\mathbb{R}_+\times\Omega\to\mathbb{R}_+$ such that: $L_1(\varphi)\le M$; $n\ge\varphi(t,y,\omega)\ge 1/n$ for some $n\in\mathbb{N}$ and all (t,y,ω); and $\varphi(t,y,\omega)=1$ if $y>n$ for all (t,ω). Here L_1 is as defined as in 8.17 with $T=1$, $\nu_1(ds\times dy)=\nu(dy)ds$, ν is Lebesgue measure on $\mathbb{R}_+$ and $\mathscr{X}=\mathbb{R}_+$. Let as before $\bar{\mathscr{A}}_b\doteq\cup_{M\in\mathbb{N}}\bar{\mathscr{A}}_{b,M}$. For $\varphi\in\bar{\mathscr{A}}_b$, the point process N_i^{φ} for $i=1,\dots,K$ is defined by 8.16 with $\bar{N}$ replaced with $\bar{N}_i$. In terms of these point processes the variational representation in Theorem 8.12 gives

$$-\varepsilon\log E\exp\left\{-\varepsilon^{-1}F(X^\varepsilon)\right\}=\inf_{\varphi}E\left[F(\bar{X}^\varepsilon)+\sum_{i=1}^{K}\int_{[0,1]\times\mathbb{R}_+}\ell(\varphi_i(t,y))dydt\right], \tag{13.28}$$

where the infimum is taken over all $\varphi=\{\varphi_i\}_{i=1}^K$ such that $\varphi_i\in\bar{\mathscr{A}}_b$ for each i, and for any such φ, $\bar{X}^\varepsilon$ is given as the solution of the equation

$$\bar{X}^\varepsilon(t)=x^\varepsilon+\varepsilon\sum_{k=1}^{K}\int_{[0,t]\times\mathbb{R}_+}\nu_k(\bar{X}^\varepsilon(s-))1_{[0,\lambda_k(\bar{X}^\varepsilon(s-))]}(y)N_k^{\varphi_k/\varepsilon}(ds\times dy) \tag{13.29}$$

for $t\in[0,1]$. For each $\varepsilon>0$, let $\varphi^\varepsilon=\{\varphi_i^\varepsilon\}_{i=1}^K$ be ε-optimal for the infimum in (13.28), so that

$$-\varepsilon\log E\exp\left\{-\varepsilon^{-1}F(X^\varepsilon)\right\}\ge E\left[F(\bar{X}^\varepsilon)+\sum_{i=1}^{K}\int_{[0,1]\times\mathbb{R}_+}\ell(\varphi_i^\varepsilon(t,y))dydt\right]-\varepsilon, \tag{13.30}$$

where $\bar{X}^\varepsilon$ is defined by (13.29), replacing φ there with φ^ε. Using the boundedness of F and a localization argument (see for example the proof of Theorem 3.17), we can assume without loss of generality that

$$\sup_{\varepsilon\in(0,1)}\sum_{i=1}^{K}\int_{[0,1]\times\mathbb{R}_+}\ell(\varphi_i^\varepsilon(t,y))dydt\le M$$

for some $M<\infty$. Using the boundedness of λ_k and linear growth of ν_k for $k=1,\dots,K$, it follows that the collection $\{(\bar{X}^\varepsilon,\varphi^\varepsilon)\}$ is tight (see for example the proof of Lemma 10.11) and using the superlinearity of ℓ we obtain the following uniform integrability property:

$$\lim_{m\to\infty}\sup_{\varepsilon\in(0,1)}\sum_{i=1}^{K}\int_{[0,1]\times\mathbb{R}_+}\varphi_i^\varepsilon(t,y)1_{\{\varphi_i^\varepsilon(t,y)\ge m\}}dydt=0.$$

Let $(\bar{X}, \varphi)$ denote the limit along a weakly convergent sequence and assume without loss of generality by using the Skorohod representation theorem that the convergence is a.s. Then it follows from (13.29) that (see for example the proof of Lemma 10.12)

$$\bar{X}(t) = x + \sum_{k=1}^{K} \int_{[0,t]\times\mathbb{R}_+} v_k(\bar{X}(s)) 1_{[0,\lambda_k(\bar{X}(s))]}(y)\varphi_k(s,y)dyds, \quad t \in [0,1]. \tag{13.31}$$

Then using the definition of I, lower semicontinuity of ℓ and Fatou's lemma, it follows that

$$\liminf_{\varepsilon\to 0} E\left[F(\bar{X}^\varepsilon) + \sum_{i=1}^{K} \int_{[0,1]\times\mathbb{R}_+} \ell(\varphi_i^\varepsilon(t,y))dy\,dt \right] \geq \inf_{\xi\in\mathscr{D}([0,1]:\mathbb{R}^d)} [F(\xi) + I(\xi)],$$

and the desired Laplace upper bound follows from (13.30). □

13.3.1.2 Laplace Lower Bound

In this section we prove that for every $F \in \mathscr{C}_b(\mathscr{D}([0,1]:\mathbb{R}^d))$

$$\liminf_{\varepsilon\to 0} \varepsilon \log E \exp\left\{-\varepsilon^{-1}F(X^\varepsilon)\right\} \geq -\inf_{\xi\in\mathscr{D}([0,1]:\mathbb{R}^d)} [I(\xi) + F(\xi)].$$

The key ingredient in the proof of the lower bound is the following uniqueness result.

Proposition 13.15 *Fix $\sigma \in (0,1)$. Given $\xi \in \mathscr{C}([0,1]:\mathbb{R}^d)$ with $I(\xi) < \infty$ there exists a $\varphi^* = \{\varphi_i^*\}_{i=1}^K \in \mathscr{U}(\xi)$ such that*

(a)

$$\sum_{i=1}^{K} \int_{[0,1]\times\mathbb{R}_+} \ell(\varphi_i^*(t,y))dy\,dt \leq I(\xi) + \sigma,$$

(b) *if $\tilde{\xi} \in \mathscr{C}([0,1]:\mathbb{R}^d)$ is another function such that $\varphi^* \in \mathscr{U}(\tilde{\xi})$ then $\xi = \tilde{\xi}$.*

Proof Since $I(\xi) < \infty$, we can find $\varphi = \{\varphi_i\}_{i=1}^K \in \mathscr{U}(\xi)$ such that

$$\sum_{i=1}^{K} \int_{[0,1]\times\mathbb{R}_+} \ell(\varphi_i(t,y))dy\,dt \leq I(\xi) + \sigma/2.$$

Using Jensen's inequality, we can assume without loss that for $i = 1,\ldots,K$ there are $\rho_i : [0,1] \to \mathbb{R}_+$ such that for all $(s,y) \in [0,1]\times\mathbb{R}_+$,

$$\varphi_i(s,y) = \rho_i(s)1_{[0,\lambda_i(\xi(s))]}(y) + 1_{(\lambda_i(\xi(s)),\infty)}(y).$$

For $a \in (0,1)$ let

$$\varphi_i^a(s,y) = \frac{\rho_i(s)}{1-a} 1_{[0,(1-a)\lambda_i(\xi(s))]}(y) + 1_{((1+a)\lambda_i(\xi(s)),\infty)}(y). \tag{13.32}$$

Then $\varphi^a \doteq \{\varphi_i^a\}_{i=1}^K$ satisfies $\varphi^a \in \mathscr{U}(\xi)$. It can be checked using the formula for ℓ that there is $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\gamma(\theta) \to 0$ as $\theta \to 0$ and

$$\sum_{i=1}^K \int_{[0,1]\times\mathbb{R}_+} [\ell(\varphi_i^a(s,y)) - \ell(\varphi_i(s,y))] dy ds \le \gamma(a)(I(\xi)+1).$$

Choose $a > 0$ sufficiently small so that the right side is bounded above by $\sigma/2$. Then the first part of the proposition holds with $\varphi^* \doteq \varphi^a$. We now show that the second part holds with this φ^* as well.

Suppose that $\tilde{\xi}$ is another function such that $\varphi^* \in \mathscr{U}(\tilde{\xi})$. We need to show that $\tilde{\xi} = \xi$. Let

$$\tau \doteq \inf\{t \in [0,1] : \xi(t) \neq \tilde{\xi}(t)\} \wedge 1.$$

It suffices to show that $\tau = 1$. Suppose to the contrary that $\tau < 1$. We will show that for some $\delta \in (0, 1-\tau]$

$$\xi(t) = \tilde{\xi}(t), \quad t \in [\tau, \tau+\delta]. \tag{13.33}$$

This contradiction will prove the desired result.

Note that $\xi(\tau) = \tilde{\xi}(\tau)$. From Condition 13.13, we can find $c_1, c_2 \in (0,\infty)$ such that $c_1 \le \lambda_k(x) \le c_2$ for all $x \in \mathbb{R}^d$ and $k = 1, \ldots, K$. Using continuity of λ_k, ξ and $\tilde{\xi}$ we can find $\delta > 0$ such that for all $s \in [\tau, \tau+\delta]$ and $k = 1, \ldots, K$

$$(1-a)\lambda_k(\xi(s)) < \lambda_k(\tilde{\xi}(s)) < (1+a)\lambda_k(\xi(s)). \tag{13.34}$$

Also, using the Lipschitz property of v_k, there is $\kappa \in (0,\infty)$ such that

$$\max_{1\le k\le K} \|v_k(x) - v_k(\tilde{x})\| \le \kappa \|x - \tilde{x}\| \text{ for all } x, \tilde{x} \in \mathbb{R}^d.$$

Then for any $s \in [0, \delta]$

$$\begin{aligned}
&\sup_{0\le u\le s} \|\xi(\tau+u) - \tilde{\xi}(\tau+u)\| \\
&\le \sum_{i=1}^K \int_{[\tau,\tau+s]\times\mathbb{R}_+} \left\| v_i(\xi(r)) 1_{[0,\lambda_i(\xi(r))]}(y) - v_i(\tilde{\xi}(r)) 1_{[0,\lambda_i(\tilde{\xi}(r))]}(y) \right\| \varphi_i^a(r,y) dy dr \\
&\le \sum_{i=1}^K \int_{[\tau,\tau+s]\times\mathbb{R}_+} \left\| v_i(\xi(r)) - v_i(\tilde{\xi}(r)) \right\| \varphi_i^a(r,y) 1_{[0,\lambda_i(\xi(r))]}(y) dy dr
\end{aligned}$$

$$\leq \kappa \int_{[\tau,\tau+s]} \sup_{0\leq u\leq r} \left\| \xi(u) - \tilde{\xi}(u) \right\| \left(\int_{\mathbb{R}_+} \sum_{i=1}^{K} \varphi_i^a(r, y) 1_{[0,\lambda_i(\xi(r))]}(y) dy \right) dr,$$

where the second inequality follows on observing that by (13.32) and (13.34), for all $r \in [\tau, \tau + \delta]$ and $y \in \mathbb{R}_+$

$$1_{[0,\lambda_i(\xi(r))]}(y)\varphi_i^a(r, y) = 1_{[0,\lambda_i(\tilde{\xi}(r))]}(y)\varphi_i^a(r, y).$$

The bound $x \leq c[\ell(x) + 1]$ for some $c < \infty$, the bound asserted in part (a), and an application of Gronwall's lemma shows that

$$\sup_{0\leq u\leq \delta} \|\xi(\tau + u) - \tilde{\xi}(\tau + u)\|$$

is identically 0. This proves (13.33) and completes the proof of the proposition. □

Remark 13.16 Proposition 13.15 is the key ingredient in the proof of the large deviation lower bound. It asserts that given any trajectory with finite cost, we can find controls that one could apply to the driving Poisson random measure that are arbitrarily close in cost to the infimum over all controls, and that give a unique output when pushed through the deterministic limiting dynamics that map a control to a controlled trajectory. Note that the perturbation of the controls in (13.32), which can be done without changing the trajectory and with a small increase in cost, conveniently sets the intensity to zero in a neighborhood of the points at which the mapping from noise space to state space is discontinuous. This provides a new technique for proving the large deviation lower bound, and it relies heavily on being able to represent the system of interest (here a pure jump process with a finite number of jump types) in terms of a fixed and exogenous set of PRMs, as well as the representation theorem for PRMs. In comparison with the mollifications used say in Sect. 4.7, the mollification or approximation here is done directly on the controls $\{\varphi_i\}$, rather than on the trajectory ξ. Recent applications of this technique include [23, 42], and for further discussion see the notes at the end of Chap. 10.

We now complete the proof of Laplace lower bound. Fix $\sigma \in (0, 1)$ and let $\xi^* \in \mathscr{C}([0, T] : \mathbb{R}^d)$ satisfy

$$F(\xi^*) + I(\xi^*) \leq \inf_{\xi\in\mathscr{C}([0,T]:\mathbb{R}^d)} [F(\xi) + I(\xi)] + \sigma.$$

Let $\varphi^* = \{\varphi_i^*\}_{i=1}^K \in \mathscr{U}(\xi^*)$ be as given by Proposition 13.15 (with ξ replaced by ξ^*). For $i = 1, \ldots, K$ define the deterministic controls

$$\varphi_i^\varepsilon(s, y) \tag{13.35}$$
$$\doteq 1 + 1_{\{y\leq 1/\varepsilon\}} \left(\varepsilon 1_{\{\varphi_i^*(s,y)\leq\varepsilon\}} + \varphi_i^*(s, y) 1_{\{\varepsilon<\varphi_i^*(s,y)<1/\varepsilon\}} + \varepsilon^{-1} 1_{\{\varphi_i^*(s,y)\geq 1/\varepsilon\}} - 1 \right).$$

Since $x \mapsto \ell(x)$ is increasing for $x \geq 1$ and decreasing for $x \leq 1$, $\ell(\varphi_i^\varepsilon(t,y)) \leq \ell(\varphi_i^*(t,y))$ for all $\varepsilon > 0, i, t$ and y. Hence if $\bar{X}^\varepsilon$ is defined through (13.29) with φ replaced with φ^ε, then

$$
\begin{aligned}
-\varepsilon \log E \exp\left\{-\varepsilon^{-1} F(X^\varepsilon)\right\} &\leq E\left[F(\bar{X}^\varepsilon) + \sum_{i=1}^K \int_{[0,1]\times\mathbb{R}_+} \ell(\varphi_i^\varepsilon(t,y))dy\,dt\right] \\
&\leq E\left[F(\bar{X}^\varepsilon) + \sum_{i=1}^K \int_{[0,1]\times\mathbb{R}_+} \ell(\varphi_i^*(t,y))dy\,dt\right],
\end{aligned}
$$

where the first inequality is from (13.28) on observing that φ_i^ε in (13.35) is in $\bar{\mathscr{A}}_b$ for each i. By repeating the arguments used for proving the upper bound we see that $\{\bar{X}^\varepsilon\}$ is tight. Furthermore, φ^ε converges to φ^*. If $\bar{X}^\varepsilon$ converges along a subsequence to $\bar{X}$ then $\bar{X}$ must satisfy (13.31) with φ replaced with φ^*. From the uniqueness in Proposition 13.15 it follows that $\bar{X} = \xi^*$. Therefore

$$
\begin{aligned}
&\limsup_{\varepsilon\to 0} -\varepsilon \log E \exp\left\{-\varepsilon^{-1} F(X^\varepsilon)\right\} \\
&\leq \limsup_{\varepsilon\to 0} E\left[F(\bar{X}^\varepsilon) + \sum_{i=1}^K \int_{[0,1]\times\mathbb{R}_+} \ell(\varphi_i^*(t,y))dydt\right] \\
&= F(\xi^*) + \sum_{i=1}^K \int_{[0,1]\times\mathbb{R}_+} \ell(\varphi_i^*(t,y))dydt \\
&\leq F(\xi^*) + I(\xi^*) + \sigma \\
&\leq \inf_{\xi\in\mathscr{D}([0,T]:\mathbb{R}^d)} [F(\xi) + I(\xi)] + 2\sigma.
\end{aligned}
$$

Since $\sigma > 0$ is arbitrary, we have the desired Laplace lower bound. □

13.3.2 Moderate Deviation Principle

We will use notation and results from Sects. 9.2.2 and 10.3. In particular, $a(\varepsilon)$ and $\varkappa(\varepsilon)$ as in (9.6) satisfy $a(\varepsilon) \to 0$ and $\varkappa(\varepsilon) \doteq \varepsilon/a^2(\varepsilon) \to 0$ as $\varepsilon \to 0$. Let $X^0 \in \mathscr{C}([0,T] : \mathbb{R}^d)$ be a solution of the equation

$$
X^0(t) = x + \int_0^t \beta(X^0(s))ds, \quad t \in [0,T], \tag{13.36}
$$

where for $x \in \mathbb{R}^d$, $\beta(x) \doteq \sum_{k=1}^K \nu_k(x)\lambda_k(x)$. We will impose the following condition, under which there is a unique solution of (13.36).

Condition 13.17 $\lambda_k : \mathbb{R}^d \to [0, \infty)$ and $v_k : \mathbb{R}^d \to \mathbb{R}^d$ are Lipschitz continuous functions for all $k = 1, \ldots, K$ and there is $c \in (0, \infty)$ such that $\lambda_k(x) \le c$ for all $k = 1, \ldots, K$ and $x \in \mathbb{R}^d$. The map $\beta : \mathbb{R}^d \to \mathbb{R}^d$ is differentiable and $D\beta$, the $d \times d$ matrix of first derivatives of β, is Lipschitz continuous. Thus for some $L_{D\beta} \in (0, \infty)$,

$$\|D\beta(x) - D\beta(y)\| \le L_{D\beta}\|x - y\| \text{ for all } x, y \in \mathbb{R}^d.$$

Note that under Condition 13.17 β is a locally Lipschitz continuous function with linear growth, and thus (13.36) admits a unique solution. In this section we establish a Laplace principle for $\{Y^\varepsilon\}$ with scaling function $\varkappa(\varepsilon)$, where

$$Y^\varepsilon = \frac{1}{a(\varepsilon)}(X^\varepsilon - X^0).$$

We now introduce the rate function associated with the collection $\{Y^\varepsilon\}$. For $\eta \in \mathscr{C}([0, T] : \mathbb{R}^d)$, let $\mathscr{V}(\eta)$ be the collection of all $u = \{u_i\}_{i=1}^K$ such that each $u_i \in \mathscr{L}^2([0, T] : \mathbb{R})$ and the equation

$$\eta(t) = \int_0^t [D\beta(X^0(s))]\eta(s)ds + \sum_{k=1}^K \int_0^t v_k(X^0(s))\sqrt{\lambda_k(X^0(s))}u_k(s)ds, \ t \in [0, T]$$

is satisfied. Define $I : \mathscr{D}([0, T] : \mathbb{R}^d) \to [0, \infty]$ by

$$I(\eta) \doteq \inf_{u \in \mathscr{V}(\eta)} \left[\frac{1}{2} \sum_{k=1}^K \int_0^1 u_k^2(s)ds \right]$$

if $\eta \in \mathscr{C}([0, T] : \mathbb{R}^d)$ (where the infimum is taken to be ∞ if the set is empty), and $I(\eta) = \infty$ for $\eta \in \mathscr{D}([0, T] : \mathbb{R}^d) \setminus \mathscr{C}([0, T] : \mathbb{R}^d)$. The following is the main theorem of this section.

Theorem 13.18 *Suppose that Condition 13.17 is satisfied. Further suppose that* $(x^\varepsilon - x)/a(\varepsilon) \to 0$ *as* $\varepsilon \to 0$. *Then* $\{Y^\varepsilon\}$ *satisfies a Laplace principle in* $\mathscr{D}([0, T] : \mathbb{R}^d)$ *with scaling function* $\varkappa(\varepsilon)$ *and rate function* I.

Remark 13.19 The rate function I has the following alternative representation. For $\eta \in \mathscr{C}([0, 1] : \mathbb{R}^d)$, let $\tilde{\mathscr{V}}(\eta)$ be the collection of all $\psi = \{\psi_i\}_{i=1}^K$ such that each $\psi_i \in \mathscr{L}^2([0, 1] \times \mathbb{R}_+)$ and the following equation is satisfied for $t \in [0, 1]$.

$$\begin{aligned}\eta(t) = &\int_0^t [D\beta(X^0(s))]\eta(s)ds \\ &+ \sum_{k=1}^K \int_{[0,t]\times\mathbb{R}_+} v_k(X^0(s))1_{[0,\lambda_k(X^0(s))]}(y)\psi_k(s, y)dyds. \end{aligned} \tag{13.37}$$

Define $\tilde{I} : \mathscr{D}([0, 1] : \mathbb{R}^d) \to [0, \infty]$ by

$$\tilde{I}(\eta) \doteq \inf_{\psi \in \tilde{\mathscr{V}}(\eta)} \left[\frac{1}{2} \sum_{k=1}^{K} \int_{[0,1]\times\mathbb{R}_+} \psi_k^2(s,y) dy ds \right]$$

if $\eta \in \mathscr{C}([0,1] : \mathbb{R}^d)$, and $\tilde{I}(\eta) = \infty$ for $\eta \in \mathscr{D}([0,1] : \mathbb{R}^d) \setminus \mathscr{C}([0,1] : \mathbb{R}^d)$. Then it can be checked that $\tilde{I} = I$. The optimal $\psi_k(s,y)$ for a given $u_k(s)$ is equal to $u_k(s) 1_{[0,\lambda_k(X^0(s))]}(y)/[\lambda_k(X^0(s))]^{1/2}$.

Remark 13.20 Another form of the rate function is analogous to that of Chap. 5. Define

$$A_{ij}(x) \doteq \sum_{k=1}^{K} \lambda_k(x) [v_k(x)]_i [v_k(x)]_j ,$$

where $[v_k(x)]_i$ is the ith component of $v_k(x)$. It is easy to check that $A_{ij}(x)$ is symmetric and nonnegative definite, and thus has a well defined square root $A^{1/2}(x)$. It can be shown that the rate function I is same as the function $\bar{I}$ defined by

$$\bar{I}(\eta) \doteq \inf_{f \in \mathscr{L}^2([0,1]:\mathbb{R}^d)} \left[\frac{1}{2} \int_0^1 \|f(s)\|^2 ds \right]$$

for $\eta \in \mathscr{C}([0,1] : \mathbb{R}^d)$, where the infimum is taken over all $f \in \mathscr{L}^2([0,1] : \mathbb{R}^d)$ such that

$$\eta(t) = \int_0^t [D\beta(X^0(s))]\eta(s) ds + \int_0^t A^{1/2}(X^0(s)) f(s) ds, \ t \in [0,1].$$

The infimum is taken to be ∞ if the set of such f is empty, and $\tilde{I}(\eta) = \infty$ for $\eta \in \mathscr{D}([0,1] : \mathbb{R}^d) \setminus \mathscr{C}([0,1] : \mathbb{R}^d)$. To see that the two rate functions are the same it is enough to check that they give the same cost in the sense that for all $p \in \mathbb{R}^d$ the Legendre-Fenchel transforms

$$\sup_{f \in \mathbb{R}^d} \left[\left\langle p, A^{1/2}(x) f \right\rangle - \frac{1}{2} \langle f, f \rangle \right]$$

and

$$\sup_{u \in \mathbb{R}^K} \left[\sum_{k=1}^{K} \left\langle p, v_k(v) \sqrt{\lambda_k(x)} u_k \right\rangle - \frac{1}{2} \sum_{k=1}^{K} u_k^2 \right]$$

coincide. Evaluating for the optimum gives

$$\frac{1}{2} \langle p, A(x) p \rangle \text{ and } \frac{1}{2} \sum_{k=1}^{K} \left\langle p, v_k(v) \sqrt{\lambda_k(x)} \right\rangle^2$$

respectively, which coincide owing to the definition of $A(x)$.

We return to the proof of Theorem 13.18, which will be given for the case $T = 1$. In order to prove the theorem we will verify the sufficient condition for a MDP formulated in Condition 9.8. We will in fact use a simpler form of this condition since the process of interest does not have a Brownian noise component and we do not seek a uniform Laplace principle here. We record this simpler version of Condition 9.8 below. Let $\mathbb{M} \doteq \Sigma([0,1] \times \mathbb{R}_+)$. Also, let $\mathbb{U}$ be a Polish space and $\{\mathscr{K}^\varepsilon\}$ be a collection of maps from $\mathbb{M}^K \to \mathbb{U}$. Let

$$\hat{S}_n \doteq \left\{ f = \{f_i\}_{i=1}^K : f_i \in \mathscr{L}^2([0,1] \times \mathbb{R}_+) \text{ and } \sum_{i=1}^K \int_{[0,1]\times\mathbb{R}_+} f_i^2(s,y)dyds \le n \right\}.$$

For $n \in \mathbb{N}$ let

$$S_{n,+}^\varepsilon \doteq \Big\{ g = \{g_i\}_{i=1}^K : g_i : [0,1] \times \mathbb{R}_+ \to \mathbb{R}_+ \\ \text{and } \sum_{i=1}^K \int_{[0,1]\times\mathbb{R}_+} \ell(g_i(s,y))dyds \le na^2(\varepsilon) \Big\}$$

and let

$$\mathscr{U}_{n,+}^\varepsilon \doteq \left\{ \varphi = \{\varphi_i\}_{i=1}^K : \varphi_i \in \bar{\mathscr{A}_b} \text{ for each } i, \text{ and } \varphi \in S_{n,+}^\varepsilon \text{ a.s.} \right\}.$$

For notational convenience when $\varphi \in \mathscr{U}_{n,+}^\varepsilon$, we write $\{\varepsilon N_i^{\varphi_i/\varepsilon}\}_{i=1}^K$ as $\varepsilon N^{\varphi/\varepsilon}$. The following is a sufficient condition for $\{\mathscr{K}^\varepsilon(\varepsilon N^{1/\varepsilon})\}$ to satisfy an MDP with scaling function $\varkappa(\varepsilon)$.

Condition 13.21 For some measurable map $\mathscr{K}^0 : (\mathscr{L}^2([0,1] \times \mathbb{R}_+))^K \to \mathbb{U}$, the following two conditions hold.

(a) For every $n \in \mathbb{N}$, the set $\left\{ \mathscr{K}^0(f) : f \in \hat{S}_n \right\}$ is a compact subset of $\mathbb{U}$.

(b) Given $n \in \mathbb{N}$ and $\varepsilon > 0$, let $\varphi^\varepsilon \in \mathscr{U}_{n,+}^\varepsilon$ and $\zeta_i^\varepsilon \doteq (\varphi_i^\varepsilon - 1)/a(\varepsilon)$. Suppose that for some $\theta \in (0,1]$ there is $m \in \mathbb{N}$ such that

$$\{\zeta_i^\varepsilon 1_{\{|\zeta_i^\varepsilon| \le \theta/a(\varepsilon)\}}\}_{i=1}^K \Rightarrow \zeta \doteq \{\zeta_i\}_{i=1}^K \text{ in } \hat{S}_m.$$

Then $\mathscr{K}^\varepsilon(\varepsilon N^{\varphi^\varepsilon/\varepsilon}) \Rightarrow \mathscr{K}^0(\zeta)$.

For $\eta \in \mathbb{U}$ define

$$I(\eta) \doteq \inf_{f \in (\mathscr{L}^2([0,1]\times\mathbb{R}_+))^K : \eta = \mathscr{K}^0(f)} \left[\frac{1}{2} \sum_{i=1}^K \int_{[0,1]\times\mathbb{R}_+} f_i^2(s,y)dyds \right].$$

It follows from Theorem 9.9 that if Condition 13.21 is satisfied, the collection $\{\mathscr{K}^\varepsilon(\varepsilon N^{1/\varepsilon})\}$ satisfies an MDP with scaling function $\varkappa(\varepsilon)$ and the rate function I defined in the last display.

We will now apply Theorem 9.9 for the case where $\mathscr{K}^\varepsilon$ is implicitly defined by $\mathscr{K}^\varepsilon(\varepsilon N^{1/\varepsilon}) = Y^\varepsilon$. For $\varphi^\varepsilon \in \mathscr{U}^\varepsilon_{n,+}$, let $\bar{Y}^\varepsilon \doteq (\bar{X}^\varepsilon - X^0)/a(\varepsilon)$, where $\bar{X}^\varepsilon$ solves (13.29) with φ replaced with φ^ε. Also, for $f \in (\mathscr{L}^2([0,1]\times\mathbb{R}_+))^K$, let

$$\eta(t) = \int_0^t [D\beta(X^0(s))]\eta(s)ds + \sum_{k=1}^K \int_{[0,t]\times\mathbb{R}_+} v_k(X^0(s))1_{[0,\lambda_k(X^0(s))]}(y) f_k(s,y)dyds. \tag{13.38}$$

Then in order to prove Theorem 13.18, in view of Remark 13.19, it suffices to show the following.

(i) If for some $n \in \mathbb{N}$ $f^m \to f$ in $\hat{S}_n$ as $m \to \infty$, and if η_m solves (13.38) with f replaced by f^m, then $\eta_m \to \eta$.

(ii) Let $\zeta_i^\varepsilon \doteq (\varphi_i^\varepsilon - 1)/a(\varepsilon)$ where $\varphi^\varepsilon \in \mathscr{U}^\varepsilon_{n,+}$. If for some $\theta \in (0,1]$, there is $m \in \mathbb{N}$ such that

$$\tilde{\zeta}^\varepsilon \doteq \{\zeta_i^\varepsilon 1_{\{|\zeta_i^\varepsilon| \le \theta/a(\varepsilon)\}}\}_{i=1}^K \Rightarrow \zeta \doteq \{\zeta_i\}_{i=1}^K \text{ in } \hat{S}_m,$$

then $\bar{Y}^\varepsilon \Rightarrow \bar{Y}$, where $\bar{Y}$ solves (13.38) with f replaced by ζ.

In rest of the section we prove (i) and (ii).

13.3.2.1 Proof of (i)

Let $\{f^m\}$, f be elements in $\hat{S}_n$ such that $f^m \to f$ as $m \to \infty$. Then using the fact that

$$(s,y) \mapsto v_k(X^0(s))1_{[0,\lambda_k(X^0(s))]}(y)$$

is in $\mathscr{L}^2([0,1]\times\mathbb{R}_+)$, it follows from Hölder's inequality and $f_k^m \to f_k$ weakly in $\mathscr{L}^2$ that as $m \to \infty$

$$\sup_{0\le t\le 1} \sum_{k=1}^K \left\| \int_{[0,t]\times\mathbb{R}_+} v_k(X^0(s))1_{[0,\lambda_k(X^0(s))]}(y)(f_k^m(s,y) - f_k(s,y))dyds \right\| \to 0.$$

We also have

$$\begin{aligned}
&\|\eta_m(t) - \eta(t)\| \\
&\quad \le \|D\beta(X^0(\cdot))\|_{\infty,1} \int_0^t \|\eta_m(s) - \eta(s)\|ds \\
&\qquad + \sup_{0\le t\le 1} \sum_{k=1}^K \left\| \int_{[0,t]\times\mathbb{R}_+} v_k(X^0(s))1_{[0,\lambda_k(X^0(s))]}(y)(f_k^m(s,y) - f_k(s,y))dyds \right\|.
\end{aligned}$$

The statement in (i) now follows from Gronwall's inequality. □

13.3.2.2 Proof of (ii)

From arguments as in Lemma 10.20 it follows that there exists an $\varepsilon_0 \in (0,1)$ such that

$$\sup_{\varepsilon\in(0,\varepsilon_0)} \sup_{\varphi\in\mathscr{U}^{\varepsilon}_{n,+}} E\|\bar{X}^{\varepsilon}\|_{\infty,1} < \infty. \tag{13.39}$$

This is a consequence of the fact that β has linear growth and the quantity in the current setting that is analogous to $M_G(y)$ in Sect. 10.3 is $1_{[0,c]}(y)$, which is clearly in the class $\mathscr{L}^1(\nu) \cap \mathscr{L}^{\rho}_{\exp}$ when ν is Lebesgue measure on $\mathbb{R}_+$ (here c is from Condition 13.17).

Next we prove the following analogue of Lemma 10.21.

Lemma 13.22 *For every $n \in \mathbb{N}$, there exists an $\varepsilon_1 \in (0,1)$ such that*

$$\left\{\|\bar{Y}^{\varepsilon}\|_{\infty,1},\ \varphi \in \mathscr{U}^{\varepsilon}_{n,+},\ \varepsilon \in (0,\varepsilon_1)\right\}$$

is a tight collection of $\mathbb{R}_+$-valued random variables and

$$\left\{\bar{Y}^{\varepsilon},\ \varphi \in \mathscr{U}^{\varepsilon}_{n,+},\ \varepsilon \in (0,\varepsilon_1)\right\}$$

is a tight collection of $\mathscr{D}([0,1]:\mathbb{R}^d)$-valued random variables, where $\bar{Y}^{\varepsilon} \doteq (\bar{X}^{\varepsilon} - X^0)/a(\varepsilon)$ and $\bar{X}^{\varepsilon}$ uses the control φ.

Proof We write $\bar{Y}^{\varepsilon}$ as

$$\bar{Y}^{\varepsilon}(t) = \bar{Y}^{\varepsilon}(0) + M^{\varepsilon}(t) + B^{\varepsilon}(t) + C^{\varepsilon}(t), \quad t \in [0,1] \tag{13.40}$$

where $\bar{Y}^{\varepsilon}(0) = a(\varepsilon)^{-1}(x^{\varepsilon} - x) \to 0$ as $\varepsilon \to 0$, and

$$M^{\varepsilon}(t) \doteq \frac{\varepsilon}{a(\varepsilon)} \sum_{k=1}^{K} \int_{[0,t]\times\mathbb{R}_+} v_k(\bar{X}^{\varepsilon}(s-)) 1_{[0,\lambda_k(\bar{X}^{\varepsilon}(s-))]}(y) N^{\varphi_k/\varepsilon}_{c,k}(ds \times dy)$$

$$B^{\varepsilon}(t) \doteq \frac{1}{a(\varepsilon)} \int_0^t (\beta(\bar{X}^{\varepsilon}(s)) - \beta(X^0(s)))ds$$

$$C^{\varepsilon}(t) \doteq \sum_{k=1}^{K} \int_{[0,t]\times\mathbb{R}_+} v_k(\bar{X}^{\varepsilon}(s)) 1_{[0,\lambda_k(\bar{X}^{\varepsilon}(s))]}(y) \psi_k(s,y) dy ds,$$

where $N^{\varphi_k/\varepsilon}_{c,k}$ is the compensated form of the point process $N^{\varphi_k/\varepsilon}_k$ and $\psi \doteq (\varphi - 1)/a(\varepsilon)$.

Exactly as in the proof of (10.37) we see, by using the moment bound (13.39), that

$$\sup_{\varphi\in\mathscr{U}^{\varepsilon}_{n,+}} E\|M^{\varepsilon}\|_{\infty,1} \to 0 \text{ as } \varepsilon \to 0.$$

Also, using the Lipschitz property of v_k, λ_k, the boundedness of λ_k, and the fact that $\sup_{0\le t\le 1}\|X^0(t)\| < \infty$, we see that for some $\kappa \in (0,\infty)$ (not depending on ε or φ) and all $s \in [0,1]$,

$$\begin{aligned}\|\beta(\bar{X}^\varepsilon(s)) - \beta(X^0(s))\| &\le \sum_{k=1}^K \|v_k(\bar{X}^\varepsilon(s))\lambda_k(\bar{X}^\varepsilon(s)) - v_k(X^0(s))\lambda_k(X^0(s))\| \\ &\le \sum_{k=1}^K \lambda_k(\bar{X}^\varepsilon(s))\|v_k(\bar{X}^\varepsilon(s)) - v_k(X^0(s))\| \\ &\quad + \sum_{k=1}^K \|v_k(X^0(s))\||\lambda_k(\bar{X}^\varepsilon(s)) - \lambda_k(X^0(s))| \\ &\le \kappa\|\bar{X}^\varepsilon(s) - X^0(s)\|.\end{aligned}$$

Therefore,

$$\|B^\varepsilon\|_{\infty,t} \le \kappa \int_0^t \|\bar{Y}^\varepsilon(s)\|ds,\ \ t \in [0,1].$$

By using $\lambda_k(x) \le c$, for some $c_1 \in (0,\infty)$, and all $\varphi \in \mathscr{U}^\varepsilon_{n,+}$, $\varepsilon \in (0,1)$,

$$E\|C^\varepsilon\|_{\infty,1} \le c_1 E\left(\left(\|\bar{X}^\varepsilon\|_{\infty,1} + 1\right)\sum_{k=1}^K \int_{[0,1]\times[0,c]} |\psi_k(s,y)|dy\,ds\right).$$

From parts (a) and (c) of Lemma 9.7, there is a $c_2 \in (0,\infty)$ such that for every $\varphi \in \mathscr{U}^\varepsilon_{n,+}$, $\varepsilon \in (0,1)$, $k = 1,\dots,K$,

$$\begin{aligned}\int_{[0,1]\times[0,c]} |\psi_k(s,y)|dy\,ds &= \int_{[0,1]\times[0,c]} |\psi_k(s,y)|1_{\{|\psi_k(s,y)|\ge 1/a(\varepsilon)\}}dy\,ds \\ &\quad + \int_{[0,1]\times[0,c]} |\psi_k(s,y)|1_{\{|\psi_k(s,y)|< 1/a(\varepsilon)\}}dy\,ds \\ &\le \int_{[0,1]\times[0,c]} |\psi_k(s,y)|1_{\{|\psi_k(s,y)|\ge 1/a(\varepsilon)\}}dy\,ds \\ &\quad + c^{1/2}\left(\int_{[0,1]\times[0,c]} |\psi_k(s,y)|^2 1_{\{|\psi_k(s,y)|< 1/a(\varepsilon)\}}dy\,ds\right)^{1/2} \\ &\le c_2.\end{aligned}$$

Combining this with the moment bound in (13.39), we obtain

$$\sup_{\varphi\in\mathscr{U}^\varepsilon_{n,+},\varepsilon\in(0,\varepsilon_0)} E\|C^\varepsilon\|_{\infty,1} < \infty.$$

The first statement in the lemma now follows by combining these estimates on M^ε, B^ε, C^ε and Gronwall's lemma. The second statement follows on estimating oscillations of $\bar{Y}^\varepsilon$ using similar estimates along with Lemma 10.18 and the first tightness property established in the lemma. Details are omitted. □

We now verify (ii). Let $\varphi^\varepsilon \in \mathscr{U}^\varepsilon_{n,+}$ and let $\zeta^\varepsilon_i \doteq (\varphi^\varepsilon_i - 1)/a(\varepsilon)$. Suppose for some $\theta \in (0, 1]$, there is $m \in \mathbb{N}$ such that

$$\tilde{\zeta}^\varepsilon \doteq \{\zeta^\varepsilon_i 1_{\{|\zeta^\varepsilon_i| \le \theta/a(\varepsilon)\}}\}_{i=1}^K \Rightarrow \zeta \doteq \{\zeta_i\}_{i=1}^K \text{ in } \hat{S}_m.$$

We need to show that

$$\bar{Y}^\varepsilon \Rightarrow \bar{Y}, \tag{13.41}$$

where $\bar{Y}^\varepsilon = (\bar{X}^\varepsilon - X^0)/a(\varepsilon)$, $\bar{X}^\varepsilon$ solves (13.29) with φ replaced by φ^ε, and $\bar{Y}$ solves (13.38) with f replaced by ζ.

From Lemma 13.22, $\{\bar{Y}^\varepsilon\}$ is tight. Note that in particular this implies $\bar{X}^\varepsilon \to X^0$. Now suppose that $(\bar{Y}^\varepsilon, \bar{X}^\varepsilon, \tilde{\zeta}^\varepsilon)$ converges to $(\bar{Y}, X^0, \tilde{\zeta})$ in distribution. Note that $\tilde{\zeta}$ and ζ have the same distribution. We use the representation analogous to (13.40), but with φ_k replaced by φ^ε_k and ψ_k replaced by ζ^ε_k. Using our assumption on β

$$B^\varepsilon(t) = \int_0^t [D\beta(X^0(s))]\bar{Y}^\varepsilon(s)ds + T^\varepsilon_1(t),$$

and where using the tightness of $\{\|\bar{Y}^\varepsilon\|_{\infty,1}\}$, as in proof of Lemma 10.24, we find $\|T^\varepsilon_1\|_{\infty,1} \to 0$ in probability. Also as seen in Lemma 10.21, $E\|M^\varepsilon\|_{\infty,1} \to 0$ in probability as well. Next, using part (a) of Lemma 9.7 to handle $\{|\zeta^\varepsilon_i| \le \theta/a(\varepsilon)\}$ and the convergence of $\tilde{\zeta}^\varepsilon$ to $\tilde{\zeta}$, we have that for each $k = 1, \ldots, K$,

$$\int_{[0,t]\times\mathbb{R}_+} v_k(X^0(s))1_{[0,\lambda_k(X^0(s))]}(y)\zeta^\varepsilon_k(s, y)dyds$$

converges to

$$\int_{[0,t]\times\mathbb{R}_+} v_k(X^0(s))1_{[0,\lambda_k(X^0(s))]}(y)\tilde{\zeta}_k(s, y)dyds.$$

Using the convergence of $\bar{X}^\varepsilon$ to X^0, the fact that $\tilde{\zeta}^\varepsilon \in \hat{S}_m$, and the Cauchy-Schwarz inequality, we have that

$$\int_{[0,t]\times\mathbb{R}_+} \left[v_k(\bar{X}^\varepsilon(s))1_{[0,\lambda_k(X^\varepsilon(s))]}(y) - v_k(X^0(s))1_{[0,\lambda_k(X^0(s))]}(y)\right]\tilde{\zeta}^\varepsilon_k(s, y)dyds \to 0.$$

Finally, using part (a) of Lemma 9.7 again and the tightness of $\{\bar{X}^\varepsilon\}$

$$\int_{[0,t]\times\mathbb{R}_+} \zeta_k^\varepsilon(s,y)1_{\{|\zeta_k^\varepsilon|>\theta/a(\varepsilon)\}} \times \left[v_k(\bar{X}^\varepsilon(s))1_{[0,\lambda_k(\bar{X}^\varepsilon(s))]}(y) - v_k(X^0(s))1_{[0,\lambda_k(X^0(s))]}(y)\right] dyds \to 0.$$

Combining the last three convergence statements identifies the limit of the term that corresponds to C^ε. We see that $\bar{Y}$ solves the equation

$$\bar{Y}(t) = \int_0^t [D\beta(X^0(s))]\bar{Y}(s)ds + \sum_{k=1}^K \int_{[0,t]\times\mathbb{R}_+} v_k(X^0(s))1_{[0,\lambda_k(X^0(s))]}(y)\tilde{\zeta}_k(s,y)dyds,$$

i.e., (13.38) with f replaced by $\tilde{\zeta}$. Since $\tilde{\zeta}$ and ζ have the same distribution, we have from unique solvability of (13.37) the desired convergence (13.41). This completes the verification of (ii) and thus the proof of Theorem 13.18. □

13.4 Notes

Section 13.2 presents our only example of the large deviations analysis of processes with what are sometimes called "discontinuous statistics" [98]. Processes of this type are common in queueing theory [96, 141, 231], and appear in many applications. In general, the large deviation properties of processes with discontinuous statistics are hard to analyze [2, 137, 158, 195]. This is especially true when the discontinuities appear on the interior of the state space, rather than the boundary. In fact, very general results with an explicit identification of the rate function only exist for the case where two regions of smooth statistical behavior are separated by an interface of codimension one [97]. Our analysis of the WSLQ model is based in part on [104], but uses the representation in terms of a PRM.

Section 13.3 considers pure jump processes with a finite number of jump types where the jump rates and jump sizes can depend on the state. The analysis is based on their representation as the solution to a stochastic differential equation driven by a corresponding number of PRMs. Using similar representations, moderate deviation principle for weakly interacting pure jump processes with countably many jump types are studied in [48]. See the Notes at the end of Chap. 10 for a discussion of the qualitative differences between the SDE model of that chapter and the SDE models of the present chapter.

Part IV
Accelerated Monte Carlo for Rare Events

In the previous three sections of the book we developed methods for identifying the rate function and proving a large deviation principle for many different types of stochastic systems. With the rate function identified, the LDP, Varadhan's lemma, and related limits can be used to derive approximations for various probabilities and expected values. The approximation requires that one solve a variational problem, which may be very challenging. If this obstacle is overcome, then one also often obtains very useful qualitative information, such as how the rare event occurred or what events contributed the most to an expected value. However, the quantitative information is often fairly crude, in that large deviation theory provides only asymptotics of logarithms, i.e., exponential rates. In certain cases one may obtain more, such as upper bounds that are valid in the prelimit, but in general one does not identify other terms in an expansion for the quantity of interest, such as a polynomial term multiplying the exponential.

In many situations one seeks approximations that are more accurate than those given just by the decay rate. Among the possible choices, Monte Carlo methods are arguably among the most valuable general-purpose numerical tools currently available. They are indeed general purpose, being a primary means by which researchers in fields such biology, chemistry, and materials science probe the atomic-level details of complex systems; researchers in statistics solve high-dimensional and complicated problems of inference; and those in electrical engineering and operations research analyze and optimize large scale networks. And they are the main computational tool used in much of mathematical finance. However, rare events are in some sense the bane of Monte Carlo simulation algorithms.

Here are two fundamental examples to illustrate why rare events present a particular challenge for simulation methods. The first is the estimation of a probability determined by a single rare event. When straightforward Monte Carlo is used, the standard deviation of a single sample, i.e., a random variable that is 1 if the event occurs and 0 otherwise, is many times larger than the probability being estimated. One can reduce the variance by averaging a number of iid samples. However, if the event is very rare, then the number required to obtain an acceptable variance will make the approach impractical. The second example occurs in the approximation of ergodic averages for some given Markov process. The straightforward approach is to simulate the process and use the empirical measure of the simulated trajectory as

a surrogate for the true invariant distribution. In this case, if convergence to equilibrium depends on the occurrence of a (perhaps large) number of relatively rare events, then the trajectory must be simulated for an extraordinarily long time before a good approximation is obtained.

Although Monte Carlo methods have benefited greatly from decades of research into questions of analysis and design, when rare events play a significant role, our understanding of these issues is limited, and a great many algorithms are justified largely by numerical evidence. This is not surprising, given the highly nonlinear and/or singular equations typically associated with systems that involve rare events. However, a natural and perhaps underutilized tool in the analysis of Monte Carlo is large deviation theory, and in the remaining chapters of this book we will show how many of the ideas and constructions developed previously can be applied for problems of algorithm analysis, design, and optimization. In this book we consider only the first class of problems, i.e., estimating probabilities and expected values that depend on one or a few rare events. The use of ideas from large deviation theory for the second class is also under development, as in [1, 2, 3].

Chapter 14
Rare Event Monte Carlo and Importance Sampling

Suppose that in the analysis of some system, the value of a probability or expected value that is largely determined by one or a few events is important. Examples include the data loss in a communication network; depletion of capital reserves in a model for insurance; motion between metastable states in a chemical reaction network; and exceedance of a regulatory threshold in a model for pollution in a waterway. In previous chapters we have described how large deviation theory gives approximations for such quantities. The approximations take the form of logarithmic asymptotics, i.e., exponential decay rates.[1] For some purposes, especially when one is seeking *qualitative* information on how a rare event occurs, these approximations may be sufficient. For other purposes they may be inadequate, and a more accurate estimate is needed.

In this situation it is natural to turn to Monte Carlo approximation. However, as we will explain in some detail, the Monte Carlo approximation of small probabilities and related expected values also has difficulties owing to the role of rare events, and the design of reliable schemes requires great care. It turns out that many of the tools and constructions used for the large deviation analysis of a given problem can be used for the problem of designing Monte Carlo schemes that are efficient and reliable.

14.1 Example of a Quantity to be Estimated

To set the context, we consider a particular problem that arises frequently in various systems, especially communication theory. Let X^n be a Markov process with small noise of the form analyzed in Chap. 4. Thus we are given iid random vector fields $\{v_i(x), i \in \mathbb{N}_0, x \in \mathbb{R}^d\}$ on some probability space, and for each $x \in \mathbb{R}^d$ $v_i(x)$ has

[1] For certain special structures one can obtain more accurate approximations, e.g., approximations which identify both the exponential rate of decay as well as "pre-exponential" terms.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_14

distribution $\theta(\cdot|x)$, where $\theta(dy|x)$ is a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d$. Then the discrete time Markov process $\{X_i^n\}_{i\in\mathbb{N}_0}$ is constructed through the recursion

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n), \quad X_0^n = x, \tag{14.1}$$

and the continuous time interpolation is defined by

$$X^n(t) = X_i^n + \left[X_{i+1}^n - X_i^n\right](nt - i), \quad t \in [i/n, i/n + 1/n], i \in \mathbb{N}_0. \tag{14.2}$$

Let P_x denote probability conditioned on $X_0^n = x$ and integration with respect to P_x by E_x. The problem of interest is then to evaluate

$$p^n(x) \doteq P_x\left\{X^n \text{ enters } B \text{ before entering } A\right\}, \tag{14.3}$$

where A is open, B is closed, and $A \cap B = \emptyset$. For reasons that will become clear later, we explicitly record the initial condition in the notation, even though for many problems there may be only one initial condition of interest.

Under Conditions 4.3 and 4.7 or Conditions 4.3 and 4.8, Theorem 4.9 shows that for every $T \in (0, \infty)$, $\{X^n\}_{n\in\mathbb{N}}$ regarded as a collection of stochastic processes over the time horizon $[0, T]$ satisfies an LDP. Let I_T denote the corresponding rate function, and recall that $I_T(\phi) = 0$ characterizes the LLN limit trajectories of $\{X^n\}_{n\in\mathbb{N}}$. We will assume that A is an attractor of the LLN limit with nonempty interior and that B is in some sense rare. See Fig. 14.1. The trajectories in the figure are assumed to satisfy $I_T(\phi) = 0$ for all $T \in (0, \infty)$, and for all initial conditions $\phi(t)$ enters A as $t \to \infty$.

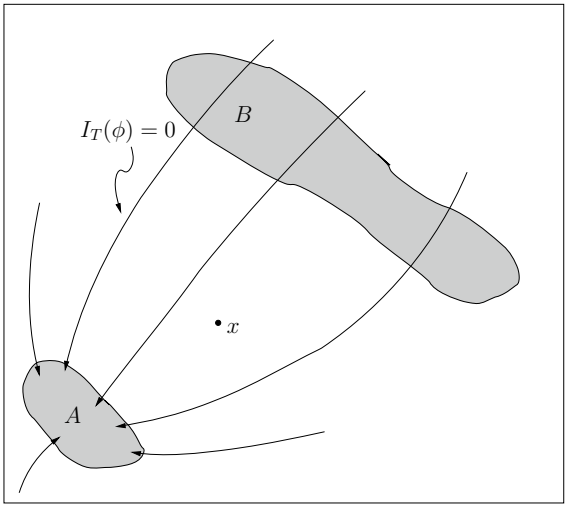

Fig. 14.1 Stability of the zero cost trajectories

Recall the notation

$$H(y,\alpha) = \log E \exp\{\langle \alpha, v_i(y)\rangle\}, \quad L(y,\beta) = \sup_{\alpha\in\mathbb{R}^d} [\langle \alpha,\beta\rangle - H(y,\alpha)]$$

from Chap. 4. Under appropriate regularity conditions the large deviation principle for $\{X^n\}$ implies

$$-\frac{1}{n}\log P_x\left\{X^n \text{ enters } B \text{ before entering } A\right\} \to V(x), \tag{14.4}$$

where

$$V(x) \doteq \inf\left[\int_0^T L(\phi(t),\dot{\phi}(t))dt : \phi \in C_{x,T}, T < \infty\right],$$

and with

$$C_{x,T} \doteq \{\phi(0) = x, \phi(t) \in B \text{ for some } t \in [0,T] \text{ and } \phi(s) \notin A \text{ for } s \in [0,t]\}.$$

The proof of (14.4) is typically carried out by reducing the analysis to that over a finite time interval and then invoking the large deviation principle for $\{X^n\}$ over finite time intervals (see Condition 15.18 and Proposition 15.19).

14.1.1 Relative Error

Recall that the problem of interest is to estimate the probability

$$p^n(x) \doteq P_x\left\{X^n \text{ enters } B \text{ before entering } A\right\}.$$

Let $C_x \doteq \cup_{T\in(0,\infty)} C_{x,T}$ be the trajectories that enter B before entering A after starting at x. To apply straightforward Monte Carlo, one would simulate K independent copies $\left\{X^{n,k}\right\}_{k=1,\ldots,K}$ of X^n, and then form the estimate

$$\hat{p}^n_K(x) \doteq \frac{1}{K}\sum_{k=1}^{K} 1_{\{X^{n,k}\in C_x\}}.$$

Note that k here is the index of the sample and *not* the time step, and that depending on the problem, the computational expense of simulating a single trajectory can vary greatly.

The variance of a single sample is

$$\begin{aligned}\mathrm{Var}\left(1_{\{X^{n,k}\in C_x\}}\right) &= E_x\left[1_{\{X^{n,k}\in C_x\}} - E_x 1_{\{X^{n,k}\in C_x\}}\right]^2 \\ &= E_x 1_{\{X^{n,k}\in C_x\}} - \left(E_x 1_{\{X^{n,k}\in C_x\}}\right)^2 \\ &= p^n(x) - [p^n(x)]^2,\end{aligned}$$

and if $p^n(x)$ is small $[p^n(x)]^2$ can be neglected. The *relative error*, which is defined by the ratio of the standard deviation of $\hat{p}^n_K(x)$ and $p^n(x)$, is then

$$\frac{\sqrt{\mathrm{Var}(\hat{p}^n_K(x))}}{p^n(x)} \approx \sqrt{\frac{p^n(x)}{K}} \cdot \frac{1}{p^n(x)} = \sqrt{\frac{1}{Kp^n(x)}}.$$

When considering rare events it is essential to use relative error as the figure of merit, since the variance can be small (or conversely big in some situations involving expected values) in absolute terms, and yet provide an estimate that is orders of magnitude off, and therefore quite inaccurate in a relative sense.

For the example problem, to obtain a relative error of roughly size 1 requires $K \approx (p^n(x))^{-1}$ samples. This is computationally infeasible when $p^n(x)$ is very small (e.g., 10^{-5}), or even when $p^n(x)$ is not so small if the computational effort needed to generate samples of X^n is great. For example, consider the problem of estimating the probability of an unusually large concentration of pollutant in a model for ground water contamination. The generation of each sample would typically involve solving a time dependent stochastic partial differential equation, and hence each sample is computationally expensive.

An alternative to standard Monte Carlo is to construct iid random variables $\gamma^n_1, \ldots, \gamma^n_K$ with $E_x\gamma^n_1 = p_n(x)$, and use the estimator

$$\hat{q}^n_K(x) \doteq \frac{\gamma^n_1 + \cdots + \gamma^n_K}{K}.$$

The performance as with ordinary Monte Carlo is determined by variance of γ^n_1, and since the estimator is unbiased [i.e., $E_x\gamma^n_1 = p_n(x)$], minimizing the variance is equivalent to minimizing $E_x\left(\gamma^n_1\right)^2$.

It is straightforward to obtain bounds on the best possible performance. For example, by Jensen's inequality and (14.4)

$$-\frac{1}{n}\log E_x\left(\gamma^n_1\right)^2 \le -\frac{2}{n}\log E_x\gamma^n_1 = -\frac{2}{n}\log p_n(x) \to 2V(x). \tag{14.5}$$

Hence the decay rate for the second moment cannot possible exceed $2V(x)$. An estimator is called **asymptotically efficient** if

$$\liminf_{n\to\infty} -\frac{1}{n}\log E_x\left(\gamma_1^n\right)^2 = 2V(x),$$

i.e., the optimal decay rate is achieved.

One could consider more stringent measures of performance, such as **bounded relative error**: there is $K < \infty$ such that

$$\limsup_{n\to\infty} \frac{\sqrt{\mathrm{Var}(\gamma_1^n)}}{p^n(x)} \le K,$$

or vanishing relative error [183]. While bounded relative error is certainly a desirable feature, it can be achieved only for the most elementary process models and events, such as the probability that a homogeneous random walk escapes from a set with simple structure. State dependence in the dynamics or even more complex situations (e.g., Markov modulated noise, multiscale dynamics) usually make it very difficult to construct schemes with (provable) bounded relative error. However, while asymptotic efficiency may be a more practical figure of merit, the logarithmic scaling can wipe out important terms in the variance that depend on other system parameters, such as another exponential scaling in terms of a time variable. Thus while more flexible and realistic than bounded relative error, it must be used with caution, and in all cases no single performance measure can replace a careful analysis of the variance and its dependence on all important system parameters. If different methods vary significantly with regard to the computational cost of implementation, then that aspect should also be factored into the performance measure.

We will discuss two well known methods used to design random variables $\{\gamma_k^n\}$ that are unbiased, which can be simulated with reasonable effort, and for which one may hope to get good performance: **importance sampling** and **splitting schemes**. For the remainder of this chapter and in Chap. 15 we focus on importance sampling (IS), and then in Chap. 16 turn to splitting. While many of the constructions needed for the successful design and analysis are essentially the same for both approaches, there are also interesting differences, some of which will be discussed at the end of Chap. 17.

We stress that for any approach to problems of rare event estimation a rigorous and *independent* analysis of performance is very important, since typical methods one would use to assess accuracy of the estimates (e.g., the empirical variance) are prone to the same difficulties and errors which can affect the estimates themselves. This point will be illustrated via a numerical example in the next section.

14.2 Importance Sampling

The basic formulae of importance sampling are as follows. Suppose that X has distribution θ, where X takes values in a Polish space S. Suppose that $G : S \to \mathbb{R}$ is Borel measurable and integrable with respect to θ, and the goal is to estimate

$m = EG(X)$. Consider an alternative sampling distribution π. It is required that θ be absolutely continuous with respect to π, so that the Radon-Nikodym derivative (also called the likelihood ratio in this context) $f(x) \doteq (d\theta/d\pi)(x)$ exists. Iid samples $Y_0, Y_1, \ldots$ with distribution π are generated, and the estimate

$$\bar{m}_K \doteq \frac{1}{K} \sum_{k=0}^{K-1} G(Y_k) f(Y_k)$$

is formed. Since

$$EG(Y_k) f(Y_k) = \int_S G(x) f(x) \pi(dx) = \int_S G(x) \theta(dx) = m,$$

$\bar{m}_K$ is an unbiased estimate of m, with a rate of convergence determined by

$$\text{var}\left[G(Y_0) f(Y_0)\right] = \int_S G(x)^2 f(x) \theta(dx) - \left[\int_S G(x) \theta(dx)\right]^2.$$

Standard Monte Carlo corresponds to $f = 1$, and the goal of importance sampling is to choose f in such a way that: (i) the variance is lowered significantly, and (ii) sampling from π is not too difficult. Note that minimizing the variance with respect to f is equivalent to minimizing the second moment, and so if posed as an optimization problem, one can use the simpler second moment in lieu of variance. Note also that without further restriction on the class of sampling measures the problem is in some sense ill-posed. For example, suppose θ is supported on $[0, \infty)$, and $\theta(dx) = g(x)\,dx$. Let $G(x) \doteq x$ so that $m = EX \neq 0$ and let $\pi(dx) \doteq m^{-1} x g(x)\,dx$. Then θ is absolutely continuous with respect to π, with $f(x) = m/x$. Furthermore,

$$\text{var}\left[Y_0 f(Y_0)\right] = \int_{[0,\infty)} x^2 f(x) \theta(dx) - m^2 = 0.$$

However, such a distribution π is of little use in practice since it requires knowledge of m, the very thing we want to estimate! Instead of this unconstrained optimization, one typically seeks to minimize over parameterized families of alternative sampling distributions.

14.2.1 Importance Sampling for Rare Events

We now return to the discrete time model of Sect. 14.1. Recall the notation

$$H(y, \alpha) = \log E \exp\left\{\langle \alpha, v_i(y) \rangle\right\}, \quad L(y, \beta) = \sup_{\alpha \in \mathbb{R}^d} \left[\langle \alpha, \beta \rangle - H(y, \alpha)\right], \quad (14.6)$$

and consider the problem of estimating $p^n(x)$ as defined in (14.3). When $p^n(x)$ is small (e.g., on the order of say 10^{-6}) ordinary Monte Carlo attempts to estimate this number as a convex combination of 0's and 1's. The goal of importance sampling (and indeed any accelerated Monte Carlo scheme) is to produce estimators whose distribution is more closely clustered around the target value of 10^{-6}.

As we have just noted, the problem of optimizing over all changes of measure is in some sense ill-posed, and thus the first question is, "what are natural changes of measure?" A hint is provided by the analysis of Chap. 4. The control measures $\bar{\mu}_i^n$ of the weak convergence approach correspond to a change of measure for the noise sequence. An a posteriori conclusion of the large deviation analysis is that *exponential* changes of measure are asymptotically optimal in the representation. (See, for example, the measures γ defined in part (g) of Lemma 4.16, and their use in the proof of the Laplace lower bound proof in Sect. 4.7.) Exponential changes of measure have a finite dimensional parameterization, and thus are convenient to work with. Recalling that $\{v_i(x), i \in \mathbb{N}\}$ are iid with distribution $\theta(dv|x)$ and associated log moment generating functions $H(x, \alpha)$, this suggests that measures of the form

$$\eta_\alpha(dv|x) = e^{\langle \alpha, v\rangle - H(x,\alpha)}\theta(dv|x)$$

be used to generate the noise sequence under the new distribution. We will show later on that changes of measure within this class are sufficient for asymptotic optimality. The parameter α can be thought of as a control, which is selected to produce good performance of the resulting Monte Carlo scheme. In this context η_α is sometimes referred to as an **exponential tilt** of θ, with α the **tilt parameter**.

While more complicated dependencies could be considered, it will turn out (for the models of Chap. 4) that allowing α to depend on time and the current state of the simulated trajectory will be sufficient for asymptotic optimality. Thus a control scheme (i.e., a change of measure) will be characterized as a collection of measurable mappings $\alpha_i^n : \mathbb{R}^d \to \mathbb{R}^d$, defined for $i \in \mathbb{N}_0$. The generation of a single sample as well as the likelihood ratio needed to estimate $p^n(x)$ then proceeds as follows.

We initialize with $Y_0^n = x$. A sequence of noises w_i^n and states Y_{i+1}^n are then generated recursively by

$$P_x\left\{w_i^n \in dv \middle| \mathscr{F}_i^n\right\} = \eta_{\alpha_i^n(Y_i^n)}(dv|Y_i^n), \text{ with } \mathscr{F}_i^n = \sigma\left(w_j^n, j = 0, \ldots, i-1\right)$$

and

$$Y_{i+1}^n = Y_i^n + \frac{1}{n}w_i^n.$$

The simulation proceeds up until

$$N^n \doteq \inf\left\{i : Y_i^n \in A \cup B\right\},$$

and we define $Y^n(t)$ to be the piecewise linear interpolation, so that $1_{\{Y^n \in C_x\}}$ means B was entered before A. The likelihood ratio is then

$$\prod_{i=0}^{N^n-1} \frac{d\theta(\cdot|Y_i^n)}{d\eta_{\alpha_i^n(Y_i^n)}(\cdot|Y_i^n)}(w_i^n) = \prod_{i=0}^{N^n-1} e^{-\left\langle \alpha_i^n(Y_i^n), w_i^n \right\rangle + H(Y_i^n, \alpha_i^n(Y_i^n))},$$

and the estimate based on a single sample is thus

$$1_{\{Y^n \in C_x\}} \prod_{i=0}^{N^n-1} e^{-\left\langle \alpha_i^n(Y_i^n), w_i^n \right\rangle + H(Y_i^n, \alpha_i^n(Y_i^n))}. \tag{14.7}$$

As discussed previously, one then simulates K independent copies of (14.7) and takes the sample average, where K depends on the variance of a single sample and the desired accuracy.

We recall that performance is determined by the variance of a single sample, and minimizing this is the same as minimizing the second moment. The second moment of (14.7) is

$$E_x \left[1_{\{Y^n \in C_x\}} \prod_{i=0}^{N^n-1} e^{-2\left\langle \alpha_i^n(Y_i^n), w_i^n \right\rangle + 2H(Y_i^n, \alpha_i^n(Y_i^n))} \right],$$

which when rewritten in terms of the distribution of the *original process* $\left\{X_i^n\right\}$ takes the form

$$E_x \left[1_{\{X^n \in C_x\}} \prod_{i=0}^{N^n-1} e^{-\left\langle \alpha_i^n(X_i^n), v_i(X_i^n) \right\rangle + H(X_i^n, \alpha_i^n(X_i^n))} \right].$$

14.2.2 Controls Without Feedback, and Dangers in the Rare Event Setting

Since one of the classical approaches to the large deviation lower bound involves a change of measure argument, it is natural to ask if there is a connection between the change of measure (equivalently control measure) used there to prove bounds for a particular event or expected value, and a change of measure that would produce a good IS scheme for that same event. Note that there are actually many changes of measure that *could* be used to prove the lower bound. Here we mean the one that is typically used in the proof, and which uses a deterministic sequence $\alpha_i^n(x)$ that depends on i but not x, which we refer to as an "open loop" control. It turns out that in some special circumstances one can achieve asymptotic optimality within the class of open loop controls (e.g., [232]), and for some time it was generally thought that using this lower bound change of measure would work well in general. This turned out to be false, and indeed the class of schemes that had been considered up to that time turned out to be, in general, inadequate. In this section we illustrate the issue through an example due to [150]. The techniques we develop to understand the particular example are broadly useful for understanding rare event importance sampling. Of

special importance is the game characterization of performance described in the next section.

The example is as follows. Suppose that $v_i(X_i^n)$ are in fact independent of X_i^n, i.e., that they are just an iid sequence with distribution θ. We further assume $d = 1$ and that $X_0^n = 0$ (for the rest of this section we write P and E rather than P_0 and E_0). Then X_i^n is a random walk, and $X_n^n = \frac{1}{n}\sum_{i=0}^{n-1} v_i$ is just the sample mean, i.e., we are in the setting of Cramér's theorem with rate function $L(\beta)$ (see Sect. 3.1.6). Let $B \subset \mathbb{R}$, and suppose we want to estimate $P\left\{X_n^n \in B\right\}$ by importance sampling.

The heuristic just described to construct an alternative sampling distribution is straightforward to implement. Let β^* solve $\inf[L(\beta) : \beta \in B]$ (and assume the infimum over the interior and closure of B are the same). If α^* is dual to β^*, i.e., if α^* is the point that maximizes in the relation $L(\beta^*) = \sup_{\alpha\in\mathbb{R}}[\alpha\beta^* - H(\alpha)]$, then as discussed in Chap. 4 [see part (g) of Lemma 4.16], the mean of $\eta_{\alpha^*}(dv) \doteq e^{\alpha^* v - H(\alpha^*)}\theta(dv)$ is exactly β^*, and $\mu_i^n = \eta_{\alpha^*}$ is the control one could use to prove the large deviation lower bound. Since this problem is over a fixed time horizon, the single sample estimate is just

$$1_{\{Y_n^n \in B\}} \prod_{i=0}^{n-1} e^{-\alpha^* w_i^n + H(\alpha^*)} = 1_{\{Y_n^n \in B\}} e^{-n\left[\alpha^* Y_n^n - H(\alpha^*)\right]}.$$

One can now describe the shortcomings of the open loop heuristic. Assume that θ is Gaussian $N(0, 1)$ and consider the *nonconvex* set $B = (-\infty, -0.25] \cup [0.2, \infty)$ (see Fig. 14.2). For this process $L(\beta) = \beta^2/2$, $H(\alpha) = \alpha^2/2$, and $\alpha^* = \beta^* = 0.2$, and the change of measure will shift the mean to this value. If all goes according to plan and the simulated trajectory ends up near β^*, then the likelihood ratio will be near $\exp\{-n\left[\alpha^*\beta^* - H(\alpha^*)\right]\} = \exp\{-nL(\beta^*)\}$. Thus the estimator is either zero or close to the large deviation approximation to the probability, which is just the sort of qualitative behavior that is needed. However, it is also possible that an event that is rare under the $\eta_{\alpha^*}(dv)$ distribution may occur, and one can end up with Y_n^n that is in the interval $(-\infty, -0.25]$. Such an occurrence is labeled the "rogue" simulation in Fig. 14.2 . When this happens, the likelihood ratio will be approximately

$$\exp\left\{-n\left[\alpha^*\bar{\beta} - H(\alpha^*)\right]\right\} = \exp\left\{n\left[0.2 \times 0.25 + \frac{1}{2}(0.2)^2\right]\right\}.$$

This quantity grows exponentially in n and, while the event itself might be rare, it happens often enough that the variance of the estimate is very large, and even larger than standard Monte Carlo!

In this example the true probability for $n = 60$ is $p^n = 8.71 \times 10^{-2}$, which can be calculated using the known distribution of X_n^n. The data in Table 14.1 reflect four trials of $K = 5000$ replications. The "standard error" is the estimated standard deviation for the entire trial, and $\hat{\mathfrak{S}}^n$ is the estimate of the second moment based on the data.

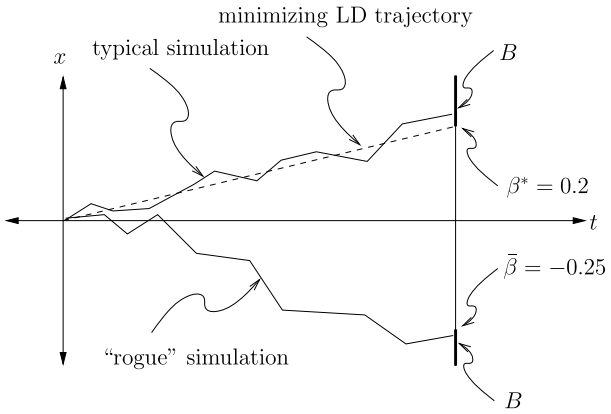

Fig. 14.2 An expected trajectory and a rogue trajectory

Table 14.1 Importance sampling implementation based on an open loop control

	No. 1	No. 2	No. 3	No. 4
Estimate $\hat{p}^n$ ($\times 10^{-2}$)	18.36	7.51	5.95	8.02
Standard error ($\times 10^{-2}$)	7.41	1.37	0.118	1.80
95% confidence interval ($\times 10^{-2}$)	[3.83, 32.89]	[4.82, 10.20]	[5.72, 6.18]	[4.49, 11.55]
Number of "rogue" trajectories	3	1	0	1
$(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$	−1.96	0.0210	1.62	−0.193

The first, second and fourth trials have 3, 1 and 1 "rogue" trajectories, respectively. In contrast the third has none. While the third estimate has a small standard error and associated confidence interval, the interval does not contain the true value. The estimate is smaller than the true value, reflecting the fact that the estimate has never sampled from the interval $(-\infty, -0.25]$, and is therefore in some sense providing an estimate of only the probability to end in $[0.2, \infty)$. Both the estimate and the estimate of the standard deviation are misleading, and it is in fact the same difficulties that affect the estimation of p_n that make the confidence interval essentially useless, though one does not a priori know this is the case. Because of this, an independent theoretical (and not only data driven) analysis of errors is important for rare event Monte Carlo estimation. All of the other trials include at least one rogue trajectory, which is needed to avoid the bias of trial 3. The estimates may be far from the true value, but in this case at least the confidence intervals are correctly indicating this fact. If the estimates were accurate, $(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$ should be close to 2 for asymptotic optimality. This appears to be to some degree valid for trial 3, but for reasons mentioned previously this is misleading.

One could argue that the difficulties encountered in this example can be avoided by splitting the problem into that of estimating two half-infinite intervals. While such an approach would work here, it will fall apart as soon as one considers problems in

higher dimensions or even slightly more complicated dynamics. What is needed is a *global* approach that properly controls the likelihood ratio for any possible simulated trajectory.

14.2.3 A Dynamic Game Interpretation of Importance Sampling

Further insight into the difficulties of IS in the rare event setting can be obtained by modeling the performance in terms of prelimit *small noise stochastic game* and limiting *deterministic differential game*. Although in this section we develop this connection only for the simple random walk model just discussed, it is easily adapted to other situations. Suppose that for the iid random walk model and problem of the last section we consider, instead of the constant control α^* suggested by the standard heuristic, a collection of sampling controls of the general form $\alpha_i^n(x)$, and in particular assume

$$\alpha_i^n(Y_i^n) = u(Y_i^n, i/n)$$

for some smooth function $u : \mathbb{R} \times [0, 1] \to \mathbb{R}$ (we assume $d = 1$ for simplicity). In this case, the second moment of a single sample, and hence the performance of the scheme, is given by the exponential integral

$$E\left[1_{\{Y_n^n \in B\}} \prod_{i=0}^{n-1} e^{-2u(Y_i^n, i/n)w_i^n + 2H(u(Y_i^n, i/n))}\right],$$

which we can rewrite in terms of the original process as

$$E\left[1_{\{X_n^n \in B\}} \prod_{i=0}^{n-1} e^{-u(X_i^n, i/n)v_i + H(u(X_i^n, i/n))}\right].$$

Note that this is the sort of Laplace functional for which relative entropy representations are derived in Chaps. 3 and 4 (see for example Proposition 3.1 and Theorem 4.5), although previously we have assumed (e.g., in Proposition 2.3) that the quantity appearing in the exponent was at least bounded either from above or below. This boundedness will not hold if the support of $\{v_i\}$ is unbounded above and below. Setting aside the issue of boundedness, the quantity in the last display is still expected to scale exponentially in n, and thus it is natural to consider the log transform. Using the same notation for the controls (measures) and controlled processes as in Sect. 4.2, we formally have

$$-\frac{1}{n}\log E\left[1_{\{X_n^n\in B\}}\prod_{i=0}^{n-1}e^{-u(X_i^n,i/n)v_i+H(u(X_i^n,i/n))}\right]=\inf_{\{\bar{\mu}_i^n\}}E\left[\frac{1}{n}\sum_{i=0}^{n-1}R\left(\bar{\mu}_i^n\|\theta\right)\right.$$
$$\left.+\frac{1}{n}\sum_{i=0}^{n-1}\left[u(\bar{X}_i^n,i/n)\bar{v}_i-H(u(\bar{X}_i^n,i/n))\right]+\infty 1_{B^c}\left(\bar{X}_n^n\right)\right].$$

Keeping in mind that to minimize the variance we will supremize the right hand side over $u(\cdot,\cdot)$, the optimal variance is then characterized in terms of a discrete time small noise stochastic game. One player (corresponding to u) is given in feedback form as a function of the state and seeks to maximize. The other player (with controls $\{\bar{\mu}_i^n\}$) arises from the representation, and seeks to minimize. This player's controls can depend on the state (in fact the whole history) as well as u, and since it seeks to minimize the cost, must drive the process into B at time n with probability one. Note that the class of open loop controls as they would be used in IS schemes correspond to eliminating state feedback in u, i.e., restricting to the form $u(x,s)=u(s)$.

One can calculate the limit in the last display using the same weak convergence methods as those used in Chap. 4 to study the LDP for $\{X^n\}$, and for a fixed bounded and continuous control u the limit is characterized by the optimization problem

$$J[u]=\inf_{\phi}\left[\int_0^1\left[u(\phi(t),t)\dot{\phi}(t)-H(u(\phi(t),t))+L(\dot{\phi}(t))\right]dt+\infty 1_{B^c}\left(\phi(1)\right)\right],$$

where the infimum is over absolutely continuous ϕ with $\phi(0)=0$.

The quantity $J[u]$ gives the rate of decay of the second moment of the IS scheme that uses the sampling control $\alpha_i^n(Y_i^n)=u(Y_i^n,i/n)$ to dynamically choose the change of measure. For the purposes of IS scheme selection, one can consider this simpler limit problem which characterizes the rate of decay. Thus we consider $U=\sup_{u(\cdot,\cdot)}J[u]$. This is a type of deterministic differential (or dynamic) game, where $\dot{\phi}$ (replacing $\{\bar{\mu}_i^n\}$) attempts to minimize (in open loop form) and u attempts to maximize (in feedback form, but u must be selected before ϕ is chosen).

Suppose we extend the definition to allow for an arbitrary initial condition (x,t) (i.e., we consider the cost over $[t,1]$ and with $\phi(t)=x$), and denote the corresponding optimal rate of decay by $U(x,t)$. Let U_t be the partial with respect to t and $DU(x,t)$ the gradient in x. Then $U(x,t)$ will be a viscosity solution to

$$U_t(x,t)+\sup_{\alpha\in\mathbb{R}}\inf_{\beta\in\mathbb{R}}\left[DU(x,t)\beta+\alpha\beta-H(\alpha)+L(\beta)\right]=0 \tag{14.8}$$

and the terminal condition

$$U(x,1)=\infty \text{ for } x\in B^c \text{ and } U(x,1)=0 \text{ for } x\in B. \tag{14.9}$$

For properties of viscosity solutions that will be used here (though these arguments are only intended to be formal), we refer to [14].

We will not delve deeply into the nuances of differential games, since this game has a special structure which allows a reduction to a much simpler problem. Using the Minimax theorem [233] and that L is the Legendre-Fenchel transform of H, we observe that

$$\begin{aligned}\sup_{\alpha\in\mathbb{R}}\inf_{\beta\in\mathbb{R}}\left[p\beta+\alpha\beta-H(\alpha)+L(\beta)\right] &= \inf_{\beta\in\mathbb{R}}\sup_{\alpha\in\mathbb{R}}\left[p\beta+\alpha\beta-H(\alpha)+L(\beta)\right]\\ &= \inf_{\beta\in\mathbb{R}}\left[p\beta+2L(\beta)\right].\end{aligned}$$

Not surprisingly then, the PDE (14.8) is closely related to ones that are connected with the large deviation rate function for the original process. Define $\mathbb{H}(p)\doteq\inf_{\beta\in\mathbb{R}}[p\beta+L(\beta)]$. This form of the Legendre transform, which is natural when discussing PDEs, is related to the form usually used in large deviation theory (e.g., a log moment generating function) by $\mathbb{H}(p)=-H(-p)$. Then the Isaacs equation (14.8) can be rewritten as

$$U_t(x,t)+2\mathbb{H}(DU(x,t)/2)=0.$$

Suppose we consider the probability $P\{X_n^n\in B\}$, but generalize to allow an arbitrary initial point x and starting time i/n. Let

$$V^n(x,i/n)\doteq-\frac{1}{n}\log P\left\{x+\frac{1}{n}\sum_{j=i}^{n-1}v_i\in B\right\}.$$

If $i/n\to t$ as $n\to\infty$, then by Cramér's theorem $V^n(x,i/n)\to V(x,t)\doteq\inf[(1-t)L(\beta):x+(1-t)\beta\in B]$, and it is straightforward to verify that $V(x,t)$ is a viscosity solution to the problem with the same terminal condition (14.9) as U and the PDE

$$V_t(x,t)+\mathbb{H}(DV(x,t))=0. \tag{14.10}$$

Using the fact that a comparison principle holds for viscosity solutions to these PDE, is follows that $U(x,t)=2V(x,t)$, which is consistent with the claim made previously (see Sect. 14.1.1) that the best possible rate of decay for the second moment of any IS is precisely twice the large deviation rate. It in fact suggests more, which is that within the class of IS schemes based on feedback and exponential changes of measure, one can in fact achieve this best decay rate.

The situation just described, which we have presented here in the context of Cramér's theorem and for a particular event, is in fact generic under the small noise large deviation scaling [116]. Note the remarkable fact that the equation for U, which models a *game*, turns out to be equivalent to the equation for V, which models a *calculus of variations* or *control* problem [14]. The Isaacs equation for U identifies (at least for smooth solutions) optimal controls for both players. Evaluating the infimum in β in (14.8) gives

$$\begin{aligned}
&\sup_{\alpha\in\mathbb{R}} \inf_{\beta\in\mathbb{R}} [DU(x,t)\beta + \alpha\beta - H(\alpha) + L(\beta)] \\
&\quad = \sup_{\alpha\in\mathbb{R}} \left[-\sup_{\beta\in\mathbb{R}} [-(DU(x,t)+\alpha)\beta - L(\beta)] - H(\alpha) \right] \\
&\quad = -\inf_{\alpha\in\mathbb{R}} [H(-DU(x,t)-\alpha) + H(\alpha)] .
\end{aligned}$$

Suppose that the distribution of v_i does not concentrate on a single point, so that H is strictly convex. Then the optimal α is given by the unique solution of $H'(\alpha) = H'(-DU(x,t)-\alpha)$, which is $\alpha = -DU(x,t)/2$. In terms of the value function associated with the large deviation control problem this is simply $\alpha = -DV(x,t)$. Although it is not needed or used, the optimal control for the large deviation player [but for the second moment, and not for the original event!] can be found by solving $L'(\beta) = -DU(x,t)/2$.

It turns out that one does not need to solve the game or control problem, and in fact the construction of suitable *subsolutions* to the associated PDE (14.10) will be sufficient for a certain level of performance, in a sense that will be made precise in Chap. 15. This is a significant simplification, because for many interesting classes of problems such subsolutions can be constructed explicitly. The reason subsolutions suffice is because the goal in algorithm design is lower bounds on the rate of decay of the second moment. The verification of these one-sided bounds require only certain inequalities, which coincide with the subsolution definition.

In the next section the definitions of classical and piecewise classical subsolution are given. It will turn out to be much easier for many problems to find appropriate piecewise classical subsolutions, so this generalization is important. We also spell out how the various subsolutions generate sampling schemes.

14.3 Subsolutions

We will describe the subsolutions needed for both finite time problems (as in Sect. 14.2.2) and exit probability problems (as in Sect. 14.1.1). We begin with the finite time problem, which generalizes the example used in Sect. 14.2.2. Processes will be of interest on a continuous time interval of the form $[0, T]$, $T < \infty$, and to simplify the notation we assume Tn is an integer. As in Section 14.1 let $\{v_i(x), i \in \mathbb{N}_0, x \in \mathbb{R}^d\}$ be iid random vector fields given on some probability space with the property that for each $x \in \mathbb{R}^d$ $v_i(x)$ has distribution $\theta(\cdot|x)$, where $\theta(dy|x)$ is a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d$. Recall the discrete time Markov process $\{X_i^n\}_{i=0,\ldots,Tn}$ defined by the recursion

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n), \quad X_0^n = x_0,$$

and the continuous time interpolation defined by

$$X^n(t) = X_i^n + \left[X_{i+1}^n - X_i^n\right](nt - i), \quad t \in [i/n, i/n + 1/n], i = 0, 1, \ldots, Tn.$$

Also, we assume $H(x, \alpha) = \log E \exp\{\langle \alpha, v_i(x)\rangle\} < \infty$ for all $x \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}^d$.

The importance sampling problem of interest is to estimate

$$P_{x_0}\left\{X^n(T) \in B\right\},$$

where $B \subset \mathbb{R}^d$. As for the one dimensional setting considered in Sect. 14.2.3, the PDE that characterizes the large deviation rate and half the optimal rate of decay for an asymptotically optimal importance sampling scheme is

$$V_t(x, t) + \mathbb{H}(x, DV(x, t)) = 0 \tag{14.11}$$

for $(x, t) \in \mathbb{R}^d \times [0, T)$, where $\mathbb{H}(x, p) = -H(x, -p)$. The terminal condition is

$$V(x, T) = \infty \text{ for } x \in B^c \text{ and } V(x, T) = 0 \text{ for } x \in B. \tag{14.12}$$

Definition 14.1 A function $\bar{V} : \mathbb{R}^d \times [0, T] \to \mathbb{R}$ is a **classical sense subsolution** (or simply a classical subsolution) if it is continuously differentiable in both variables and if

$$\bar{V}_t(x, t) + \mathbb{H}(x, D\bar{V}(x, t)) \geq 0$$

for all $(x, t) \in \mathbb{R}^d \times [0, T)$ and

$$\bar{V}(x, T) \leq \infty \text{ for } x \in B^c \text{ and } \bar{V}(x, T) \leq 0 \text{ for } x \in B.$$

Note that the condition $\bar{V}(x, T) \leq \infty$ for $x \in B^c$ is vacuous. Let $\wedge_{j=1}^J a_j$ denote the minimum of real numbers a_j, $j = 1, \ldots, J$.

Definition 14.2 A function $\bar{V} : \mathbb{R}^d \times [0, T] \to \mathbb{R}$ is a **piecewise classical sense subsolution** (or simply a piecewise classical subsolution) if the following hold. There are $J \in \mathbb{N}$ and functions $\bar{V}^{(j)} : \mathbb{R}^d \times [0, T] \to \mathbb{R}$, $j = 1, \ldots, J$, that are continuously differentiable in both variables and satisfy

$$\bar{V}_t^{(j)}(x, t) + \mathbb{H}(x, D\bar{V}^{(j)}(x, t)) \geq 0$$

for all $(x, t) \in \mathbb{R}^d \times [0, T)$. Moreover $\bar{V}(x, t) \doteq \wedge_{j=1}^J \bar{V}^{(j)}(x, t)$ satisfies

$$\bar{V}(x, T) \leq \infty \text{ for } x \in B^c \text{ and } \bar{V}(x, T) \leq 0 \text{ for } x \in B.$$

Example 14.3 Consider again the iid random walk example of Sect. 14.2.2, where $H(\alpha) = \log E e^{\alpha v_i}$ and $\{v_i\}_{i \in \mathbb{N}}$ are iid real random variables with mean zero. Without loss of generality we take the time horizon $T = 1$. The set B in the example was of the form $(-\infty, \bar{\beta}] \cup [\beta^*, \infty)$, with $\bar{\beta} < 0 < \beta^*$. The solution to (14.11) and (14.12) is

$$V(x,t) = \inf\left[(T-t)L(\beta) : x + (T-t)\beta \in (-\infty, \bar{\beta}] \cup [\beta^*, \infty)\right].$$

For this example it is natural to look for a piecewise classical subsolution as the minimum of two functions. One can easily construct solutions to the PDE by assuming the simple form $-ax + bt + c$ and requiring that $b + \mathbb{H}(-a) = b - H(a) = 0$ hold. If $\hat{\alpha}$ and $\hat{\beta}$ are convex dual points, i.e.,

$$L(\hat{\beta}) = \sup_{\alpha \in \mathbb{R}} \left[\alpha\hat{\beta} - H(\alpha)\right] = \hat{\alpha}\hat{\beta} - H(\hat{\alpha}),$$

we obtain the solution $-\hat{\alpha}(x - \hat{\beta}) + (L(\hat{\beta}) - \hat{\alpha}\hat{\beta})[1-t]$, which corresponds to the terminal condition $-\hat{\alpha}(x - \hat{\beta})$. Note that since $Ev_i = 0$ Jensen's inequality implies $H(\hat{\alpha}) \geq 0$, and so $\hat{\alpha}\hat{\beta} \geq 0$. Thus $\hat{\beta} > 0$ if and only if $\hat{\alpha} > 0$.

We conclude that the two solutions

$$\bar{V}^{(1)}(x,t) = -\alpha^*(x - \beta^*) + (L(\beta^*) - \alpha^*\beta^*)[1-t],$$
$$\bar{V}^{(2)}(x,t) = -\bar{\alpha}(x - \bar{\beta}) + (L(\bar{\beta}) - \bar{\alpha}\bar{\beta})[1-t],$$

which correspond to the terminal conditions indicated in Fig. 14.3, generate the piecewise classical subsolution $\bar{V} \doteq \bar{V}^{(1)} \wedge \bar{V}^{(2)}$. Note that since the (α, β) pairs are convex dual points, α^* and $\bar{\alpha}$ generate changes of measure with the means β^* and $\bar{\beta}$, respectively. See Fig. 14.4. The dotted line in the figure represents points (x,t) for which $\bar{V}^{(1)}(x,t) = \bar{V}^{(2)}(x,t)$. Note that the *subsolution* $\bar{V}(x,t)$ has a much simpler structure than the *solution* $V(x,t)$, but it also has the same (maximal) value at $(0,0)$, namely $\left[L(\beta^*) \wedge L(\bar{\beta})\right]$.

Consider next the problem of entering a rare set B before a typical set A (Fig. 14.5). Thus the importance sampling problem is to estimate

$$P_{x_0}\left\{X^n \text{ enters } B \text{ before entering } A\right\}.$$

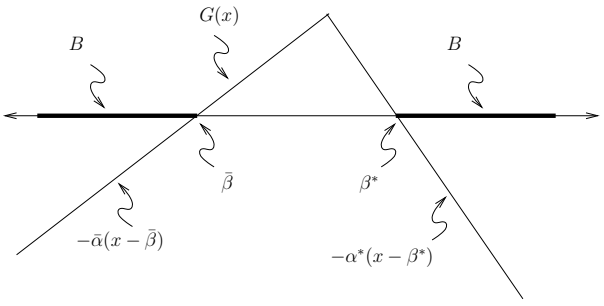

Fig. 14.3 Terminal condition corresponding to a subsolution

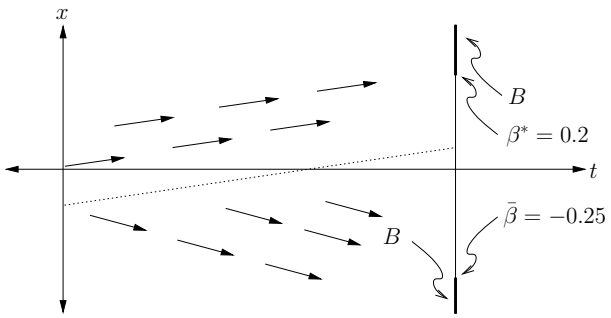

Fig. 14.4 Partition of the domain by a piecewise classical subsolution

The definitions for classical and piecewise classical subsolutions are similar to the finite time case. The relevant PDE is

$$\mathbb{H}(x, DV(x)) = 0, \tag{14.13}$$

with the boundary condition

$$V(x) = 0 \text{ for } x \in \partial B. \tag{14.14}$$

Definition 14.4 A function $\bar{V} : \mathbb{R}^d \to \mathbb{R}$ is a **classical sense subsolution** (or simply a classical subsolution) of (14.13)–(14.14) if it is continuously differentiable and if

$$\mathbb{H}(x, D\bar{V}(x)) \geq 0$$

for all $x \in (A \cup B)^c$, and if

$$\bar{V}(x) \leq 0 \text{ for } x \in B.$$

Definition 14.5 A function $\bar{V} : \mathbb{R}^d \to \mathbb{R}$ is a **piecewise classical sense subsolution** (or simply a piecewise classical subsolution) if the following hold. For some $J \in \mathbb{N}$ there are functions $\bar{V}^{(j)} : \mathbb{R}^d \to \mathbb{R}$, $j = 1, \ldots, J$, that are continuously differentiable and satisfy

$$\mathbb{H}(x, D\bar{V}^{(j)}(x)) \geq 0$$

for all $x \in (A \cup B)^c$. Moreover $\bar{V}(x) \doteq \wedge_{j=1}^{J} \bar{V}^{(j)}(x)$ satisfies

$$\bar{V}(x) \leq 0 \text{ for } x \in B.$$

Remark 14.6 (boundary condition for ∂A) In general one should specify a boundary condition for A as well. Since we have taken A to be open, the appropriate

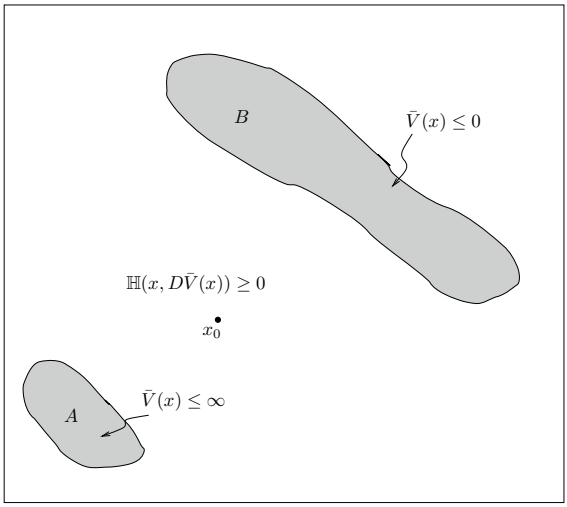

Fig. 14.5 Subsolution for the exit problem

boundary condition is the one which corresponds to a "state space constraint" [234]. For example, if ∂A were smooth with $x \in \partial A$ and n an outward normal to A at x, then the classical formulation of the state space constraint is

$$\inf_{\beta:\langle\beta,n\rangle\ge 0} [\langle DV(x), \beta\rangle + L(x,\beta)] = 0,$$

which reflects the fact that any candidate trajectory in the definition of $V(x)$ cannot enter A. Since our approach to rare event simulation is based on the construction of suitable classical and piecewise classical subsolutions, this boundary condition is vacuous. Indeed, we will assume that $\bar{V}$ is a subsolution in the sense of either Definition 14.4 or 14.5, and as a consequence the boundary condition for a subsolution will hold automatically. For example, in the context of Definition 14.4

$$\begin{aligned}\inf_{\beta:\langle\beta,n\rangle\ge 0} \left[\left\langle D\bar{V}(x), \beta\right\rangle + L(x,\beta)\right] &\ge \inf_{\beta\in\mathbb{R}^d} \left[\left\langle D\bar{V}(x), \beta\right\rangle + L(x,\beta)\right] \\ &= -H(x, -D\bar{V}(x)) \\ &= \mathbb{H}(x, D\bar{V}(x)) \\ &\ge 0.\end{aligned}$$

For a piecewise classical subsolution the concavity of $\mathbb{H}(x, p)$ gives the analogous bound. If instead we had assumed that A is closed with the attractor in the interior of A, the appropriate boundary condition ($V(x) = \infty$ for $x \in \partial A$) would again be vacuous when used to characterize a subsolution ($\bar{V}(x) \le \infty$ for $x \in \partial A$). The fact that these boundary conditions are vacuous can also be seen from the proofs of asymptotic optimality, where they play no role.

Of course there are many other types of events and (risk-sensitive) expected values that one could consider, and the interested reader can find the appropriate definitions of subsolutions for many of these in the references and Sect. 14.5. However, the two examples of this section will suffice to illustrate the main points.

14.4 The IS Scheme Associated to a Subsolution

We next discuss importance sampling schemes associated with a particular subsolution. Consider first the finite time problem. As discussed at the end of Sect. 14.2.3, if a smooth solution $V(x,t)$ to the HJB equation were available, then the correct change of measure if the current state of the simulated trajectory is at Y_i^n would be to replace the original distribution on the noise $v_i(Y_i^n)$, i.e., $\theta(dv|Y_i^n)$, by

$$\eta_\alpha(dv|Y_i^n) = e^{\langle\alpha,v\rangle - H(Y_i^n,\alpha)}\theta(dv|Y_i^n) \text{ with } \alpha = -DV(Y_i^n, i/n).$$

If one is using a classical subsolution $\bar{V}$ to design a scheme we follow exactly the same recipe, and the resulting second moment, rewritten in terms of the original random variables and process model, will equal

$$\mathfrak{S}^n(\bar{V}) \doteq E_{x_0}\left[1_{\{X_{Tn}^n \in B\}} \prod_{i=0}^{Tn-1} e^{\left\langle D\bar{V}(X_i^n, i/n), v_i(X_i^n)\right\rangle + H(X_i^n, -D\bar{V}(X_i^n, i/n))}\right]. \tag{14.15}$$

Rigorous asymptotic bounds on $\mathfrak{S}^n(\bar{V})$ will be derived in Sect. 15.2. It is shown in Theorem 15.1 that the decay rate of the second moment is bounded below by $V(x_0, 0)$ (the large deviation decay rate for the starting point x_0) plus $\bar{V}(x_0, 0)$. If $\bar{V}(x_0, 0) = V(x_0, 0)$ (the maximum possible value) then we have asymptotic optimality.

If dealing with a piecewise classical sense subsolution, the situation is different. In such a case the gradient $D\bar{V}$ is not smooth, and the analysis used to prove asymptotic performance bounds on $\mathfrak{S}^n(\bar{V})$ for the smooth case does not apply. In this case we mollify $\bar{V}$ and consider two associated importance sampling schemes. To be precise, for a small parameter $\delta > 0$ the standard mollification

$$\bar{V}^\delta(x,t) \doteq -\delta \log\left(e^{-\frac{1}{\delta}\bar{V}^{(1)}(x,t)} + \cdots + e^{-\frac{1}{\delta}\bar{V}^{(J)}(x,t)}\right) \tag{14.16}$$

is used. The properties of this mollification are summarized in the following lemma. The straightforward proof is omitted.

Lemma 14.7 *Let $\bar{V}^\delta$ be as in (14.16) where each function $\bar{V}^{(j)}$, $j = 1, \ldots, J$ is continuously differentiable on $\mathbb{R}^d \times [0, T]$. Define the weights*

$$\rho_j^\delta(x,t) \doteq \frac{e^{-\frac{1}{\delta}\bar{V}^{(j)}(x,t)}}{e^{-\frac{1}{\delta}\bar{V}^{(1)}(x,t)} + \cdots + e^{-\frac{1}{\delta}\bar{V}^{(J)}(x,t)}}.$$

Then

$$D\bar{V}^{\delta}(x,t) = \sum_{j=1}^{J} \rho_j^{\delta}(x,t) D\bar{V}^{(j)}(x,t) \text{ and } \bar{V}_t^{\delta}(x,t) = \sum_{j=1}^{J} \rho_j^{\delta}(x,t) \bar{V}_t^{(j)}(x,t). \tag{14.17}$$

Moreover

$$e^{-\frac{1}{\delta}\bar{V}(x,t)} \leq e^{-\frac{1}{\delta}\bar{V}^{\delta}(x,t)} \leq J e^{-\frac{1}{\delta}\bar{V}(x,t)},$$

and therefore

$$\bar{V}(x,t) \geq \bar{V}^{\delta}(x,t) \geq \bar{V}(x,t) - \delta \log J. \tag{14.18}$$

Recall that given an initial condition x_0, we seek a subsolution for which the value at $(x,t) = (x_0, 0)$ is as large as possible. From the convexity of H and the properties (14.17) it is easily checked that $\bar{V}^{\delta}$ is a classical subsolution in the sense of Definition 14.1 whenever $\bar{V}$ is a piecewise subsolution in the sense of Definition 14.2. The inequality (14.18) together with Theorem 15.1 in Chap. 15 then says that the mollification may lead to a loss of performance (a lowering of the decay rate of the second moment) that is at most $\delta \log J$ (see Theorem 15.1). Thus the role of the mollification is to define a mixture whose performance is very close to that of a classical subsolution, without giving up the flexibility and convenience of piecewise subsolutions. There are (at least) two schemes generated by a subsolution of the form (14.16), which we call the *ordinary implementation* and the *randomized implementation*.

Ordinary Implementation. Using the fact that $\bar{V}^{\delta}$ is a classical subsolution whenever $\bar{V}$ is a piecewise subsolution, we follow the standard procedure for classical subsolutions. Given that the state of the current simulated trajectory is Y_i^n, we use the sampling distribution $\eta_{\alpha}(dv|Y_i^n) = e^{\langle \alpha, v\rangle - H(Y_i^n, \alpha)}\theta(dv|Y_i^n)$ with tilt parameter $\alpha = -D\bar{V}^{\delta}(Y_i^n, i/n)$ to generate a random variable w_i^n with the given (conditional) distribution. The state of the system is then updated according to $Y_{i+1}^n = Y_i^n + w_i^n/n$, and we repeat. The likelihood ratio is

$$R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,Tn-1}) = \prod_{i=0}^{Tn-1} e^{\langle D\bar{V}^{\delta}(Y_i^n, \frac{i}{n}), w_i^n\rangle + H(Y_i^n, -D\bar{V}^{\delta}(Y_i^n, \frac{i}{n}))}, \tag{14.19}$$

and the resulting estimator is $1_{\{Y^n(T)\in B\}} R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,Tn-1})$, where $Y^n(t)$ is the continuous time interpolation.

Randomized Implementation. In this case, the estimator is constructed as follows. Given that the state of the current simulated trajectory is Y_i^n, we generate an independent random variable $\kappa_i^n \in \{1, \ldots, J\}$ with probabilities $\rho_j^{\delta}(Y_i^n, i/n)$, and if $\kappa_i^n = j$ then use the sampling distribution with the tilt parameter $\alpha = -D\bar{V}^{(j)}(Y_i^n, i/n)$ to generate w_i^n. In this case the likelihood ratio is

$$R^n(\{Y_i^n, w_i^n\}_{i=0,\dots,Tn-1}) \tag{14.20}$$
$$= \prod_{i=0}^{Tn-1} \left(\sum_{j=1}^{J} \rho_j^\delta \left(Y_i^n, \frac{i}{n} \right) e^{-\langle D\bar{V}^{(j)}(Y_i^n, \frac{i}{n}), w_i^n \rangle - H(Y_i^n, -D\bar{V}^{(j)}(Y_i^n, \frac{i}{n}))} \right)^{-1},$$

and the estimator takes the same form as in the ordinary case.

For both implementations the resulting second moment, rewritten in terms of the original process and noises, is

$$\mathfrak{S}^n(\bar{V}^\delta) \doteq E_{x_0}\left[1_{\{X^n(T)\in B\}} R^n(\{X_i^n, v_i(X_i^n)\}_{i=0,\dots,Tn-1}) \right]. \tag{14.21}$$

where R^n is given by (14.19) or (14.20) depending on which implementation is used. Since $\bar{V}^\delta$ is a classical subsolution, the randomized case includes the ordinary case with $J = 1$ and taking $\bar{V}^{(1)} = \bar{V}^\delta$. It is shown in Theorem 15.1 that the decay rate of the second moment for both implementations is bounded below by $V(x_0, 0)$ (the large deviation decay rate for the starting point x_0) plus $\bar{V}^\delta(x_0, 0)$.

Example 14.8 (Example 14.3 continued) In Example 14.3 a piecewise subsolution was constructed for the problem of Sect. 14.2.2 with a nonconvex set B. We apply this subsolution for the same data ($\beta^* = 0.2$ and $\bar{\beta} = -0.25$) as in Sect. 14.2.2. As before, each trial is based on $K = 5,000$ simulated trajectories. We give the number of "rogue" trajectories (those ending in $(-\infty, -0.25]$) even though that name is no longer appropriate. Recall that the true value for $n = 60$ is $p^n = 8.70 \times 10^{-2}$. Table 14.2 presents data using the ordinary implementation. The estimates are much more stable across the different trials, with confidence intervals that are both small and which contain the true value. Table 14.3 gives the analogous data for the randomized implementation, which is qualitatively very similar to that of the ordinary case. Table 14.4 considers the same model and escape set for the randomized implementation, but for various values of n. The analogous results for the ordinary implementation are omitted since they are similar. Each trial used $K = 20,000$ simulated trajectories. As with $n = 60$, the results are stable and accurate. Note that the ratio $(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$ is increasing in n (though since $\delta > 0$ is fixed it will never reach 2), and that the number of "rogue" trajectories is decreasing in n, reflecting

Table 14.2 Ordinary implementation of mollified subsolution with $\delta = 0.02$

	No. 1	No. 2	No. 3	No. 4
Estimate $\hat{p}^n$ ($\times 10^{-2}$)	8.55	8.73	8.72	8.61
Standard error ($\times 10^{-2}$)	0.183	0.184	0.182	0.182
95% confidence interval ($\times 10^{-2}$)	[8.19, 8.91]	[8.37, 9.10]	[8.36, 9.08]	[8.25, 8.97]
Number of "rogue" trajectories	751	727	833	807
$(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$	1.51	1.52	1.53	1.52

Table 14.3 Randomized implementation of mollified subsolution with $\delta = 0.02$

	No. 1	No. 2	No. 3	No. 4
Estimate $\hat{p}^n$ ($\times 10^{-2}$)	9.02	8.76	8.62	8.91
Standard error ($\times 10^{-2}$)	0.183	0.182	0.181	0.183
95% confidence interval ($\times 10^{-2}$)	[8.66, 9.38]	[8.40, 9.11]	[8.26, 8.97]	[8.55, 9.26]
Number of "rogue" trajectories	802	782	823	883
$(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$	1.54	1.53	1.52	1.53

Table 14.4 Randomized implementation of mollified subsolution with $\delta = 0.02$

	$n = 100$	$n = 200$	$n = 500$
Exact value p^n	2.90×10^{-2}	2.54×10^{-3}	3.88×10^{-6}
Estimate $\hat{p}^n$	2.93×10^{-2}	2.59×10^{-3}	3.81×10^{-6}
Standard error	3.76×10^{-4}	4.82×10^{-5}	1.35×10^{-7}
95% confidence interval	$[2.86, 3.00] \times 10^{-2}$	$[2.49, 2.68] \times 10^{-3}$	$[3.55, 4.08] \times 10^{-6}$
Number of "rogue" trajectories	2176	935	107
$(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$	1.59	1.65	1.74

the fact that the probability associated with $(-\infty, -0.25]$ conditioned on ending in B is decreasing in n.

Remark 14.9 (role of smoothness) The theoretical bounds on performance derived in Chap. 15 make use of the fact that $\bar{V}^\delta$ smooth, and in particular that it is a classical sense subsolution (and not just a viscosity sense subsolution [14, 134]). A natural question is whether this smoothness is necessary. From the perspective of implementation it is certainly convenient, since the change of measure for the increments is based on the gradient of the subsolution. However, one could ask if there is some generalized implementation (e.g., based on sub or superdifferentials) that might allow for less regular subsolutions. Such a construction would require that in the analysis of the second moment we consider the large deviation theory for processes with "discontinuous statistics." The theory for such processes is not well understood in great generality, and in particular there is no rigorous analysis of importance sampling for nonsmooth subsolutions. Given the subtlety in applying importance sampling to rare event estimation, it seems prudent to use the mollification presented previously, which is very easy to implement and for which a rigorous analysis is available. This difference in the properties of subsolutions is one of the key qualitative distinctions between importance sampling and the analogous splitting algorithms to be considered in Chap. 16, for which a weak sense subsolution is known to be sufficient.

Remark 14.10 (achieving asymptotic optimality) Since the mollification can reduce the value of the subsolution at the starting point [i.e., $\bar{V}^\delta(x_0, 0) < V(x_0, 0)$ is possible even when $\bar{V}(x_0, 0) = V(x_0, 0)$], this would seem to be a significant drawback for

importance sampling. However, while there may be other issues to consider when comparing importance sampling and splitting, it is easy to remedy this objection, and in general one can allow $\delta \to 0$ as $n \to \infty$ so as to achieve asymptotic optimality. This issue is discussed in Remark 15.7 and Theorem 15.14.

Remark 14.11 (randomized versus ordinary) When dealing with noise models such as those of (14.1) one may prefer the ordinary implementation over the randomized implementation, since the appropriate change of measure is simply defined by an exponential tilt, and there is no need to generate random variables according to the weights $\rho_j^\delta(\cdot, i/n)$. Note that for these models the distribution of the noise, conditioned on the state X_i^n, is independent in the time variable. For more complex models (e.g., the Markov modulated models discussed in Sect. 7.3) there may be an advantage to using the randomized implementation, since the change of measure is more complex, and requires, for each distinct value of the gradient, the solution of an eigenvalue problem. In particular, if the component functions $\bar{V}^{(j)}$ all have a constant gradient then one must solve at most J eigenvalue problems for the randomized implementation, while the ordinary implementation will typically require that such a problem be solved for each time $i = 0, \ldots, Tn - 1$ of the simulation. An example of this sort appears in Sect. 14.5.5.

Remark 14.12 The implementation of the importance sampling scheme and resulting form of the second moment are entirely analogous for the problem of hitting a rare set before a typical set, save that the scheme has no explicit dependence on time, and Tn is replaced by the first exit time N^n.

14.5 Generalizations

In this section we briefly comment on generalizations with respect to various aspects of the model, including expected values besides probabilities, continuous time models, and more complex noise models. Some generalizations that are very straightforward (e.g., when the local rate function also depends on time) are not discussed.

14.5.1 Functionals Besides Probabilities

Straightforward and natural generalizations in the context of both the finite time problem and the problem of hitting a rare set prior to a typical set involve the computation of risk-sensitive functionals. For example, in the setting of the finite time problem, we may want to compute a quantity such as

$$V^n(x, 0) = -\frac{1}{n} \log E_x \exp\left\{-nF(X_{Tn}^n)\right\},$$

where F is a suitably regular (e.g., continuous) function, and where for convenience in the notation we assume Tn is an integer. Under appropriate conditions $V^n(x,0) \to V(x,0)$, where

$$V(x,t) \doteq \inf\left[\int_t^T L(\phi(s),\dot{\phi}(s))ds + F(\phi(T)) : \phi(t) = x\right], \tag{14.22}$$

and the only difference in the definition of the various forms of subsolution occur in the terminal condition. Thus in Definition 14.1, the condition $\bar{V}(x,T) \le 0$ for $x \in B$ is replaced by $\bar{V}(x,T) \le F(x)$ for $x \in \mathbb{R}^d$.

A single sample of the estimator, with $R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,Tn-1})$ defined by either (14.19) or (14.20) depending on which implementation is used, is

$$F(Y_{Tn}^n)R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,Tn-1}).$$

14.5.2 Continuous Time

When considering continuous time process models a basic issue is numerical implementation. For example, trajectories of the solution to an SDE are usually approximated, e.g., by the Euler-Maruyama method. Since this returns the problem to the discrete time setting, it can be dealt with using the same notions of importance sampling and subsolutions as those already given. (Note that there is still the problem of quantifying the impact of the time discretization, but that is a topic we do not consider here.) In contrast, for continuous time models that are of pure jump form there is no need to discretize time, and one can formulate both the importance sampling and related analysis directly in continuous time.

To keep the presentation brief we will consider just one class of models, but the ideas can easily be generalized. Thus suppose that X^n is a continuous time Markov process of the following form. There is $J \in \mathbb{N}$ and bounded and Lipschitz continuous functions

$$v_j : \mathbb{R}^d \to \mathbb{R}^d, \quad r_j : \mathbb{R}^d \to (0,\infty), \quad j = 1,\ldots,J,$$

where $nr_j(x)$ is the jump intensity of a jump to the point $x + v_j(x)/n$, given that $X^n(t) = x$. Thus X^n has the infinitesimal generator

$$(\mathscr{L}^n f)(x) \doteq \sum_{j=1}^J nr_j(x)\left[f(x + v_j(x)/n) - f(x)\right]$$

for bounded functions $f : \mathbb{R}^d \to \mathbb{R}$. Hence the process waits at the location x for an exponentially distributed time τ with inverse mean $\sum_{j=1}^J nr_j(x)$. After τ units of time, it jumps to the location $x + v_j(x)/n$ with probability proportional to $r_j(x)$

for $j = 1, \ldots, J$. The weighed serve-the-longest queue model of Chap. 13 is of this sort, except that $r_j(x)$ can be equal to zero for some x values.

If we consider the problem of hitting a rare set before a typical one, the continuous time aspect is unimportant, and this problem can be reduced to the discrete time problems considered previously by working with the imbedded discrete time process. This is the approach taken in [105, 110, 117]. In the notation of this chapter, the discrete time model corresponds to

$$\theta(A|x) = \sum_{j=1}^{J} r_j(x)\delta_{v_j(x)}(A) \Big/ \sum_{j=1}^{J} r_j(x).$$

This simplification is not possible with the finite time problem, since a rare outcome depends on the holding time and not just on which jump type is selected at the time of a transition. In this case, we need to stay in the continuous time framework.

The processes X^n take values in $\mathscr{D}([0, T] : \mathbb{R}^d)$, and the local rate function for the sequence $\{X^n\}_{n\in\mathbb{N}}$ is given by

$$L(x, \beta) \doteq \inf \left[\sum_{j=1}^{J} r_j(x)\ell(\bar{r}_j/r_j(x)) : \sum_{j=1}^{J} \bar{r}_j v_j(x) = \beta, \bar{r}_j \in [0, \infty), j = 1, \ldots, J \right],$$

where $x \in \mathbb{R}^d$, $\beta \in \mathbb{R}^d$ and as usual $\ell(z) \doteq z \log z - z + 1$ for $z \in [0, \infty)$. The analogue of the log moment generating function is given by

$$\begin{aligned} H(x, \alpha) &= \sup_{\beta\in\mathbb{R}^d} [\langle \alpha, \beta\rangle - L(x, \beta)] \\ &= \sup_{\bar{r}_j\in[0,\infty), j=1,\ldots,J} \left[\sum_{j=1}^{J} \bar{r}_j \left\langle v_j(x), \alpha \right\rangle - \sum_{j=1}^{J} r_j(x)\ell(\bar{r}_j/r_j(x)) \right] \\ &= \sum_{j=1}^{J} r_j(x) \left[e^{\langle v_j(x),\alpha\rangle} - 1 \right], \end{aligned}$$

with the supremum achieved at $\bar{r}_j = r_j(x)e^{\langle v_j(x),\alpha\rangle}$. As before the PDE that is relevant takes the form (14.11), where $\mathbb{H}(x, p) = -H(x, -p)$, and the terminal condition is as in the discrete time setting. Given a classical subsolution $\bar{V}(x, t)$, the simulated process under the ordinary implementation uses the rates $\bar{r}_j(x, t) = r_j(x)e^{\langle v_j(x),-D\bar{V}(x,t)\rangle}$. The estimate is then $1_{\{Y^n(T)\in B\}}R^n(Y^n)$, where

$$\log R^n(Y^n) = \int_0^T \sum_{j=1}^{J} r_j(Y^n(t)) \left[e^{\langle v_j(Y^n(t)),-D\bar{V}(Y^n(t),t)\rangle} - 1 \right] dt$$

$$- \sum_{i:t_i^n \le T} \left[\langle v_{j_i^n}(Y^n(t_i^n-)), D\bar{V}(Y^n(t_i^n-), t_i^n-) \rangle \right],$$

with t_i^n the jump times of Y^n and with j_i^n identifying the type of jump.

Remark 14.13 For the problem of hitting a rare set before a typical set one could also use the PDE (14.13) and boundary condition (14.14), with

$$\mathbb{H}(x, p) = -\sum_{j=1}^{J} r_j(x)[e^{-\langle v_j(x), p \rangle} - 1].$$

This form of $\mathbb{H}$ differs from the discrete time analogue, but characterizes the same set of subsolutions if used for an exit type problem.

14.5.3 Level Crossing

Problems such as level crossings, which appear in ruin problems from insurance, are of the same general sort as that of hitting a rare set before hitting a typical set. The main distinction is that the "typical set" is not part to the state space, and instead corresponds to the process drifting infinitely far in some direction. For example, consider once again the discrete time setting, suppose that $\theta(dy|x) = \theta(dy)$, $H(\alpha) < \infty$ for all $\alpha \in \mathbb{R}^d$ and also that

$$\int_{\mathbb{R}^d} y_k \theta(dy) < 0 \text{ for } k = 1, \ldots, d. \tag{14.23}$$

Then each component of $X^n(t)$ tends to $-\infty$ as $t \to \infty$. Let $M_k \in (0, \infty)$ for $i = 1, \ldots, d$ and consider the problem of estimating the level crossing probability

$$P\left\{ \sup_{m \in \mathbb{N}} \max_{k=1,\ldots,d} \frac{1}{M_k} \sum_{i=0}^{m-1} (v_i)_k \ge n \right\},$$

where v_i are iid with distribution θ and $(v_i)_k$ denotes the kth component of v_i. This quantity is the same as

$$P_0\left\{ \sup_{t \in [0,\infty)} \max_{k=1,\ldots,d} \frac{[X^n(t)]_k}{M_k} \ge 1 \right\},$$

and can be thought of has hitting the rare set G^c, where $G \doteq \times_{k=1}^d (-\infty, M_k)$, before wandering off to $-\infty$ in each component (the "typical" set). With this analogy in place, the definitions of subsolution and their use are exactly as before. In particular, if $H(\alpha)$ is the log moment generating function of θ and $\mathbb{H}(p) = -H(-p)$, then a smooth function $\bar{V} : \mathbb{R}^d \to \mathbb{R}$ is a classical subsolution if

$$\mathbb{H}(D\bar{V}(x)) \geq 0 \text{ for } x \in G \text{ and } \bar{V}(x) \leq 0 \text{ for } x \in G^c. \tag{14.24}$$

The (now state dependent) alternative sampling distribution for the next increment w_i^n given $Y_i^n = x$ is $e^{-\langle D\bar{V}(x),v\rangle - H(-D\bar{V}(x))}\theta(dv)$, and the estimate is

$$1_{\left\{Y^n_{\bar{N}^n} \in G^c\right\}} \prod_{i=0}^{\bar{N}^n - 1} e^{\langle D\bar{V}(Y_i^n), w_i^n\rangle + H\left(-D\bar{V}(Y_i^n)\right)},$$

where $\bar{N}^n \doteq \inf\left\{i : Y_i^n \in G^c\right\}$. For such problems it is natural to consider piecewise classical subsolutions with one component $\bar{V}^{(k)}$ for each index $k = 1, \ldots, d$. $\bar{V}^{(k)}(x)$ should be of the form $= -\langle \alpha^{(k)}, x\rangle + c^{(k)}$, where $\alpha^{(k)}$ is of the form $a^{(k)}e_k$, $H(\alpha^{(k)}) = 0$, and $c^{(k)} = a^{(k)}M_k$. One can check that under (14.23), for each $k = 1, \ldots, d$ there is exactly one positive number $a^{(k)}$ such that $H(a^{(k)}e_k) = 0$, that with these choices $\bar{V}(x) \doteq \min_{k=1,\ldots,d} \bar{V}^{(k)}(x)$ is a piecewise classical subsolution with $V(0) = \bar{V}(0) = \min_{k=1,\ldots,d} a^{(k)}M_k$. For any problem where the simulation time is potentially unbounded it is important to know that a proposed scheme is practical. In the present setting, for the process that is simulated the increments have conditional distribution $e^{-\langle D\bar{V}^\delta(x),v\rangle - H(-D\bar{V}^\delta(x))}\theta(dv)$. The mean of this distribution points towards the target set, and it follows that $\bar{N}^n < \infty$ and $Y^n_{\bar{N}^n} \in G^c$ with probability one. One can in fact show more, for example that $E\bar{N}^n < \infty$.

14.5.4 Path Dependent Events

In some situations one may be interested in probabilities and related quantities in which the occurrence or not of the rare event is determined by the path of X^n over an interval $[0, T]$. To simplify notation we will consider a homogeneous random walk as in the last section [i.e., $\theta(dy|x) = \theta(dy)$], $T = 1$, and the case of one dimension. Then an example of this type of problem is to compute

$$E_0\left[e^{-nF(X^n(1))} 1_{\{\max_{t\in[0,1]} X^n(t) \geq h\}}\right], \tag{14.25}$$

where $h \in (0, \infty)$ and F is bounded and continuous. Let $l < h$ and define $\tau_h^n \doteq \inf\{t \geq 0 : X^n(t) \geq h\}$ and $\tau_l^n \doteq \inf\{t \geq \tau_h^n : X^n(t) \leq l\}$. A second example is computing $P_0\left\{\tau_l^n \leq 1\right\}$. Of course this is only particularly difficult if the indicated events are rare, and to make this so we assume $\int_{\mathbb{R}} y\theta(dy) < h$.

It is easy to write down the variational problem for the large deviation approximations to these quantities. For example, for the expected value in (14.25) the corresponding variational problem is

$$\inf\left[\int_0^1 L(\dot{\phi}(t))dt + F(\phi(1)) : \phi(0) = 0, \phi(s) \geq h \text{ for some } s \in [0, 1]\right],$$

where the infimum is over absolutely continuous ϕ. To identify the PDE that is related to this problem we introduce a state variable that will indicate whether or not h has been crossed. Denote the simulated process by $Y^n(t)$ and consider the associated indicator process $Z^n(t) \doteq 1_{[h,\infty)}(\max_{s\in[0,t]} Y^n(s))$. Suppose we are given that $(Z^n(t), Y^n(t)) = (1, x)$. Then the event $\{\max_{t\in[0,1]} Y^n(t) \geq h\}$ is certain, and importance sampling schemes for time instants after t can be generated by subsolutions of the PDE

$$\bar{V}_t(1, x, t) + \mathbb{H}(D\bar{V}(1, x, t)) \geq 0, \quad x \in \mathbb{R}, t \in (0, 1), \tag{14.26}$$

with terminal condition

$$\bar{V}(1, x, 1) \leq F(x), \quad x \in \mathbb{R} \tag{14.27}$$

(here we use variables (z, x, t), and $\bar{V}(1, x, t)$ indicates that $z = 1$). If on the other hand we are given $(Z^n(t), Y^n(t)) = (0, x)$, then for the cost to be finite the event $Y^n(s) \geq h$ must occur for some $s \in [t, 1]$, and by the usual logic of dynamic programming the asymptotic optimal future costs after that time will be bounded below by any subsolution $\bar{V}(1, \cdot, \cdot)$ to (14.26). The characterization of a subsolution for times prior to this event is given by

$$\bar{V}_t(0, x, t) + \mathbb{H}(D\bar{V}(0, x, t)) \geq 0, \quad x \in (-\infty, h), t \in (0, 1), \tag{14.28}$$

and

$$\bar{V}(0, x, t) \leq \bar{V}(1, x, t), \quad x \in [h, \infty), t \in (0, 1). \tag{14.29}$$

Note that one must construct the subsolutions in the order first $\bar{V}(1, x, t)$, then $\bar{V}(0, x, t)$. Given classical subsolutions $\bar{V}(0, x, t)$ and $\bar{V}(1, x, t)$, the simulated trajectory $\{Y^n(t)\}$ is defined as follows. Given that the state of the current simulated trajectory is Y_i^n, we use the sampling distribution $\eta_\alpha(dv|Y_i^n) = e^{\langle\alpha,v\rangle - H(\alpha)}\theta(dv)$ with tilt parameter $\alpha = -D\bar{V}(0, Y_i^n, i/n)$ to generate a random variable w_i^n with the given (conditional) distribution if $i < N^n$, where $N^n \doteq \inf\{i : Y_i^n \geq h\} \wedge n$. If $i \geq N^n$ we instead use the tilt parameter $\alpha = -D\bar{V}(1, Y_i^n, i/n)$ to generate a random variable w_i^n. The state of the system is then updated according to $Y_{i+1}^n = Y_i^n + w_i^n/n$, and we repeat. The likelihood ratio is

$$\begin{aligned} R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,n-1}) = {} & \textstyle\prod_{i=0}^{N^n-1} e^{\langle D\bar{V}(0,Y_i^n,\frac{i}{n}),w_i^n\rangle + H(-D\bar{V}(0,Y_i^n,\frac{i}{n}))} \\ & \times \textstyle\prod_{i=N^n}^{n-1} e^{\langle D\bar{V}(1,Y_i^n,\frac{i}{n}),w_i^n\rangle + H(-D\bar{V}(1,Y_i^n,\frac{i}{n}))}, \end{aligned}$$

and the resulting estimator is

$$e^{-nF(Y^n(1))} 1_{\{\max_{t\in[0,1]} Y^n(t)\geq h\}} R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,n-1}),$$

where $Y^n(t)$ is the continuous time interpolation.

The corresponding set of PDEs for $P_0\left\{\tau_l^n \le 1\right\}$ is similar, save (14.26) and (14.27) are replaced by

$$\bar{V}_t(1,x,t) + \mathbb{H}(D\bar{V}(x,1,t)) \ge 0, \quad x \in (l,\infty), t \in (0,1),$$

and

$$\bar{V}(1,x,t) \le 0, \quad x \in (-\infty, l], t \in [0,1].$$

This construction can be generalized in many directions. For example, with a level crossing problem as in the last section one could consider the event that a particular level is crossed (i.e., one component exceeds its threshold) prior to a level crossing in some other coordinate direction.

14.5.5 Markov Modulated Models

As a final example we consider problems where there are two times scales, as was the case with the models of Sect. 7.3. To keep the discussion simple we consider the case

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(\Xi_{i+1}), \quad X_0^n = x_0, \Xi_1 = \xi,$$

with X_i^n taking values in $\mathbb{R}^d$ and the probability of interest a level crossing as in Sect. 14.5.3. However, the constructions generalize greatly, and other examples can be found in [116]. Recall from Sect. 7.3 that $\{\Xi_i\}_{i\in\mathbb{N}}$ is an S-valued Markov chain with transition probability kernel p and that $\{v_i(\xi)\}_{i\in\mathbb{N}_0}$ is a sequence of iid random vector fields with distribution given by $\theta(\cdot|\xi)$. We assume that the moment generating functions $Ee^{\langle\alpha, v_i(\xi)\rangle}$ are bounded from above uniformly in $\xi \in S$, and the other conditions of Sect. 7.3. The local rate function for this model is

$$L(\beta) \doteq$$
$$\inf\left[\int_S R(\nu(\cdot|\xi)\|\theta(\cdot|\xi))\mu(d\xi) + R\left(\gamma \| \mu \otimes p\right) : \int_{S\times\mathbb{R}^d} y\nu(dy|\xi)\mu(d\xi) = \beta\right],$$

where the infimum is over $\gamma \in \mathscr{P}(S\times S)$ such that $[\gamma]_1 = [\gamma]_2 = \mu$ and stochastic kernels $\nu(dw|\xi)$ on $\mathbb{R}^d$ given S.

Let $\mathbb{H}(p) = \inf_{\beta\in\mathbb{R}^d}[\langle p,\beta\rangle + L(\beta)]$. Then the correct notion of subsolution for this problem is again (14.24). There is an alternative characterization of $\mathbb{H}(p) = -H(-p)$ in terms of an eigenvector/eigenvalue problem. For $\alpha \in \mathbb{R}^d$ let $H(\alpha)$ and $r(\cdot;\alpha)$ solve

$$\int_S \int_{\mathbb{R}^d} e^{\langle\alpha,w\rangle}\theta(dw|\eta)r(\eta;\alpha)p(\xi,d\eta) = e^{H(\alpha)}r(\xi;\alpha), \quad \xi \in S,$$

where $r(\cdot;\alpha) : S \to [0,\infty)$ is the corresponding eigenfunction [116]. One can interpret $H(\alpha)$ in terms of a large time risk-sensitive (i.e., multiplicative) cost, in that

$$\frac{1}{k}\log E\left[e^{n\langle\alpha,X_k^n\rangle}\,\middle|\,X_0^n=0,\,\Xi_1=\xi\right]\to H(\alpha)\text{ as }k\to\infty,$$

and $r(\xi;\alpha)$ plays the role of the cost potential. One can in fact prove this limit using the weak convergence arguments of Chap. 6.

Given a subsolution $\bar{V}$, we generate processes $\{(Y_i^n,\Theta_{i+1}^n)\}$ by setting $Y_0^n=x_0$ and $\Theta_1^n=\xi$, using

$$e^{-\langle D\bar{V}(Y_i^n),w\rangle-H(-D\bar{V}(Y_i^n))}\theta(dw|\eta)\frac{r(\eta;-D\bar{V}(Y_i^n))}{r(\Theta_i^n;-D\bar{V}(Y_i^n))}p(\Theta_i^n,d\eta)$$

to identify the conditional distribution of w_i^n and Θ_{i+1}^n given Y_i^n and Θ_i^n, and then setting $Y_{i+1}^n=Y_i^n+w_i^n/n$. The estimator for the level crossing problem is then

$$1_{\left\{Y_{\bar{N}^n}^n\in(\times_{k=1}^d(-\infty,M_k))^c\right\}}\prod_{i=0}^{\bar{N}^n-1}e^{\langle D\bar{V}(Y_i^n),w_i^n\rangle+H(-D\bar{V}(Y_i^n))}\frac{r(\Theta_i^n;-D\bar{V}(Y_i^n))}{r(\Theta_{i+1}^n;-D\bar{V}(Y_i^n))},$$

where $\bar{N}^n\doteq\inf\left\{i:Y_i^n\in(\times_{k=1}^d(-\infty,M_k))^c\right\}$. As in the iid case the resulting algorithm is practical, in that $E\bar{N}^n<\infty$.

14.6 Notes

The references [6, 190, 224] present Monte Carlo methods in a general setting, and also discuss various aspects of rare event estimation. A nice overview of the use of Monte Carlo in the rare event setting specifically can be found in [223], which discusses other methods that are widely used, such as interacting particle methods (see also [51, 78]) and the cross entropy method for the design of importance sampling (see also [225]).

As noted previously the first paper to apply importance sampling in the rare event context is Siegmund [232]. The material of this chapter is mostly taken from [114, 116, 150], though the last section includes examples from other papers as well. The notion of Lyapunov inequality as used in [27] is closely related to that of subsolution in the context of importance sampling, and more information on this connection can be found in [28].

We consider only the light tailed cases (i.e., distributions for which moment generating functions are finite at least in a neighborhood of the origin). Problems with heavy tailed distributions are also important. A survey of developments up to 2012 on importance sampling for rare event estimation that includes the heavy tailed case is [26], and more recent developments for the heavy tailed case (including new classes of problem formulation not discussed previously) appear in [54].

For background on the Hamilton-Jacobi-Bellman equations used in this chapter we refer to [14, 134].

Chapter 15
Performance of an IS Scheme Based on a Subsolution

In Chap. 14 we considered the problem of rare event simulation associated with small noise discrete time Markov processes of the form analyzed in Chap. 4. Two types of events were emphasized: those that are described by process behavior on a bounded time interval (finite time problems) and those that concern properties of the process over unbounded time horizons (e.g., exit probability problems). For both families of problems, importance sampling schemes based on classical and piecewise classical-sense subsolutions of certain Hamilton–Jacobi–Bellman equations were proposed. In the current chapter we provide asymptotic performance bounds for such schemes. The main result for the finite-time problem is Theorem 15.1, whereas the performance bound for the exit probability problem is given in Theorem 15.10. The proofs of Theorems 15.1 and 15.10 appear in Sects. 15.2 and 15.3, respectively. One can weaken the conditions that are assumed, and also generalize the arguments in many directions. In particular, generalizations to cover the examples of Sect. 14.5 can all be carried out using arguments analogous to those presented in Sects. 15.2 and 15.3.

15.1 Statement of Resulting Performance

We recall that the performance of any Monte Carlo approximation scheme is characterized by the variance of a single sample, and that since the schemes we consider are unbiased, minimizing the variance is equivalent to minimizing the second moment. We also recall from Eq. (14.5) that under suitable regularity assumptions on the event of interest, the best possible rate of decay for this second moment is precisely twice the large deviation rate for the quantity being estimated.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_15

We will identify a lower bound on the decay rate for any scheme constructed in terms of a subsolution as described in Sect. 14.4. As we will see, the bound has a simple expression, and moreover, the proof of this result will follow from almost the same argument used to prove the large deviation upper bound. Under additional regularity one can also characterize the limit of the second moment; see Remark 15.16.

We first state the result for the finite-time problem and then the corresponding result for the exit problem. The process model will be the state-dependent random walk model of Chap. 4. However, for the reasons just given, the proof carries over to other process models once one has established the corresponding large deviation theory. As noted previously, generalizations to the various functionals described in Sect. 14.5 can be established using arguments of the same type.

We follow the notation from Chaps. 4 and 14 for the small noise Markov process X^n of this chapter. In particular, iid random vector fields $\{v_i(x), i \in \mathbb{N}_0, x \in \mathbb{R}^d\}$ and a stochastic kernel $\theta(dy|x)$ on $\mathbb{R}^d$ given $\mathbb{R}^d$ are as in Sect. 14.1, and $\{X_i^n\}_{i\in\mathbb{N}_0}$, $\{X^n(t)\}_{t\geq 0}$ are defined through (14.1) and (14.2), respectively. Recall also that for $x_0 \in \mathbb{R}^d$, P_{x_0} denotes the probability measure under which $X_0^n = X^n(0) = x_0$.

We first consider the finite-time problem. Recall that the finite-time importance sampling problem of interest is to estimate $P_{x_0}\{X^n(T) \in B\}$, where B is a closed set in $\mathbb{R}^d$ and $T \in (0, \infty)$. Also recall that the PDE that characterizes the large deviation rate, or equivalently, half the optimal rate of decay for an asymptotically optimal importance sampling scheme, is given by (14.11) with the terminal condition as in (14.12). The importance sampling schemes proposed in Sect. 14.4 use classical and piecewise classical subsolutions of such equations as defined in Definitions 14.1 and 14.2 respectively. For $T \in [0, \infty)$, recall the rate function I_T associated with the LDP for X^n on $\mathscr{C}([0, T] : \mathbb{R}^d)$. It is given by

$$I_T(\phi) \doteq \int_0^T L(\phi(t), \dot{\phi}(t))dt, \ \ \phi \in \mathscr{AC}([0, T] : \mathbb{R}^d),$$

where L is as in (14.6), and $I_T(\phi) = \infty$ for all other ϕ. Let

$$V(x_0, 0) \doteq \inf [I_T(\phi) : \phi(0) = x_0, \phi(T) \in B].$$

The following theorem gives the performance bound for this problem. Recall that the second moment of the single-sample estimate for a scheme based on such a subsolution $\bar{V}$ is denoted by $\mathfrak{S}^n(\bar{V})$, and is defined in (14.15) in the classical case and (14.21) in the piecewise classical case.

Theorem 15.1 *Suppose that Condition 4.3 holds. Let $\bar{V}$ be a classical subsolution for the PDE (14.11) with terminal condition (14.12) such that*

$$\sup_{x\in\mathbb{R}^d,\, t\in[0,T]} \left[\|D\bar{V}(x,t)\| + |\bar{V}_t(x,t)|\right] < \infty,$$

and define the IS scheme as in Sect. 14.4. Then the second moment for this scheme satisfies

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathfrak{S}^n(\bar{V}) \geq V(x_0,0) + \bar{V}(x_0,0). \tag{15.1}$$

Suppose that $\bar{V} = \wedge_{j=1}^{J} \bar{V}^{(j)}$ *is a piecewise classical subsolution such that for each* $j = 1, \ldots, J$,

$$\sup_{x\in\mathbb{R}^d,\, t\in[0,T]} \left[\|D\bar{V}^{(j)}(x,t)\| + |\bar{V}_t^{(j)}(x,t)| \right] < \infty,$$

and let $\bar{V}^{\delta}$ *denote the mollification defined by (14.16). If one is using either the ordinary or randomized implementation as described in Sect. 14.4, then (15.1) holds with* $\bar{V}$ *replaced by* $\bar{V}^{\delta}$.

Remark 15.2 According to (15.1), the performance of the scheme based on any subsolution is measured by the value of the subsolution at the starting point, with larger values giving better performance. When there is a comparison principle for the PDE, a subsolution can never be greater than the solution, and the best possible value is $\bar{V}(x_0,0) = V(x_0,0)$. In this case, the decay rate of the second moment is $2\bar{V}(x_0,0)$, the best possible, and thus the scheme is asymptotically efficient in the sense of Sect. 14.1. There are proofs under many different sets of conditions of comparison principles for solutions to viscosity solutions. In this finite-time context, examples include [14, Theorem 3.7] and [134, Theorem II.9.1].

Remark 15.3 The subsolution property relaxes various equalities to inequalities, but only in directions such that a verification argument can rigorously bound the second moment of the corresponding estimator. However, one can ask whether the subsolution property is needed, and in particular whether a scheme based on a function that is not a subsolution can be expected to perform in some predictable way, with the decay rate of the second moment characterized by the value of this function at the starting location $(x_0, 0)$. Example 14.3 shows that in general, this is not the case. One can construct a function $\bar{V}$ that has the same value as the solution at $(x_0, 0)$, and which will generate the importance sampling scheme used in the example (it is, in fact, $\bar{V}^{(1)}$). Thus the relation between the asymptotics of the second moment and the value at the starting point of the function that generates the scheme is in general valid only for subsolutions.

Remark 15.4 (*Bounding the state space*) The assumed bounds on derivatives of $\bar{V}$ will hold just by continuity if one replaces $\mathbb{R}^d$ by a large but compact set D and considers in place of the original problem that of estimating $P_{x_0}\{X^n(t) \in D$ for $t \in [0, T)$ and $X^n(T) \in B\}$. This "bounding" of the state space for computational purposes is common, and it is easy to obtain a priori bounds on how large D must be so that $P_{x_0}\{X^n(t) \notin D$ for some $t \in [0, T]\}$ is many orders of magnitude smaller than $P_{x_0}\{X^n(T) \in B\}$. The corresponding definition of a subsolution differs by a condition that holds vacuously (see Remark 14.6), and so any subsolution for the original problem is a subsolution for this problem as well.

Remark 15.5 Note that for an asymptotically efficient scheme, the equality between subsolution and solution is required only at the starting point $(x_0, 0)$. As we will see, for many problems, subsolutions with the optimal value at a given point can be easier to find than the solution. In fact, there are in general many subsolutions with the optimal value at the starting point.

Remark 15.6 The subsolution $\bar{V}(x_0, 0) = 0$ corresponds to standard Monte Carlo, and Theorem 15.1 gives the very poor bound $V(x_0, 0)$ on the rate of decay. Thus any subsolution with $\bar{V}(x_0, 0) > 0$ will improve on standard Monte Carlo, though it is also possible that a scheme could correspond to $\bar{V}(x_0, 0) < 0$ and do even worse than standard Monte Carlo!

Remark 15.7 (*Obtaining asymptotic optimality*) Recall from (14.18) that the mollification $\bar{V}^\delta$ corresponding to a piecewise classical subsolution $\bar{V}$ satisfies the inequality

$$\bar{V}^\delta(x_0, 0) \geq \bar{V}(x_0, 0) - \delta \log J.$$

Consider a sequence $\delta_n > 0$ such that $\delta_n \to 0$ and $n\delta_n \to \infty$. Assuming greater regularity of the component pieces of $\bar{V}^\delta$, it can be shown that the inequality in (15.1) continues to hold with $\bar{V}$ on the left side of (15.1) replaced by $\bar{V}^{\delta_n}$. For the precise statement, see Theorem 15.14.

We next consider the exit probability problem. Consider the ordinary differential equation

$$\dot{\xi}(t) = \int_{\mathbb{R}^d} y\theta(dy|\xi(t)), \ \xi(0) = x. \tag{15.2}$$

Under Condition 4.3, the mapping $x \mapsto \int_{\mathbb{R}^d} y\theta(dy|x)$ is well defined and continuous. Suppose the ODE (15.2) has a unique solution $\{\xi_x(\cdot)\}$ in $\mathscr{C}([0, \infty) : \mathbb{R}^d)$ for every $x \in \mathbb{R}^d$. Let $x^* \in \mathbb{R}^d$ be an asymptotically stable equilibrium point of the ODE, and suppose that A is an open set with $x^* \in A$ and B is a disjoint closed set as in Sect. 14.3, with the problem of interest being to estimate the probability that X_t^n enters B before A after starting as $x_0 \in (A \cup B)^c$.

Remark 15.8 To simplify the analysis we assume as in Remark 15.4 that a "bounding" of the computational domain has already been carried out, so that $(A \cup B)^c$ is bounded, and thus by continuity, $\left\| D\bar{V}(x) \right\|$ is uniformly bounded on $(A \cup B)^c$ for every classical-sense subsolution. To be precise, if $(A \cup B)^c$ is not bounded then we choose a compact set D such that the probability of exiting D prior to entering A or B is negligible compared with the probability of entering B before A. Given such a set, we *redefine* A to be the set $A \cup [D^c \backslash B]$. Then automatically $(A \cup B)^c \subset D$ is bounded. With such a redefinition of A, every classical-sense subsolution or piecewise classical-sense subsolution for the original problem retains this property, since the added requirement is vacuous (see Remark 14.6).

We are interested in the probability that given $X^n(0) = x_0 \in (A \cup B)^c$, the process $X^n(\cdot)$ enters B before reaching A. Thus letting

$$N^n \doteq \inf\{j \in \mathbb{N}_0 : X_j^n \in A \cup B\} \text{ and } \tau^n \doteq N^n/n,$$

we consider the probability of the event $\{X^n(\tau^n) \in B\}$. We will also assume Condition 15.9 below, which in particular ensures that $\tau^n < \infty$ a.s. and therefore $X^n(\tau^n)$ is well defined.

For $x \in (A \cup B)^c \subset D$, define

$$V(x) \doteq \inf[I_T(\phi) : \phi(0) = x, T \in [0, \infty) \text{ and for some } t \le T \\ \phi(t) \in B \text{ and } \phi(s) \in A^c \text{ for all } 0 \le s \le t].$$

Given a classical subsolution $\bar{V}$ of the PDE 14.3 with boundary condition 14.14 (see Definition 14.4), the importance sampling scheme to estimate $P_x\{X^n(\tau^n) \in B\}$ is constructed as in Sect. 14.4. Thus if the current state of the simulated trajectory is Y_i^n, then we replace the original distribution on the noise $v_i(Y_i^n)$, i.e., $\theta(dy|Y_i^n)$, by

$$\eta_\alpha(dv|Y_i^n) = e^{\langle \alpha, v\rangle - H(Y_i^n, \alpha)}\theta(dv|Y_i^n) \text{ with } \alpha = -D\bar{V}(Y_i^n).$$

Using a bound on certain exponential moments of N^n that is stated in (15.32), the second moment of the estimator can be written in terms of the original random variables and process model as

$$\mathfrak{S}^n(\bar{V}) = E_{x_0}\left[1_{\{X_{N^n}^n \in B\}} \prod_{i=0}^{N^n-1} e^{\langle D\bar{V}(X_i^n), v_i(X_i^n)\rangle + H(X_i^n, -D\bar{V}(X_i^n))}\right]. \tag{15.3}$$

Similarly, if $\bar{V}(x) = \wedge_{i=1}^J \bar{V}^{(i)}(x)$ is a piecewise classical subsolution, then the mollification $\bar{V}^\delta(x)$ is defined as in 14.16 but with $\bar{V}^{(j)}(x,t)$ replaced by $\bar{V}^{(j)}(x)$. If the ordinary implementation is used, then the second moment is given by (15.3) with $\bar{V}$ replaced by $\bar{V}^\delta$. If the randomized implementation is used, then the second moment takes the form

$$\mathfrak{S}^n(\bar{V}^\delta) \doteq \\ E_{x_0}\left[1_{\{X_{N^n}^n \in B\}} \prod_{i=0}^{N^n-1} \left(\sum_{j=1}^J \rho_j^\delta\left(X_i^n, \frac{i}{n}\right) e^{-\langle D\bar{V}^{(j)}(X_i^n), v_i(X_i^n)\rangle - H\left(X_i^n, -D\bar{V}^{(j)}(X_i^n)\right)}\right)^{-1}\right].$$

Besides other uses, the following bound ensures that all moments of the time till the simulation terminates are finite. As shown in Proposition 15.19, the bound in Condition 15.9 follows from classical Freidlin–Wentzell arguments [140] under quite general conditions. In the condition, D is a closed bounded set that contains $(A \cup B)^c$.

Condition 15.9 *There exist $c > 0$, $T_0 \in (0, \infty)$ and $n_0 \in \mathbb{N}$ such that for all $n \ge n_0$, $T < \infty$, and $x \in D$,*

$$P_x\{\tau^n > T\} \le \exp\{-cn(T - T_0)\}.$$

The following theorem gives the asymptotic performance bound for the exit probability problem.

Theorem 15.10 *Assume Conditions 4.3 and 15.9 and that $(A \cup B)^c$ is bounded. Let $\bar{V}$ be a classical subsolution for the PDE (14.13) with boundary condition (14.14), construct the corresponding IS estimator as in Sect. 14.4, and consider its second moment $\mathfrak{S}^n(\bar{V})$, for which we have the representation (15.3). Then for every $x_0 \in (A \cup B)^c$, this second moment satisfies*

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathfrak{S}^n(\bar{V}) \geq V(x_0) + \bar{V}(x_0). \tag{15.4}$$

Suppose that $\bar{V} = \wedge_{j=1}^{J} \bar{V}^{(j)}$ is a piecewise classical subsolution and let $\bar{V}^\delta$ denote the mollification defined by (14.16). If one is using either the ordinary or randomized implementation as described in Sect. 14.4, then (15.4) holds with $\bar{V}$ replaced by $\bar{V}^\delta$.

15.2 Performance Bounds for the Finite-Time Problem

In this section we prove Theorem 15.1. As in Chap. 4, we simplify notation by giving the proof for $T = 1$.

We first consider the simpler case of a classical subsolution. Recall from (14.15) that when rewritten in terms of the original process model, the second moment of the scheme defined in terms of a subsolution $\bar{V}$ takes the form

$$\mathfrak{S}^n(\bar{V}) = E_{x_0}\left[1_{\{X_n^n \in B\}} \prod_{i=0}^{n-1} e^{\langle D\bar{V}(X_i^n, i/n), v_i(X_i^n)\rangle + H(X_i^n, -D\bar{V}(X_i^n, i/n))} \right], \tag{15.5}$$

where H is as in (14.6). To obtain this expression for the second moment, we used a change of measure to rewrite the expectation under the original probability law. We now use a second change of measure to give another expression for $\mathfrak{S}^n(\bar{V})$, which will allow a convenient use of the representation obtained in Sect. 4.2. This change of measure will rewrite $\mathfrak{S}^n(\bar{V})$ as an exponential integral with an exponent that is bounded from below. An alternative approach would be to extend the representation of Sect. 4.2 to a class of exponents that are not bounded from one side or the other (i.e., above or below), but we consider the approach based on the second change of measure simpler.

For $n \in \mathbb{N}$ and $i = 0, 1, \ldots, n-1$, define stochastic kernels $\gamma_i^n(dy|x)$ on $\mathbb{R}^d$ given $\mathbb{R}^d$ by

$$\gamma_i^n(dy|x) \doteq e^{\langle D\bar{V}(x, i/n), y\rangle - H(x, D\bar{V}(x, i/n))} \theta(dy|x). \tag{15.6}$$

Note that under Condition 4.3, $\gamma_i^n(\cdot|x)$ is a well-defined probability measure for every x. Let $\{\check{v}_i^n(x), i = 0, 1, \ldots, n-1, x \in \mathbb{R}^d\}$ be independent random vector fields on some probability space with the property that for each $x \in \mathbb{R}^d$, $\check{v}_i^n(x)$ has distribution

$\gamma_i^n(\cdot|x)$. Construct a time-inhomogeneous Markov chain $\{\check{X}_i^n\}_{i=0,\dots,n}$ through the recurrence (14.1) with v_i replaced by $\check{\nu}_i^n$. Then in terms of the sequence $\{\check{X}_i^n\}$, the second moment $\mathfrak{S}^n(\bar{V})$ can be rewritten as

$$\mathfrak{S}^n(\bar{V}) = E_{x_0}\left[1_{\left\{\check{X}_n^n \in B\right\}} \prod_{i=0}^{n-1} e^{H(\check{X}_i^n, D\bar{V}(\check{X}_i^n, i/n)) + H(\check{X}_i^n, -D\bar{V}(\check{X}_i^n, i/n))}\right]. \tag{15.7}$$

Note that by the boundedness assumption on $D\bar{V}$ and the property that

$$\sup_{x\in\mathbb{R}^d} \sup_{\|\alpha\|\le M} H(x,\alpha) < \infty, \quad \text{for all } M \in (0,\infty), \tag{15.8}$$

the functional F to which the representation in (4.4) would be applied to get a representation for $\mathfrak{S}^n(\bar{V})$ is bounded from below (although it is not bounded from above due to the indicator function). Hence Theorem 4.5 can be applied. This gives the representation

$$\begin{aligned} &-\frac{1}{n}\log \mathfrak{S}^n(\bar{V}) \\ &\quad = \frac{1}{n} \inf_{\{\bar{\mu}_i^n\}} E_{x_0}\left[\sum_{i=0}^{n-1}\left[-H\left(\bar{X}_i^n, D\bar{V}\left(\bar{X}_i^n, i/n\right)\right) - H\left(\bar{X}_i^n, -D\bar{V}\left(\bar{X}_i^n, i/n\right)\right)\right]\right. \\ &\qquad\qquad \left. + \sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\,\gamma_i^n(\cdot|\bar{X}_i^n)\right) + \infty 1_{\{\bar{X}_n^n \in B^c\}}\right], \end{aligned} \tag{15.9}$$

where $\{\bar{\mu}_i^n\}$ and $\{\bar{X}_i^n\}$ are as defined in Construction 4.4. We remark that although the representation (4.4) is written for a time-homogeneous Markov chain with a fixed stochastic kernel $\theta(dy|x)$, the representation needed here for a time-inhomogeneous Markov chain given through a time-dependent sequence of stochastic kernels $\gamma_i^n(dy|x)$ can be obtained in a completely analogous manner.

Remark 15.11 At this point, we have introduced many related processes, due to the fact that we use two changes of measure and a representation. To recapitulate, we have the original process $\{X_i^n\}$, the process $\{Y_i^n\}$ that is actually simulated and is related to $\{X_i^n\}$ by a change of measure, the process $\{\check{X}_i^n\}$, which is also related to $\{X_i^n\}$ by a change of measure and used to make the derivation of a representation for the second moment simpler, and the process $\{\bar{X}_i^n\}$ that appears in the representation (15.9). For the present purpose of performance analysis, it is only the last process and the representation (15.9) that are relevant.

We now show that for every sequence of controls and controlled processes $\{\bar{\mu}_i^n\}$ and $\{\bar{X}_i^n\}$,

$$\liminf_{n\to\infty} \frac{1}{n} E_{x_0}\Bigg[\sum_{i=0}^{n-1}\left[-H\left(\bar{X}_i^n, D\bar{V}\left(\bar{X}_i^n, i/n\right)\right) - H\left(\bar{X}_i^n, -D\bar{V}\left(\bar{X}_i^n, i/n\right)\right)\right] \tag{15.10}$$

$$+\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\,\gamma_i^n(\cdot|\bar{X}_i^n)\right) + \infty 1_{\{\bar{X}_n^n \in B^c\}}\Bigg] \geq V(x_0,0) + \bar{V}(x_0,0).$$

For this it suffices to argue that every subsequence has a further subsequence (labeled once more by n) along which (15.10) holds with lim inf replaced by lim sup. Note that we can assume without loss of generality that along this sequence,

$$\sup_{n\in\mathbb{N}} \frac{1}{n} E_{x_0}\left[\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\,\gamma_i^n(\cdot|\bar{X}_i^n)\right)\right] < \infty \text{ and } \bar{X}_n^n \in B \text{ a.s.} \tag{15.11}$$

Henceforth, we fix such a sequence and suppose that the properties (15.11) are satisfied along this sequence.

Recall from Construction 4.4 the noise sequence $\{\bar{v}_i^n\}$ with conditional distributions $\{\bar{\mu}_i^n\}$ and the sequences of measure-valued random variables $\{\bar{L}^n\}$, $\{\bar{\mu}^n\}$, $\{\lambda^n\}$. To prove the tightness of the sequence $\{(\bar{X}^n, \bar{L}^n, \bar{\mu}^n)\}$, we will use Lemma 4.11. The following lemma gives the key estimate that allows the application of Lemma 4.11.

Lemma 15.12 *Assume the conditions of Theorem 15.1 and (15.11). Then*

$$\sup_{n\in\mathbb{N}} \frac{1}{n} E_{x_0}\left[\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n)\right)\right] < \infty.$$

Proof Note that for each $x \in \mathbb{R}^d$ and $\mu \in \mathscr{P}(\mathbb{R}^d)$, (15.6) implies

$$\begin{aligned} &R\left(\mu(\cdot)\,\|\theta(\cdot|x)\right) \qquad (15.12)\\ &\quad = R\left(\mu(\cdot)\,\|\,\gamma_i^n(\cdot|x)\right) + \int_{\mathbb{R}^d} \log\left(\frac{d\gamma_i^n(\cdot|x)}{d\theta(\cdot|x)}(y)\right)\mu(dy)\\ &\quad = R\left(\mu(\cdot)\,\|\,\gamma_i^n(\cdot|x)\right) + \int_{\mathbb{R}^d} \left\langle D\bar{V}(x, i/n), y\right\rangle \mu(dy) - H(x, D\bar{V}(x, i/n)). \end{aligned}$$

Thus, in view of the bound (15.11), the boundedness of $D\bar{V}$, and (15.8), it suffices to show that

$$\sup_n \frac{1}{n} E_{x_0} \sum_{i=0}^{n-1} \|\bar{v}_i^n\| < \infty. \tag{15.13}$$

For this we use a minor modification of Lemma 4.11 with θ replaced by γ_i^n. A similar argument to that used for the proof of (4.8) [see (4.10)] shows that

$$\sup_{n\in\mathbb{N}} \frac{1}{n} E_{x_0} \sum_{i=0}^{n-1} \|\bar{v}_i^n\| = \sup_{n\in\mathbb{N}} E_{x_0} \left[\int_{\mathbb{R}^d \times [0,1]} \|y\| \, \bar{L}^n(dy \times dt) \right]$$
$$\leq \sup_{n\in\mathbb{N}} \sup_{x\in\mathbb{R}^d} \max_{i=0,1,\ldots,n-1} \int_{\mathbb{R}^d} e^{\|y\|} \gamma_i^n(dy|x)$$
$$+ \sup_{n\in\mathbb{N}} \frac{1}{n} E_{x_0} \left[\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot) \,\middle\|\, \gamma_i^n(\cdot|\bar{X}_i^n)\right) \right].$$

The last term is of course bounded by (15.11). Recall the relation between γ_i^n and θ in (15.6). The desired bound (15.13) now follows on noting (for more details see the proof of Lemma 3.9) that due to the boundedness of $D\bar{V}$ and (15.8), for each $m < \infty$ there are $C, M < \infty$ such that for all $i = 0, 1, \ldots, n-1$ and $x \in \mathbb{R}^d$,

$$\sup_{\|\alpha\|\leq m} \int_{\mathbb{R}^d} e^{\langle \alpha, y\rangle} \gamma_i^n(dy|x) = \sup_{\|\alpha\|\leq m} \int_{\mathbb{R}^d} e^{\langle \alpha, y\rangle + \langle D\bar{V}(x, i/n), y\rangle - H(x, D\bar{V}(x,i/n))} \theta(dy|x)$$
$$\leq \sup_{\|\alpha\|\leq M,\, x\in\mathbb{R}^d} C \int_{\mathbb{R}^d} e^{\langle \alpha, y\rangle} \theta(dy|x).$$

Since (15.8) implies that the last quantity is bounded, this completes the proof of the lemma. □

We can now complete the proof of Theorem 15.1. Note that each summand in the third term on the left side of (15.10) can be written as

$$R\left(\bar{\mu}_i^n(\cdot) \,\middle\|\, \gamma_i^n(\cdot|\bar{X}_i^n)\right) = R\left(\bar{\mu}_i^n(\cdot) \,\middle\|\, \theta(\cdot|\bar{X}_i^n)\right) + \int_{\mathbb{R}^d} \log\left(\frac{d\theta(\cdot|\bar{X}_i^n)}{d\gamma_i^n(\cdot|\bar{X}_i^n)}(y) \right) \bar{\mu}_i^n(dy). \tag{15.14}$$

Using (15.6), the second term on the right side of (15.14) takes the form

$$\int_{\mathbb{R}^d} \log\left(\frac{d\theta(\cdot|\bar{X}_i^n)}{d\gamma_i^n(\cdot|\bar{X}_i^n)}(y) \right) \bar{\mu}_i^n(dy)$$
$$= -\int_{\mathbb{R}^d} \left\langle D\bar{V}(\bar{X}_i^n, i/n), y \right\rangle \bar{\mu}_i^n(dy) + H\left(\bar{X}_i^n, D\bar{V}\left(\bar{X}_i^n, i/n\right)\right). \tag{15.15}$$

The last term in (15.15) nicely cancels a term in (15.10), and thus in view of the discussion below (15.10), it suffices to show that

$$\liminf_{n\to\infty} \frac{1}{n} E_{x_0} \left[\sum_{i=0}^{n-1} \left[-\int_{\mathbb{R}^d} \left\langle D\bar{V}(\bar{X}_i^n, i/n), y \right\rangle \bar{\mu}_i^n(dy) - H\left(\bar{X}_i^n, -D\bar{V}\left(\bar{X}_i^n, i/n\right)\right) \right] \right.$$
$$\left. + \sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot) \,\middle\|\, \theta(\cdot|\bar{X}_i^n)\right) + \infty 1_{\{\bar{X}_n^n \in B^c\}} \right] \geq V(x_0, 0) + \bar{V}(x_0, 0). \tag{15.16}$$

As a side remark, note that since a conditioning argument allows $\int_{\mathbb{R}^d} y\bar{\mu}_i^n(dy)$ to be replaced by $\bar{v}_i^n$, this is exactly the expression one would have obtained by formally applying the representation to $\mathfrak{S}^n(\bar{V})$, in spite of the fact that the exponent in $\mathfrak{S}^n(\bar{V})$ is not bounded either above or below.

Since $\bar{V}$ is a subsolution, for $i = 0, 1, \ldots, n-1$ we have

$$-H\left(\bar{X}_i^n, -D\bar{V}\left(\bar{X}_i^n, i/n\right)\right) = \mathbb{H}\left(\bar{X}_i^n, D\bar{V}\left(\bar{X}_i^n, i/n\right)\right) \geq -\bar{V}_t(\bar{X}_i^n, i/n). \quad (15.17)$$

Then (15.17) implies

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n-1} E_{x_0}\left[-\left\langle D\bar{V}(\bar{X}_i^n, i/n), \bar{v}_i^n\right\rangle - H\left(\bar{X}_i^n, -D\bar{V}\left(\bar{X}_i^n, i/n\right)\right)\right] \\ \geq -\frac{1}{n}\sum_{i=0}^{n-1} E_{x_0}\left[\left\langle D\bar{V}(\bar{X}_i^n, i/n), \bar{v}_i^n\right\rangle + \bar{V}_t(\bar{X}_i^n, i/n)\right]. \quad (15.18)\end{aligned}$$

Using the estimate from Lemma 15.12, we can apply Lemma 4.11, which says that $\left\{(\bar{X}^n, \bar{L}^n, \bar{\mu}^n, \lambda^n)\right\}_{n\in\mathbb{N}}$ is tight and also that the uniform integrability estimate in (4.8) holds. The estimate in Lemma 15.12 also allows us to apply Lemma 4.12, which says that every weak limit $(\bar{X}, \bar{L}, \bar{\mu}, \lambda)$ has the properties that $\bar{\mu}$ can be disintegrated as $\bar{\mu}(dy|t)dt$ for a suitable kernel on $[0, 1]$ given $\mathbb{R}^d$, and that (4.15) and (4.16) hold, namely,

$$\begin{aligned}\bar{X}(t) = \int_{\mathbb{R}^d\times[0,t]} y\bar{L}(dy \times ds) + x_0 = \int_{\mathbb{R}^d\times[0,t]} y\bar{\mu}(dy|s)ds + x_0, \ t \in [0, 1], \\ \lambda(A \times B) = \int_B \theta(A|\bar{X}(t))dt, \ A \in \mathscr{B}(\mathbb{R}^d), B \in \mathscr{B}([0, 1]).\end{aligned} \quad (15.19)$$

Also, in view of (15.11) and since B is closed, we have $\bar{X}(1) \in B$ a.s. In proving the inequality (15.16), we can assume without loss of generality that along this sequence, $(\bar{X}^n, \bar{L}^n, \bar{\mu}^n, \lambda^n)$ converges weakly to $(\bar{X}, \bar{L}, \bar{\mu}, \lambda)$. Using Fatou's lemma and the properties of the limit points, we have, just as in the proof of (4.18),

$$\limsup_{n\to\infty} E_{x_0}\left[\frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\middle\|\,\theta(\cdot|\bar{X}_i^n)\right)\right] \geq E_{x_0}\left[\int_0^1 L(\bar{X}(t), \dot{\bar{X}}(t))dt\right]. \quad (15.20)$$

Let $\hat{X}^n$ be the piecewise constant interpolation

$$\hat{X}^n(t) = \bar{X}_i^n, \quad t \in [i/n, i/n + 1/n), \ i = 0, 1, \ldots, n-1. \quad (15.21)$$

Noting that

$$\sup_{0\leq t\leq 1} \|\hat{X}^n(t) - \bar{X}^n(t)\| \leq \frac{1}{n} \sup_{0\leq t\leq 1} \|\bar{X}^n(t)\|,$$

we see that $\hat{X}^n$ also converges weakly in $\mathscr{D}([0,1]:\mathbb{R}^d)$ to $\bar{X}$. From (15.18), we have

$$\frac{1}{n}\sum_{i=0}^{n-1} E_{x_0}\left[-\left\langle D\bar{V}(\bar{X}_i^n, i/n), \bar{v}_i^n\right\rangle - H\left(\bar{X}_i^n, -D\bar{V}\left(\bar{X}_i^n, i/n\right)\right)\right] + O(1) \tag{15.22}$$
$$\geq -E_{x_0}\left[\int_{\mathbb{R}^d\times[0,1]} \left(\left\langle D\bar{V}\left(\hat{X}^n(t),t\right), y\right\rangle + \bar{V}_t\left(\hat{X}^n(t),t\right)\right)\bar{L}^n(dy\times dt)\right],$$

where the error term $O(1)$ is due to replacing $D\bar{V}\left(\bar{X}_i^n, i/n\right)$ by $D\bar{V}(\hat{X}^n(t),t) = D\bar{V}(\bar{X}_i^n, t)$ for $t \in [i/n, i/n + 1/n)$, and we have used the continuity of $D\bar{V}(x,t)$, $\bar{V}_t(x,t)$ and tightness of $\{\bar{X}^n\}$. Using the uniform integrability property (4.8) of $\bar{L}^n$ and since $\bar{L} = \bar{\mu}$, we have

$$\begin{aligned}
&\limsup_{n\to\infty} -E_{x_0}\left[\int_{\mathbb{R}^d\times[0,1]} \left(\left\langle D\bar{V}\left(\hat{X}^n(t),t\right), y\right\rangle + \bar{V}_t\left(\hat{X}^n(t),t\right)\right)\bar{L}^n(dy\times dt)\right] \\
&= -E_{x_0}\left[\int_{\mathbb{R}^d\times[0,1]} \left(\left\langle D\bar{V}\left(\bar{X}(t),t\right), y\right\rangle + \bar{V}_t\left(\bar{X}(t),t\right)\right)\bar{\mu}(dy|t)dt\right] \\
&= -E_{x_0}\left[\int_0^1 \left(\left\langle D\bar{V}\left(\bar{X}(t),t\right), \dot{\bar{X}}(t)\right\rangle + \bar{V}_t\left(\bar{X}(t),t\right)\right)dt\right],
\end{aligned} \tag{15.23}$$

where in the last equality we have used (15.19). Finally, by combining (15.20), (15.22), and (15.23), we see that the left side of (15.16) is bounded below by

$$E_{x_0}\left[\int_0^1 \left(-\left\langle D\bar{V}\left(\bar{X}(t),t\right), \dot{\bar{X}}(t)\right\rangle - \bar{V}_t\left(\bar{X}(t),t\right)\right)dt + \int_0^1 L(\bar{X}(t), \dot{\bar{X}}(t))dt\right]. \tag{15.24}$$

Using that $\bar{X}(0) = x_0$, $\bar{X}(1) \in B$ a.s., and (since $\bar{V}$ is a subsolution) $\bar{V}(x,1) \leq 0$ for $x \in B$, by the (ordinary) chain rule, we have

$$\int_0^1 \left(-\left\langle D\bar{V}\left(\bar{X}(t),t\right), \dot{\bar{X}}(t)\right\rangle - \bar{V}_t\left(\bar{X}(t),t\right)\right)dt \geq \bar{V}(x_0,0) \text{ a.s.}$$

Also, from the definition of $V(x_0,0)$ and using again that $\bar{X}(1) \in B$ a.s., we obtain

$$\int_0^1 L(\bar{X}(t), \dot{\bar{X}}(t))dt \geq V(x_0,0) \text{ a.s.}$$

From the last two observations, we see that the expression in (15.24) is bounded below by $\bar{V}(x_0,0) + V(x_0,0)$, which proves (15.16) and completes the proof of the theorem for the case that $\bar{V}$ is a classical subsolution.

We now consider the second part of the theorem, namely the case that $\bar{V}$ is a piecewise classical subsolution. Let $\delta > 0$ and let $\bar{V}^\delta$ be the corresponding mollification defined by (14.16). The proof for the ordinary implementation is immediate on recalling that $\bar{V}^\delta$ is a classical subsolution. Consider now the randomized implementation.

We will make use of the identities

$$D\bar{V}^{\delta}(x,t)=\sum_{j=1}^{J}\rho_j^{\delta}(x,t)D\bar{V}^{(j)}(x,t) \text{ and } \bar{V}_t^{\delta}(x,t)=\sum_{j=1}^{J}\rho_j^{\delta}(x,t)\bar{V}_t^{(j)}(x,t). \tag{15.25}$$

As noted in Sect. 14.4, the second moment of the scheme based on $\bar{V}^{\delta}$ is equal to

$$\mathfrak{S}^n(\bar{V}^{\delta})=E_{x_0}\left[1_{\{X^n(1)\in B\}}R^n(\{X_i^n,\nu_i(X_i^n)\}_{i=0,\dots,n-1})\right],$$

where

$$R^n(\{X_i^n,\nu_i(X_i^n)\}_{i=0,\dots,n-1})=\prod_{i=0}^{n-1}\left(\sum_{j=1}^{J}\rho_j^{\delta}\left(X_i^n,i/n\right)e^{-\left\langle D\bar{V}^{(j)}\left(X_i^n,i/n\right),\nu_i(X_i^n)\right\rangle-H\left(X_i^n,-D\bar{V}^{(j)}\left(X_i^n,i/n\right)\right)}\right)^{-1}.$$

By Jensen's inequality,

$$\sum_{j=1}^{J}\rho_j^{\delta}\left(X_i^n,i/n\right)e^{-\left\langle D\bar{V}^{(j)}\left(X_i^n,i/n\right),\nu_i(X_i^n)\right\rangle-H\left(X_i^n,-D\bar{V}^{(j)}\left(X_i^n,i/n\right)\right)}$$
$$\geq e^{-\sum_{j=1}^{J}\rho_j^{\delta}\left(X_i^n,i/n\right)\left(\left\langle D\bar{V}^{(j)}\left(X_i^n,i/n\right),\nu_i(X_i^n)\right\rangle-H\left(X_i^n,-D\bar{V}^{(j)}\left(X_i^n,i/n\right)\right)\right)}.$$

Thus in view of (15.25), $\mathfrak{S}^n(\bar{V}^{\delta})$ is bounded above by

$$E_{x_0}\left[1_{\{X_n^n\in B\}}\prod_{i=0}^{n-1}\left(e^{\sum_{j=1}^{J}\rho_j^{\delta}\left(X_i^n,i/n\right)\left[\left\langle D\bar{V}^{(j)}\left(X_i^n,i/n\right),\nu_i(X_i^n)\right\rangle+H\left(X_i^n,-D\bar{V}^{(j)}\left(X_i^n,i/n\right)\right)\right]}\right)\right]$$
$$=E_{x_0}\left[1_{\{X_n^n\in B\}}\prod_{i=0}^{n-1}\left(e^{\left\langle D\bar{V}^{\delta}\left(X_i^n,i/n\right),\nu_i(X_i^n)\right\rangle+\sum_{j=1}^{J}\rho_j^{\delta}\left(X_i^n,i/n\right)H\left(X_i^n,-D\bar{V}^{(j)}\left(X_i^n,i/n\right)\right)}\right)\right].$$

Define $\{\gamma_i^n\}$ by (15.6) with $\bar{V}$ replaced by $\bar{V}^{\delta}$. Then using the fact that $D\bar{V}^{\delta}$ is uniformly bounded by (15.25) and the assumption made for the second part of Theorem 15.1, we have, exactly as in the proof for the case of a classical subsolution,

$$-\frac{1}{n}\log\mathfrak{S}^n(\bar{V}^{\delta}) \tag{15.26}$$
$$\geq\frac{1}{n}\inf_{\{\bar{\mu}_i^n\}}E_{x_0}\left[\sum_{i=0}^{n-1}R\left(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n)\right)+\infty 1_{\{\bar{X}_n^n\in B^c\}}-\sum_{i=0}^{n-1}\left\langle D\bar{V}\left(\bar{X}_i^n,i/n\right),\bar{\nu}_i^n\right\rangle\right.$$
$$\left.-\sum_{i=0}^{n-1}\sum_{j=1}^{J}\rho_j^{\delta}\left(\bar{X}_i^n,i/n\right)H\left(\bar{X}_i^n,-D\bar{V}^{(j)}\left(\bar{X}_i^n,i/n\right)\right)\right].$$

Also, since each $\bar{V}^{(j)}$ is a subsolution, for each $i = 0, 1, \ldots, n-1$, we have

$$-\sum_{j=1}^{J} \rho_j^{\delta}\left(\bar{X}_i^n, \frac{i}{n}\right) H\left(\bar{X}_i^n, -D\bar{V}^{(j)}\left(\bar{X}_i^n, i/n\right)\right) \geq -\sum_{j=1}^{J} \rho_j^{\delta}\left(\bar{X}_i^n, \frac{i}{n}\right) \bar{V}_t^{(j)}(\bar{X}_i^n, i/n)$$
$$= -\bar{V}_t^{\delta}(\bar{X}_i^n, i/n), \tag{15.27}$$

where the equality is a consequence of (15.25). The proof of the second part of the theorem can now be completed exactly as was the proof for the case of a classical subsolution. □

Remark 15.13 Note that the proof of Theorem 15.1 uses little of the particular properties of the underlying process, given that one has used the weak convergence approach to establish the large deviation upper bound, and thus it can be adapted with few changes to other process models.

The following theorem proves the statement in Remark 15.7.

Theorem 15.14 *Assume the conditions of Theorem 15.1 and also that the second derivatives of the form $\partial^2 \bar{V}^{(j)}(x,t)/\partial x_i \partial x_k$ and $\partial^2 \bar{V}^{(j)}(x,t)/\partial x_i \partial t$ exist and are continuous and uniformly bounded for $x \in \mathbb{R}^d$, $t \in [0,T]$, and $j = 1, \ldots, J$. For $\delta > 0$, let $\bar{V}^{\delta}$ be the mollification defined by (14.16) and let $\{\delta_n\}$ be a positive sequence such that $\delta_n \to 0$ and $n\delta_n \to \infty$. Define the IS scheme as in Sect. 14.4 with $\bar{V}^{\delta}$ replaced by $\bar{V}^{\delta_n}$. Then*

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathfrak{S}^n(\bar{V}^{\delta_n}) \geq V(x_0, 0) + \bar{V}(x_0, 0).$$

Proof Replacing δ by δ_n, (15.26) and (15.27), and the analogue of (15.15) with $\bar{V}$ replaced by $\bar{V}^{\delta_n}$ together imply

$$-\frac{1}{n} \log \mathfrak{S}^n(\bar{V}^{\delta_n}) \geq \frac{1}{n} \inf_{\{\bar{\mu}_i^n\}} E_{x_0}\left[\sum_{i=0}^{n-1}\left[-\left\langle D\bar{V}^{\delta_n}(\bar{X}_i^n, i/n), \bar{\nu}_i^n\right\rangle - \bar{V}_t^{\delta_n}(\bar{X}_i^n, i/n)\right]\right.$$
$$\left. + \sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot) \,\|\, \theta(\cdot|\bar{X}_i^n)\right) + \infty 1_{\{\bar{X}_n^n \in B^c\}}\right].$$

Let $r^n(t) \doteq r_1^n(t) + r_2^n(t)$, where

$$r_1^n(t) \doteq \left\langle D\bar{V}^{\delta_n}(\bar{X}_i^n, i/n), \bar{\nu}_i^n\right\rangle - \left\langle D\bar{V}^{\delta_n}(\bar{X}^n(t), t), \dot{\bar{X}}^n(t)\right\rangle$$
$$r_2^n(t) \doteq \bar{V}_t^{\delta_n}(\bar{X}_i^n, i/n) - \bar{V}_t^{\delta_n}(\bar{X}^n(t), t)$$

for $t \in [i/n, i/n + 1/n)$. Then

$$
\begin{aligned}
&-\frac{1}{n}\log \mathfrak{S}^n(\bar{V}^{\delta_n}) \\
&\quad \geq \inf_{\{\bar{\mu}_i^n\}} E_{x_0}\left[\int_0^1 \left[-\left\langle D\bar{V}^{\delta_n}(\bar{X}^n(t),t), \dot{\bar{X}}^n(t)\right\rangle - \bar{V}_t^{\delta_n}(\bar{X}^n(t),t)\right]dt \right. \\
&\qquad\qquad \left. - \int_0^1 r^n(t)dt + \frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\middle\|\,\theta(\cdot|\bar{X}_i^n)\right) + \infty 1_{\{\bar{X}_n^n \in B^c\}}\right].
\end{aligned}
$$

Since there is $C < \infty$ such that $\|D\bar{V}^{\delta_n}(x,t)\| \leq C$ for all x, t and n of interest, it follows that for $M \in (0,\infty)$,

$$
\begin{aligned}
\left|r_1^n(t)\right| &\leq \left\|D\bar{V}^{\delta_n}(\bar{X}_i^n, i/n) - D\bar{V}^{\delta_n}(\bar{X}^n(t),t)\right\| \left\|\bar{v}_i^n\right\| \qquad (15.28)\\
&\leq 2C\left\|\bar{v}_i^n\right\| 1_{\{\|\bar{v}_i^n\| \geq M\}} + \left\|D\bar{V}^{\delta_n}(\bar{X}_i^n, i/n) - D\bar{V}^{\delta_n}(\bar{X}^n(t),t)\right\| M 1_{\{\|\bar{v}_i^n\| < M\}}.
\end{aligned}
$$

Recall that by assumption there is $K \in (0,\infty)$ such that the terms $|\partial \bar{V}^{(j)}(x,t)/\partial x_i|$, $|\partial \bar{V}^{(j)}(x,t)/\partial t|$, $|\partial^2 \bar{V}^{(j)}(x,t)/\partial x_i \partial x_k|$, and $|\partial^2 \bar{V}^{(j)}(x,t)/\partial x_i \partial t|$ are bounded by K for all $(x,t) \in \mathbb{R}^d \times [0,T]$ and all i, k, and j. Using the definition of $\bar{V}^{\delta}$ in (14.16), we see that for some $\bar{K} \in (0,\infty)$, for all $n \in \mathbb{N}$, we have

$$
\sup_{(x,t)\in\mathbb{R}^d\times[0,T]} \max_{i,k}\left\{\left|\partial^2 \bar{V}^{\delta_n}(x,t)/\partial x_i \partial x_k\right| + \left|\partial^2 \bar{V}^{\delta_n}(x,t)/\partial x_i \partial t\right|\right\} \leq \frac{\bar{K}}{\delta_n}.
$$

Using this bound for the second term on the right side of (15.28), we have

$$
\begin{aligned}
&\left\|D\bar{V}^{\delta_n}(\bar{X}_i^n, i/n) - D\bar{V}^{\delta_n}(\bar{X}^n(t),t)\right\| M 1_{\{\|\bar{v}_i^n\| < M\}} \\
&\quad \leq 1_{\{\|\bar{v}_i^n\| < M\}}\left(d\|\bar{X}_i^n - \bar{X}^n(t)\| + \frac{1}{n}\right)\frac{\bar{K}M}{\delta_n} \\
&\quad \leq \frac{(dM+1)M\bar{K}}{n\delta_n}.
\end{aligned}
$$

Once more we can assume without loss of generality that

$$
\sup_{n\in\mathbb{N}} \frac{1}{n} E_{x_0}\left[\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\middle\|\,\theta(\cdot|\bar{X}_i^n)\right)\right] < \infty.
$$

Thus by Lemma 4.11,

$$
\limsup_{n\to\infty} E_{x_0}\left|\int_0^1 r_1^n(t)dt\right| \leq \limsup_{M\to\infty}\limsup_{n\to\infty} 2C\frac{1}{n} E_{x_0}\sum_{i=0}^{n-1}\left\|\bar{v}_i^n\right\| 1_{\{\|\bar{v}_i^n\| \geq M\}} = 0.
$$

A similar bound applies to r_2^n, whose details are omitted. The chain rule, the constraint $\bar{X}^n(1) \in B$ w.p.1, and $\bar{V}^{\delta_n}(x,1) \leq 0$ for $x \in B$ then imply

$$-\frac{1}{n}\log \mathfrak{S}^n(\bar{V}^{\delta_n}) \geq \bar{V}^{\delta_n}(x_0,0) \tag{15.29}$$
$$+\inf_{\{\bar{\mu}_i^n\}} E_{x_0}\left[\int_0^1 r^n(t)dt + \frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\middle\|\,\theta(\cdot|\bar{X}_i^n)\right) + \infty 1_{\{\bar{X}_n^n \in B^c\}}\right],$$

and the proof now proceeds exactly as for the proof of Theorem 15.1.

Remark 15.15 (*Nonasymptotic bounds*) If we in addition assume that there is $\sigma > 0$ such that $E\exp\{\sigma \|v_i(x)\|^2\} < \infty$ for all $x \in \mathbb{R}^d$, then nonasymptotic bounds can be derived. A sketch of the argument is as follows. For the piecewise classical case, the mollification $\bar{V}^\delta$ has all second derivatives bounded by a constant times $1/\delta$. Hence one has the bound

$$\left|r^n(t)\right| \leq \frac{1}{n\delta_n}\left(K_0 + K_1\left\|\bar{v}_i^n\right\| + K_2\left\|\bar{v}_i^n\right\|^2\right)$$

for $t \in [i/n, i/n + 1/n)$ and suitable $K_j < \infty$, $j = 0, 1, 2$. Therefore,

$$\left|E_{x_0}\int_0^1 r^n(t)dt\right| \leq \frac{1}{n\delta_n}E_{x_0}\left[K_0 + K_1\frac{1}{n}\sum_{i=0}^{n-1}\|\bar{v}_i^n\| + K_2\frac{1}{n}\sum_{i=0}^{n-1}\|\bar{v}_i^n\|^2\right].$$

The claim that there is a uniform bound on

$$E_{x_0}\left[\frac{1}{n}\sum_{i=0}^{n-1}\|\bar{v}_i^n\|^2\right]$$

follows from the same argument used to get (15.13), where we use the assumption $\sup_x Ee^{\sigma\|v_i(x)\|^2} < \infty$ for some $\sigma > 0$ in place of $\sup_x Ee^{\sigma\|v_i(x)\|} < \infty$. Thus (15.29) becomes

$$\begin{aligned}
&-\frac{1}{n}\log \mathfrak{S}^n(\bar{V}^{\delta_n}) \\
&\geq \bar{V}^{\delta_n}(x_0,0) - \frac{K}{n\delta_n} + \inf_{\{\bar{\mu}_i^n\}} E_{x_0}\left[\frac{1}{n}\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\middle\|\,\theta(\cdot|\bar{X}_i^n)\right) + \infty 1_{\{\bar{X}_n^n \in B^c\}}\right] \\
&= \bar{V}^{\delta_n}(x_0,0) - \frac{K}{n\delta_n} - \frac{1}{n}\log P_{x_0}\left\{X^n(1) \in B\right\},
\end{aligned}$$

where the constant $K < \infty$ can be estimated based on the problem data. The worst possible "loss" due to the mollification is of the form

$$-\delta_n \log J - \frac{K}{n\delta_n},$$

and that occurs when x_0 is close to a point where $\bar{V}$ is not smooth. To minimize this worst possible loss, one would minimize $\delta \to \delta \log J + K/(n\delta)$, which suggests δ_n proportional to $n^{-1/2}$, leading to a loss of $[\log J + K]/n^{1/2}$. If x_0 is away from the places where $\bar{V}$ is not smooth, then the difference between $\bar{V}^\delta$ and $\bar{V}$ scales like $e^{-c/\delta}$. Minimizing a quantity of the form $e^{-c/\delta} + K/n\delta$ leads to $\delta_n = c/[\log n + \log(c/K)]$ and a loss proportional to $\log n/n$.

If mollification is not needed, then the nonasymptotic bound is of the form $\bar{V}(x_0, 0) - K/n - \log P_{x_0}\{X^n(1) \in B\}/n$ for all points, which appears qualitatively better.

Remark 15.16 (*Limits rather than bounds*) One can ask whether a limit for the decay rate of the second moment associated with a subsolution $\bar{V}$ is possible. The answer is yes, though one must strengthen the assumptions to include the types of regularity properties needed on the process and set B so that a large deviation lower bound holds. We consider the case of a classical subsolution. Using (15.9), (15.12), and (15.14) gives the representation

$$
\begin{aligned}
&-\frac{1}{n}\log \mathfrak{S}^n(\bar{V}) \\
&\quad = \frac{1}{n}\inf_{\{\bar{\mu}_i^n\}} E_{x_0}\left[\sum_{i=0}^{n-1}\left[-\int_{\mathbb{R}^d}\left\langle D\bar{V}(\bar{X}_i^n, i/n), y\right\rangle \bar{\mu}_i^n(dy) + \mathbb{H}\left(\bar{X}_i^n, D\bar{V}\left(\bar{X}_i^n, i/n\right)\right)\right]\right. \\
&\qquad\qquad \left. + \sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n)\right) + \infty 1_{\{\bar{X}_n^n \in B^c\}}\right],
\end{aligned}
$$

where $D\bar{V}(x,t)$ is bounded and continuous, as is $\mathbb{H}(x, D\bar{V}(x,t))$. Hence this is in the form of the representation of a Laplace functional for the pair (L^n, X^n). Arguing as in Chap. 4, one can establish the limit

$$
\begin{aligned}
&\lim_{n\to\infty} -\frac{1}{n}\log \mathfrak{S}^n(\bar{V}) \\
&\quad = \inf_{\phi}\left[\int_0^1 \left(\mathbb{H}\left(\phi(t), D\bar{V}(\phi(t), t)\right) - \left\langle D\bar{V}(\phi(t), t), \dot{\phi}(t)\right\rangle\right) dt\right. \\
&\qquad\qquad \left. + \int_0^1 L(\phi(t), \dot{\phi}(t))dt + \infty 1_{\{\phi(1)\in B^c\}}\right],
\end{aligned}
$$

where the infimum is over absolutely continuous ϕ with $\phi(0) = x_0$. The additional condition that is needed is that there exists a nearly infimizing ϕ such that $\phi(1) \in B^\circ$. If $\bar{V}$ is a subsolution, it is easy to check using the subsolution property that one recovers the bound given previously, i.e., the infimum in the last display is bounded below by $V(x_0, 0) + \bar{V}(x_0, 0)$.

15.3 Performance Bounds for the Exit Probability Problem

We give the proof of Theorem 15.10 in this section. Details are given only for the case of a classical subsolution, since the proof for mollified piecewise classical subsolutions can be completed in the same way as in Sect. 15.2. Recalling that $(A \cup B)^c \subset D$ is bounded and by modifying $\bar{V}$ outside a neighborhood of D if needed, we can assume without loss of generality that

$$\sup_{x \in \mathbb{R}^d} [|V(x)| + \|DV(x)\|] < \infty. \tag{15.30}$$

By the subsolution property of $\bar{V}$, for $i = 0, 1, \ldots, N^n - 1$, we have

$$H(X_i^n, -D\bar{V}(X_i^n)) = -\mathbb{H}(X_i^n, D\bar{V}(X_i^n)) \le 0.$$

Thus (15.3) implies

$$\mathfrak{S}^n(\bar{V}) \le E_{x_0}\left[1_{\{X_{N^n}^n \in B\}} \prod_{i=0}^{N^n-1} e^{\langle D\bar{V}(X_i^n), v_i(X_i^n)\rangle}\right]. \tag{15.31}$$

An immediate consequence of Condition 15.9 is that for the given n_0 and some $\gamma \in (0, \infty)$,

$$K_\gamma \doteq \sup_{n \ge n_0} \sup_{x \in D} e^{-ncT_0} E_x[e^{\gamma N^n}] < \infty. \tag{15.32}$$

Let $X_0^n = x_0 \in (A \cup B)^c$. We use Taylor's theorem to write

$$\begin{aligned}
&\bar{V}(X_{N^n}^n) - \bar{V}(x_0) \\
&\quad = \frac{1}{n} \sum_{i=0}^{N^n-1} \left\langle D\bar{V}(X_i^n), v_i(X_i^n)\right\rangle + \frac{1}{n} \sum_{i=0}^{N^n-1} \left\langle D\bar{V}(Z_i^n) - D\bar{V}(X_i^n), v_i(X_i^n)\right\rangle,
\end{aligned}$$

where Z_i^n is on the line segment joining X_i^n and X_{i+1}^n, and therefore $\left\|Z_i^n - X_i^n\right\| \le \left\|v_i(X_i^n)\right\|/n$. Thus

$$\sum_{i=0}^{N^n-1} \left\langle D\bar{V}(X_i^n), v_i(X_i^n)\right\rangle \le 2n\|\bar{V}\|_\infty - \sum_{i=0}^{N^n-1} \left\langle D\bar{V}(Z_i^n) - D\bar{V}(X_i^n), v_i(X_i^n)\right\rangle. \tag{15.33}$$

Let $T \in (0, \infty)$ and define

$$\mathfrak{S}_{1T}^n(\bar{V}) \doteq e^{2n\|\bar{V}\|_\infty} E_{x_0}\left[1_{\{\tau^n > T\}} \prod_{i=0}^{N^n-1} e^{\langle D\bar{V}(X_i^n) - D\bar{V}(Z_i^n), v_i(X_i^n)\rangle}\right],$$

$$\mathfrak{S}^n_{2T}(\bar{V}) \doteq E_{x_0}\left[1_{\{\tau^n \le T,\, X^n_{N^n} \in B\}} \prod_{i=0}^{N^n-1} e^{\langle D\bar{V}(X^n_i), v_i(X^n_i)\rangle}\right].$$

Then (15.31) and (15.33) imply $\mathfrak{S}^n(\bar{V}) \le \mathfrak{S}^n_{1T}(\bar{V}) + \mathfrak{S}^n_{2T}(\bar{V})$. We first show for large but finite T that $\mathfrak{S}^n_{1T}(\bar{V})$ is unimportant, thus reducing back to the finite-time case.

Reduction to the case of finite time. Since $\bar{V}$ is uniformly continuous on D, given $h \in (0, 1)$ there is $\delta(h) > 0$ such that for all $x, y \in D$,

$$\|x - y\| \le \delta(h) \Rightarrow \|D\bar{V}(x) - D\bar{V}(y)\| \le h.$$

For $h > 0$, $n \in \mathbb{N}$, and $v \in \mathbb{R}^d$, let

$$f_{n,h}(v) \doteq e^{2\|D\bar{V}\|_\infty \|v\|} 1_{\{\|v\| \ge n\delta(h)\}} + e^{h\|v\|} 1_{\{\|v\| < n\delta(h)\}}.$$

Since for $v \in \mathbb{R}^d$,

$$\left|\left\langle D\bar{V}(Z^n_i) - D\bar{V}(X^n_i), v\right\rangle\right| \le \min\left\{2\|D\bar{V}\|_\infty \|v\|, \|D\bar{V}(Z^n_i) - D\bar{V}(X^n_i)\|\|v\|\right\},$$

we have from the choice of $\delta(h)$ that

$$e^{\langle D\bar{V}(X^n_i) - D\bar{V}(Z^n_i), v_i(X^n_i)\rangle} \le f_{n,h}(v_i(X^n_i)). \tag{15.34}$$

Let $f_h(v) \doteq e^{h\|v\|}$. Then for every $h \in (0, 1)$, $f_{n,h}$ converges pointwise to f_h as $n \to \infty$.

We claim that as $n \to \infty$,

$$\int_{\mathbb{R}^d} f^2_{n,h}(v)\theta(dv|x) \to \int_{\mathbb{R}^d} f^2_h(v)\theta(dv|x), \text{ uniformly in } x \in D. \tag{15.35}$$

To prove the claim, it suffices to show that if $x_n \to x$, with $x_n, x \in D$, then

$$\int_{\mathbb{R}^d} f^2_{n,h}(v)\theta(dv|x_n) \to \int_{\mathbb{R}^d} f^2_h(v)\theta(dv|x).$$

From continuity of $x \mapsto \theta(\cdot|x)$ and the uniform bound on the moment-generating function (Condition 4.3), as $n \to \infty$,

$$\int_{\mathbb{R}^d} f^2_h(v)\theta(dv|x_n) \to \int_{\mathbb{R}^d} f^2_h(v)\theta(dv|x). \tag{15.36}$$

Also, for $K \in (0, \infty)$, $h \le 1$ implies

$$\int_{\mathbb{R}^d} |f_{n,h}^2(v) - f_h^2(v)|\theta(dv|x_n) \le \int_{\{\|v\|\le K\}} |f_{n,h}^2(v) - f_h^2(v)|\theta(dv|x_n) + 2\int_{\{\|v\|\ge K\}} e^{2(1+2\|D\bar{V}\|_\infty)\|v\|}\theta(dv|x_n).$$

For $n > K/\delta(h)$, the first term is zero. Using Condition 4.3 again, the second term approaches 0 as $K \to \infty$, uniformly in n. From this it follows that the left side of the last display approaches 0 as $n \to \infty$. The uniform convergence in (15.35) now follows from this and (15.36).

Recall γ introduced in (15.32). Using Condition 4.3 once more, we can choose $h \in (0, 1)$ such that

$$\sup_{x\in D}\int_{\mathbb{R}^d} e^{2h\|v\|}\theta(dy\,|\,x) < e^{\gamma/4}.$$

Recall the parameter n_0 from Condition 15.9. Combining the last display with (15.35), for $h \in (0, 1)$ there is $n_1 \in [n_0, \infty)$ such that for all $n \ge n_1$,

$$\sup_{x\in D}\int_{\mathbb{R}^d} f_{n,h}^2(v)\theta(dy\,|\,x) < e^{\gamma/2}. \tag{15.37}$$

We now bound $\mathfrak{S}_{1T}^n(\bar{V})$ using

$$e^{-2n\|\bar{V}\|_\infty}\mathfrak{S}_{1T}^n(\bar{V}) \le E_{x_0}\left[1_{\{\tau^n>T\}}\prod_{i=0}^{N^n-1} f_{n,h}(v_i(X_i^n))\right]$$
$$\le \left(E_{x_0}\left[1_{\{\tau^n>T\}}e^{\gamma N^n/2}\right]\right)^{1/2}\left(E_{x_0}\left[\prod_{i=0}^{N^n-1} e^{-\gamma/2}f_{n,h}^2(v_i(X_i^n))\right]\right)^{1/2},$$

where the first inequality uses (15.34) and the second follows from the Cauchy-Schwarz inequality. Note that (15.37) implies $U_m \doteq \prod_{i=0}^{m-1} e^{-\gamma/2}f_{n,h}^2(v_i(X_i^n))$, $m \in \mathbb{N}$, and $U_0 \doteq 1$ is a nonnegative supermartingale, and therefore

$$E_{x_0}\left[\prod_{i=0}^{N^n-1} e^{-\gamma/2}f_{n,h}^2(v_i(X_i^n))\right] = E_{x_0}[U_{N^n}] \le 1.$$

Thus using $1_{\{\tau^n>T\}} \le e^{\gamma N^n/2-\gamma Tn/2}$, for all $n \ge n_1$,

$$e^{-2n\|\bar{V}\|_\infty}\mathfrak{S}_{1T}^n(\bar{V}) \le \left(E_{x_0}\left[1_{\{\tau^n>T\}}e^{\gamma N^n/2}\right]\right)^{1/2}$$
$$\le e^{-\gamma nT/4}\left(E_{x_0}\left[e^{\gamma N^n}\right]\right)^{1/2}$$
$$\le e^{-\gamma nT/4}\left(K_\gamma e^{cnT_0}\right)^{1/2},$$

where the last inequality uses (15.32). Thus

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathfrak{S}^n_{1T}(\bar V)\le-\frac{\gamma T}{4}+\frac{cT_0}{2}+2\|\bar V\|_\infty.$$

In proving Theorem 15.10, we can assume without loss of generality that

$$K\doteq-\limsup_{n\to\infty}\frac{1}{n}\log\mathfrak{S}^n(\bar V)<\infty.$$

Choosing T large enough that $\frac{\gamma T}{4}-\frac{cT_0}{2}-2\|\bar V\|_\infty>K$, we thus have

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log\mathfrak{S}^n(\bar V)&\le\limsup_{n\to\infty}\frac{1}{n}\log\left(\mathfrak{S}^n_{1T}(\bar V)+\mathfrak{S}^n_{2T}(\bar V)\right)\\&\le\max\left\{\limsup_{n\to\infty}\frac{1}{n}\log\mathfrak{S}^n_{1T}(\bar V),\limsup_{n\to\infty}\frac{1}{n}\log\mathfrak{S}^n_{2T}(\bar V)\right\}\\&=\limsup_{n\to\infty}\frac{1}{n}\log\mathfrak{S}^n_{2T}(\bar V).\end{aligned}$$

Completion of the proof. We can now focus on the finite-time problem associated with $\mathfrak{S}^n_{2T}(\bar V)$, using very much the same ideas as those of the last section. Letting $N^n_T\doteq N^n\wedge\lfloor nT\rfloor$, we can rewrite $\mathfrak{S}^n_{2T}(\bar V)$ as

$$\mathfrak{S}^n_{2T}(\bar V)\doteq E_{x_0}\left[1_{\{X^n_{N^n_T}\in B\}}\prod_{i=0}^{N^n_T-1}e^{\langle D\bar V(X^n_i),v_i(X^n_i)\rangle}\right].$$

Recalling (15.30), we see that

$$\gamma(dy|x)\doteq e^{\langle D\bar V(x),y\rangle-H(x,D\bar V(x))}\theta(dy|x)$$

defines a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d$. Let $\check v_i$, $\check X^n_{i+1}$, $i=0,1,\dots,\lfloor nT\rfloor$, be defined as below (15.6). Then just as (15.5) became (15.7), $\mathfrak{S}^n_{2T}(\bar V)$ can be written as

$$\mathfrak{S}^n_{2T}(\bar V)=E_{x_0}\left[1_{\{\check X^n_{\check N^n_T}\in B\}}\prod_{i=0}^{\check N^n_T-1}e^{H(\check X^n_i,D\bar V(\check X^n_i))}\right],$$

where $\check\tau^n$, $\check N^n$, $\check N^n_T$ are the analogues of τ^n, N^n, N^n_T, defined in terms of the sequence $\check X^n$. Applying the representation in (4.4) as in (15.9) yields

$$\begin{aligned}-\frac{1}{n}\log\mathfrak{S}^n_{2T}(\bar V)=\frac{1}{n}\inf_{\{\bar\mu^n_i\}}E_{x_0}\Bigg[&\sum_{i=0}^{\bar N^n_T-1}-H\left(\bar X^n_i,D\bar V\left(\bar X^n_i\right)\right)\\&+\sum_{i=0}^{\lfloor nT\rfloor-1}R\left(\bar\mu^n_i(\cdot)\,\|\,\gamma(\cdot|\bar X^n_i)\right)+\infty 1_{\left\{\hat X^n\in C^c_T\right\}}\Bigg],\end{aligned}$$

where $\bar{N}_T^n$ is the analogue of N_T^n defined in terms of the sequence $\{\bar{X}_i^n\}$, C_T is the set of trajectories in $\mathscr{D}([0,T]:\mathbb{R}^d)$ that enter B at some time $t \le T$ without having entered A previously, and $\hat{X}^n$ is the piecewise constant interpolation of $\{\bar{X}_i^n\}$ as defined in (15.21). We note that without loss of generality, the second sum in the last display can be replaced by a sum up to $\bar{N}_T^n - 1$ by replacing $\bar{\mu}_i^n$ with

$$\bar{\mu}_i^n(\cdot)1_{\{i<\bar{N}_T^n\}} + \gamma(\cdot|\bar{X}_i^n)1_{\{i\ge\bar{N}_T^n\}}.$$

We now show that for every sequence of controls and controlled processes $\{(\bar{\mu}_i^n, \bar{X}_i^n)\}$,

$$\liminf_{n\to\infty} \frac{1}{n} E_{x_0}\left[\sum_{i=0}^{\bar{N}_T^n-1} -H\left(\bar{X}_i^n, D\bar{V}\left(\bar{X}_i^n\right)\right)\right. \tag{15.38}$$
$$\left. + \sum_{i=0}^{\bar{N}_T^n-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\,\gamma(\cdot|\bar{X}_i^n)\right) + \infty 1_{\left\{\hat{X}^n\in C_T^c\right\}}\right] \ge V(x_0) + \bar{V}(x_0).$$

For this, once more a subsequential argument implies we can replace lim inf by lim sup, and then we can assume without loss of generality that

$$\sup_{n\in\mathbb{N}} \frac{1}{n} E_{x_0}\left[\sum_{i=0}^{\bar{N}_T^n-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\,\gamma(\cdot|\bar{X}_i^n)\right)\right] = \sup_{n\in\mathbb{N}} \frac{1}{n} E_{x_0}\left[\sum_{i=0}^{\lfloor nT\rfloor-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\,\gamma(\cdot|\bar{X}_i^n)\right)\right]$$
$$< \infty$$

and $\hat{X}^n \in C_T$ a.s. for every $n \in \mathbb{N}$.

Using the fact that $D\bar{V}$ is bounded, exactly as in the proof of Lemma 15.12, we see that

$$\sup_n \frac{1}{n} E_{x_0}\left[\sum_{i=0}^{n-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\,\theta(\cdot|\bar{X}_i^n)\right)\right] < \infty.$$

From this, the property that $\bar{\tau}^n \le T$ a.s., and Lemmas 15.12 and 4.12, we get as before that

$$\left\{(\bar{X}^n, \bar{L}^n, \bar{\mu}^n, \lambda^n, \bar{\tau}^n)\right\}_{n\in\mathbb{N}}$$

is tight. Moreover, the uniform integrability estimate in (4.8) holds, and every weak limit $(\bar{X}, \bar{L}, \bar{\mu}, \lambda, \bar{\tau})$ has the properties noted in (15.19). By taking a convergent subsequence, we can assume that $(\bar{X}^n, \bar{L}^n, \bar{\mu}^n, \lambda^n, \bar{\tau}^n)$ converges weakly to $(\bar{X}, \bar{L}, \bar{\mu}, \lambda, \bar{\tau})$. Since $\bar{X}(\bar{\tau}) \in B$ a.s., $\bar{X}(0) \notin B$, and B is closed, it follows that $\bar{\tau} > 0$ a.s. Also, since $\bar{\tau}^n \le T$ a.s., we have $\bar{\tau} \le T$. Using Fatou's lemma and a minor modification of the argument in (4.18), we obtain

$$\begin{aligned}
&\liminf_{n\to\infty} E_{x_0}\left[\frac{1}{n}\sum_{i=0}^{\bar N_T^n-1} R\left(\bar\mu_i^n(\cdot)\,\|\theta(\cdot|\bar X_i^n)\right)\right]\\
&\quad=\liminf_{n\to\infty} E_{x_0}\left[\bar\tau^n R\left(\frac{1}{\bar\tau^n}\bar\mu^n(dy\times(dt\cap[0,\bar\tau^n]))\,\middle\|\,\frac{1}{\bar\tau^n}\lambda^n(dy\times(dt\cap[0,\bar\tau^n]))\right)\right]\\
&\quad\ge E_{x_0}\left[\bar\tau R\left(\frac{1}{\bar\tau}\bar\mu(dy\times(dt\cap[,\bar\tau]))\,\middle\|\,\frac{1}{\bar\tau}\lambda(dy\times(dt\cap[0,\bar\tau]))\right)\right]\\
&\quad= E_{x_0}\left[\int_{[0,\bar\tau]} R\left(\bar\mu(\cdot|t)\,\|\theta(\cdot|\bar X(t))\right)dt\right]\\
&\quad\ge E_{x_0}\left[\int_{[0,\bar\tau]} L(\bar X(t),\dot{\bar X}(t))dt\right].
\end{aligned}\tag{15.39}$$

Also, using (15.15), we have

$$\begin{aligned}
&\frac{1}{n}E_{x_0}\sum_{i=0}^{\bar N_T^n-1}\int_{\mathbb{R}^d}\log\left(\frac{d\theta(\cdot|\bar X_i^n)}{d\gamma(\cdot|\bar X_i^n)}(y)\right)\bar\mu_i^n(dy)-\frac{1}{n}E_{x_0}\sum_{i=0}^{\bar N_T^n-1}H\left(\bar X_i^n, D\bar V\left(\bar X_i^n\right)\right)\\
&\quad=-\frac{1}{n}E_{x_0}\sum_{i=0}^{\bar N_T^n-1}\int_{\mathbb{R}^d}\left\langle D\bar V(\bar X_i^n), y\right\rangle\bar\mu_i^n(dy)\\
&\quad=-E_{x_0}\int_{\mathbb{R}^d\times[0,\bar\tau^n]}\left\langle D\bar V(\hat X^n(t)), y\right\rangle\bar L^n(dy\times dt).
\end{aligned}\tag{15.40}$$

Using the uniform integrability of $\bar L^n$ as in (15.23), we have

$$\begin{aligned}
&\liminf_{n\to\infty} -E_{x_0}\int_{\mathbb{R}^d\times[0,\bar\tau^n]}\left\langle D\bar V(\hat X^n(t)), y\right\rangle\bar L^n(dy\times dt)\\
&\qquad=-E_{x_0}\int_{[0,\bar\tau]}\langle D\bar V(\bar X(t)),\dot{\bar X}(t)\rangle dt.
\end{aligned}$$

Combining (15.40) with (15.39) and using (15.14) [with γ_i^n replaced by γ], we have

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}E_{x_0}\left[-\sum_{i=0}^{\bar N_T^n-1}H\left(\bar X_i^n, D\bar V\left(\bar X_i^n\right)\right)+\sum_{i=0}^{\bar N_T^n-1}R\left(\bar\mu_i^n(\cdot)\,\|\gamma(\cdot|\bar X_i^n)\right)+\infty 1_{\left\{\hat X^n\in C_T^c\right\}}\right]\\
&\quad\ge E_{x_0}\left[-\int_0^{\bar\tau}\left\langle D\bar V\left(\bar X(t)\right),\dot{\bar X}(t)\right\rangle dt+\int_0^{\bar\tau}L(\bar X(t),\dot{\bar X}(t))dt\right].
\end{aligned}$$

Note that C_T is a closed set in $\mathscr{C}([0,T]:\mathbb{R}^d)$, and so $\bar X\in C_T$ a.s. Also recall that $\bar X(\bar\tau)\in B$ a.s. as well. From these observations and the definition of V, it follows that

$$\int_0^{\bar\tau}L(\bar X(t),\dot{\bar X}(t))dt\ge V(x_0).$$

By the chain rule,

$$-\int_0^{\bar{\tau}} \left\langle D\bar{V}\left(\bar{X}(t)\right), \dot{\bar{X}}(t)\right\rangle dt = \bar{V}(x_0) - \bar{V}(\bar{X}(\tau)) \geq \bar{V}(x_0),$$

where the last inequality uses the fact that $\bar{V}(x) \leq 0$ for all $x \in B$. Combining the last two displays, we have (15.38), and the result follows. □

Remark 15.17 The argument used for the escape time problem consists in showing that for large enough $T < \infty$, the contribution to the second moment due to escape after time T is negligible, and then using the same arguments as those used for the finite-time problem. Because of this, it is straightforward to show that natural analogues of Theorem 15.14, which shows how one can let $\delta \to 0$ as $n \to \infty$ for mollified piecewise classical subsolutions; Remark 15.15, which gives nonasymptotic bounds on the second moment; and Remark 15.16, which gives a limit for the normalized logarithm of the second moment rather than a bound, all hold here as well.

We close this section by giving a set of sufficient conditions under which Condition 15.9 holds. Recall that ξ_x was defined as a solution to the ODE (15.2) with initial condition x. We assume uniqueness of the solution to the ODE in this condition, but if this does not hold, then one can formulate an analogous condition that requires all solutions to be attracted to x^*. The proof of the proposition is similar to that of [140, Lemma 4.2.2], and thus only a sketch is provided. For a set $K \subset \mathbb{R}^d$, let $d(y, K) \doteq \inf[\|y - x\| : x \in K]$, and for sets K_1 and K_2, let $d(K_1, K_2) \doteq \inf[d(y, K_2) : y \in K_1]$. In the statement of the condition, $I_{x,T}(\phi)$ equals $I_T(\phi)$ if $\phi(0) = x$ and is ∞ otherwise. Let $\mathscr{C}([0, \infty) : \mathbb{R}^d)$ denote the space of continuous functions from $[0, \infty)$ to $\mathbb{R}^d$. This is a Polish space when equipped with the topology of local uniform convergence (i.e., uniform convergence on every compact interval).

Condition 15.18 *The following properties hold.*

(a) The map $x \mapsto \xi_x$ is continuous from $\mathbb{R}^d$ to $\mathscr{C}([0, \infty) : \mathbb{R}^d)$.
(b) Let $D^\kappa \doteq \{y \in \mathbb{R}^d : d(y, D) < \kappa\}$. For some $\kappa > 0$ and all $x \in D^\kappa$, $\xi_x(t) \to x^*$ as $t \to \infty$.
(c) For every $T \in (0, \infty)$, $M \in (0, \infty)$, and compact $K \subset \mathbb{R}^d$,

$$\cup_{x \in K} \{\phi \in \mathscr{C}([0, T] : \mathbb{R}^d) : I_T(\phi) \leq M, \phi(0) = x\}$$

is a compact subset of $\mathscr{C}([0, T] : \mathbb{R}^d)$.
(d) For every $T < \infty$, the sequence $\{X^n\}$ satisfies the large deviation upper bound on $\mathscr{C}([0, T] : \mathbb{R}^d)$ with rate functions $\{I_{x,T}, x \in \mathbb{R}^d\}$, uniformly on compacts (in the sense of Definition 1.11).

Proposition 15.19 *Assume Conditions 4.3 and 15.18. Then there exist* $c, T_0 \in (0,\infty)$ *and* $n_0 \in \mathbb{N}$ *such that for all* $n \geq n_0$, $T < \infty$, *and* $x \in D$, *one has*

$$P_x\{\tau^n > T\} \leq \exp\{-cn(T - T_0)\}.$$

Proof Fix $\bar{\kappa} \in (0, \kappa)$ such that $d(B(x^*, \bar{\kappa}), A^c) > \bar{\kappa}$, where for $\delta > 0$, $B(x^*, \bar{\kappa})$ is the open ball of radius δ centered at x^*. For $x \in \mathbb{R}^d$, let

$$T(x) \doteq \inf\{t > 0 : \xi_x(t) \in B(x^*, \bar{\kappa})\}.$$

Since $x \mapsto \xi_x$ is continuous, the map $x \mapsto T(x)$ is upper semicontinuous, and consequently $T_0 \doteq \max_{x \in D^{\bar{\kappa}}} T(x) < \infty$. Define the closed set

$$\mathscr{C}_{\bar{\kappa}}(T_0) \doteq \{\phi \in \mathscr{C}([0, T_0] : \mathbb{R}^d) : \phi(t) \in D^{\bar{\kappa}} \setminus B(x^*, \bar{\kappa}) \text{ for all } t \in [0, T_0]\}.$$

Note that $\mathscr{C}_{\bar{\kappa}}(T_0) \cap \{\xi_x(\cdot \wedge T_0),\ x \in D^{\bar{\kappa}}\} = \emptyset$. Combining this with the fact that for $x \in \mathbb{R}^d$, $L(x, v) > 0$ if $v \neq \int_{\mathbb{R}^d} y\theta(dy|x)$, we see that $I_{T_0}(\phi) > 0$ for every $\phi \in \mathscr{C}_{\bar{\kappa}}(T_0)$. We claim that for some $r > 0$,

$$I_{T_0}(\phi) > r \text{ for all } \phi \in \mathscr{C}_{\bar{\kappa}}(T_0).$$

Indeed, if the claim is false, by part (c) of Condition 15.18 there is $\phi^* \in \mathscr{C}_{\bar{\kappa}}(T_0)$ such that $I_{T_0}(\phi^*) = 0$, which is a contradiction.

For $a < \infty$, $x \in \mathbb{R}^d$, and $T < \infty$, let

$$\Phi_{x,T}(a) \doteq \left\{\phi \in \mathscr{C}([0, T] : \mathbb{R}^d) : \phi(0) = x, I_T(\phi) \leq a\right\}.$$

Note that for all $x \in D$, $\Phi_{x,T_0}(r) \cap \mathscr{C}_{\bar{\kappa}}(T_0) = \emptyset$, and therefore $\phi \in \Phi_{x_0,T_0}(r)$ that implies $\phi(t)$ enters $B(x^*, \bar{\kappa})$ for some $t \in [0, T_0]$. Also, since $d(B(x^*, \bar{\kappa}), A^c) \geq \bar{\kappa}$, and $d(D, (D^{\bar{\kappa}})^c) \geq \bar{\kappa}$, it follows that

$$d(X^n(\cdot \wedge T_0), \Phi_{x_0,T_0}(r)) \geq \bar{\kappa}, \text{ on the set } \{\tau^n > T_0\}.$$

Using part (e) of Condition 15.18, it follows that there exists $n_0 \in \mathbb{N}$ such that for every $x_0 \in D$ and $n \geq n_0$,

$$P_{x_0}\{\tau^n > T_0\} \leq P_{x_0}\left\{d(X^n(\cdot \wedge T_0), \Phi_{x_0,T_0}(r)) \geq \bar{\kappa}\right\} \leq \exp\{-nr/2\}.$$

A standard argument using the Markov property now shows that for all such n and x_0,

$$P_{x_0}\{\tau^n > T\} \leq \exp\left\{-\frac{nr}{2}\left(\frac{T}{T_0} - 1\right)\right\},$$

and the result follows. □

15.4 Notes

The theory presented in this chapter is based on similar results that have appeared in the papers [103, 105, 110, 114], but it improves on them in various ways. The improvements include limits on the normalized log of the second moment of the estimator, where previously only one-sided bounds had been obtained, nonasymptotic bounds, etc. The first paper to use the convenient double change of measure for the analysis of importance sampling was [112].

Chapter 16
Multilevel Splitting

An alternative to importance sampling in estimating rare events and related functionals is *multilevel splitting*. In the context of estimating probabilities of a set $\mathscr{C}$ in path space, the multilevel splitting philosophy is to simulate particles that evolve according to the law of $\{X_i\}$, and at certain times split those particles considered more likely to lead to a trajectory that belongs to the set $\mathscr{C}$. For example, $\mathscr{C}$ might be the trajectories that reach some unlikely set B before hitting a likely set A, after starting in neither A nor B. In this case, the splitting will favor migration toward B. Splitting can also be used to enhance the sampling of regions that are important for a given integral. In all cases, particles which are split are given an appropriate weighting to ensure that the algorithm remains unbiased.

Broadly speaking, there are two types of multilevel splitting algorithms, those with killing and those without, where *stopping* is distinguished from killing. In the example just mentioned, particles are stopped upon entry into either A or B. Killing involves abandoning a particle prior to entry into either A or B, presumably because continuation of the trajectory is not worth the computational effort. Care must be taken that any killing will not introduce bias.

To the authors' knowledge, there is only one type of multilevel splitting algorithm without killing—the splitting algorithm (see [148] for further references). The standard implementation of this algorithm requires a sequence of sets $C_J \supset C_{J-1} \supset \cdots \supset C_0$, splitting thresholds called **splitting thresholds**, and a sequence of positive integers $R_{J-1}, \ldots, R_0$, splitting rates called **splitting rates**. A single particle is started at the initial position $x_0 \in C_J \backslash C_{J-1}$ and evolves according to the law of $\{X_i\}$. When a particle enters a set C_j for the first time, it produces $R_j - 1$ offspring. After splitting has occurred, all particles evolve independently of each other. Each particle is stopped according to whatever stopping rule is associated with $\{X_i\}$, and the algorithm terminates when all the particles generated have been stopped. The probability of interest is approximated by $N/\prod_{i=0}^{J-1} R_i$, where N is the number of particles simulated whose trajectories belong to $\mathscr{C}$. A more general version of this algorithm lets the splitting rates R_i take nonnegative real values, in which case the number of offspring is randomized.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_16

A large deviation analysis of ordinary splitting is given in [76], which shows that it performs quite well when the thresholds are chosen properly. Although the splitting algorithm can be very effective, there is one clear source of inefficiency in dealing with rare events. The vast majority of the particles generated will not have trajectories that belong to the set $\mathscr{C}$, and so much of the computational effort is devoted to generating trajectories that do not make any direct contribution. Multilevel splitting algorithms with killing were introduced as a way to mitigate this problem. One of the first such algorithms was the RESTART (Repetitive Simulation Trials After Reaching Threshold) algorithm, introduced in [241, 242] (for others, see [208] and the references therein). Its implementation is identical to that of the standard splitting algorithm except that particles are split every time they enter a splitting threshold (and not just at the first entrance time, as with the standard splitting scheme), and particles are killed when they exit the splitting threshold in which they were born. The initial particle is assumed to be born in the set C_J, which by convention is equal to the state space of the process $\{X_i\}$, and so this particle is never killed.

The standard version of the RESTART algorithm requires that the splitting rates be integer-valued and that the process not cross more than one splitting threshold at each time step. However, its definition implicitly allows one to design an algorithm in which the process can cross more than one threshold in each time step. The issue of allowing the process to cross more than one threshold in any given time step was first addressed explicitly in the context of the DPR (Direct Probability Redistribution) algorithm, introduced in [152, 153].

In this chapter we will develop a theory for multilevel splitting, and in particular the RESTART/DPR algorithm, which parallels the theory for importance sampling that was developed in the last two chapters. Since the algorithm is notationally much more complicated than importance sampling, to simplify the presentation we consider only the case of estimating small probabilities, and refer to [77] for expected values.

Although the statements of performance analysis for splitting are often similar to those for importance sampling, it should be noted that there is an important distinction between the types of subsolution required for the two methods. For importance sampling, one needs functions that are *classical*-sense subsolutions. In contrast, splitting-type schemes require only a subsolution in a weak sense (see Definition 16.12). Indeed, this is in some sense expected, since importance sampling uses the gradient of the subsolution to construct the algorithm, while splitting uses only the function itself. For some problems it is easier to construct weak-sense subsolutions, in which case splitting-type schemes can be easier to apply. These and related issues will be discussed and illustrated by examples in Chap. 17.

When comparing importance sampling and splitting, one must recognize that the work used to produce samples need not be the same, and in fact, depending on circumstances, one method can be strongly favored over the other. However, when one uses subsolutions to design splitting schemes, the comparison simplifies somewhat, especially if the performance measure for importance sampling is the exponential decay rate. Suppose one were to consider, say, a work-normalized relative error [see (16.21)]. We will show that the computing cost of splitting grows

subexponentially when a subsolution is used. Thus the main issue in comparing splitting to importance sampling is to compare decay rates. This assumes that the implementation of importance sampling is relatively straightforward, i.e., given a subsolution, it is easy to compute the needed alternative sampling distribution. Since splitting simulates using only the original dynamics, it may be preferred when this is not the case, as in some multiscale models.

In the rest of this chapter, unless explicitly stated otherwise, by a splitting algorithm we mean the RESTART/DPR algorithm. We focus on this version of splitting, since in our experience it is usually preferable to ordinary splitting, and we will just note that analogous versions of all the statements presented here apply to ordinary splitting [76], and with much easier proofs. In Sect. 16.3, we derive formulae for the computational cost and second moment of the algorithm. These will be used in the asymptotic analysis, and Sect. 16.4 considers the asymptotic problem. In Sect. 16.4, a method for designing RESTART/DPR algorithms based on the subsolution framework is developed. Expressions for the asymptotic work-normalized error of such algorithms are derived using the formulas developed in the previous section, and subsolutions that lead to asymptotically optimal performance are identified. The formulation of splitting that is appropriate for finite-time problems of the same type as that considered in the context of importance sampling in Chaps. 14 and 15 is presented in Sect. 16.5.

16.1 Notation and Terminology

Let $\{X_i\}_{i\in\mathbb{N}_0}$ be a Markov chain with state space $\mathbb{R}^d$ for some $d \in \mathbb{N}$. Although we will later consider processes $\{X_i\}_{i\in\mathbb{N}_0}$ as elements of a sequence that satisfies a large deviation property, for notational simplicity the large deviation index is initially suppressed. Until Sect. 16.5, we focus on estimating

$$P_{x_0}\{X_M \in B\}, \tag{16.1}$$

where $M \doteq \inf\{i : X_i \in A \cup B\}$, and as in Sect. 14.1, A is open, B is closed, and $A \cap B = \emptyset$. Although not necessary, to simplify some arguments we will assume as in Remark 15.8 that $(A \cup B)^c$ is bounded and that its closure is denoted by D.

The following notation will be used. Branching processes, which take values in $\cup_{n=1}^{\infty}(\mathbb{R}^d)^n$, are denoted by $\{Z_i\}_{i\in\mathbb{N}_0}$. Each branching process has an $\mathbb{N}$-valued process $\{N_i\}_{i\in\mathbb{N}_0}$ associated with it, where N_i equals the number of particles present in the branching process at time i. As will be explained later in the section, particles are born through branching of existing particles when they reach certain thresholds, while particles die when they exit certain regions.

For each $i \in \mathbb{N}_0$ and $j = 1, \ldots, N_i$, $Z_{i,j}$ denotes the state of the jth particle at time i. We also define a measure on $\mathbb{R}^d$ associated with such a branching process by a random measure associated with branching processes

$$\bar{\delta}_{Z_i} \doteq \sum_{j=1}^{N_i} \delta_{Z_{i,j}}.$$

Note that this is typically not a probability measure, and it is referred to as the unnormalized empirical measure. If $Z_{i,j}$ is in either A or B, then it is killed at the next time step, and so will be counted only once in this measure. Note that this killing is distinct from the killing introduced for algorithmic efficiency.

Splitting schemes are often defined in terms of "importance functions." Later on, these importance functions will be identified with subsolutions translated by a constant, and we use U to denote such an object.[1] To be precise, an importance function is a continuous mapping $U : \mathbb{R}^d \to \mathbb{R}$ that is bounded from below. As we will see, it is only the relative values of $U(x)$ at different points that matter, and so we assume for simplicity of notation that $U(x) \geq 0$ for all $x \in \mathbb{R}^d$. There is also a parameter $\Delta \in (0, \infty)$ such that $R \doteq e^{\Delta} \in \{2, 3, \ldots\}$, and we define closed sets C_j by

$$C_j \doteq \{x \in D : U(x) \leq j\Delta\}$$

for $0 \leq j \leq J - 1 \doteq \lceil U(x_0)/\Delta \rceil - 1$ and $C_J \doteq D$. Note that $x_0 \notin C_{J-1}$. We also define a piecewise *constant* function $\bar{U}$ by setting $\bar{U}(x) = 0$ for $x \in C_0$ and $j\Delta$ if $x \in C_j \backslash C_{j-1}$

$$\bar{U}(x) = j\Delta \text{ for } x \in C_j \backslash C_{j-1},\ j - 0, 1, \ldots, J,$$

where we follow the convention $C_{-1} = \emptyset$. After we introduce the large deviation scaling, it will be possible to obtain a collection of importance functions corresponding to a collection of values of the large deviation index from a single "generating" function in a convenient manner. While it would be possible to allow the splitting rate R or the spacing Δ between levels to depend on j, we will not do so once this scaling is used, and so to simplify notation, we have chosen not to do so here.

Later on, U will be derived from a subsolution $\bar{V}$ that satisfies a boundary condition (i.e., $\bar{V}(x) \leq 0$ for $x \in B$). One possibility will be to let $c \doteq \min[\bar{V}(x) : x \in B]$, with $c \leq 0$ due to the boundary condition, and then let $U(x)$ be equal to $[\bar{V}(x) - c] \vee 0$. In this case, as illustrated in Fig. 16.1, $C_0 \cap B$ may be smaller than B. Although the process stops when B is entered, if it crosses into C_1 or C_2 without entering B, the branching will continue. Thus the number of thresholds crossed before entering B depends on where B is entered. An alternative is to take $U(x)$ equal to $\bar{V}(x) \vee 0$, in which case $\bar{V}(x) \leq 0$ for $x \in B$ implies $B \subset C_0$, as in Fig. 16.2. The rate of decay of the second moment will be the same for both schemes, though one might expect slightly better performance from the first scheme.

Given an importance function U and $x \in D$, let $\sigma(x)$ be the unique integer j such that $x \in C_j \backslash C_{j-1}$ unique integer j such that $x \in C_j \backslash C_{j-1}$. We let $\bar{U}_k$ denote

[1] While in the usual definition, importance functions *increase* as one approaches B, with the identification with subsolutions it will be convenient to have them decrease.

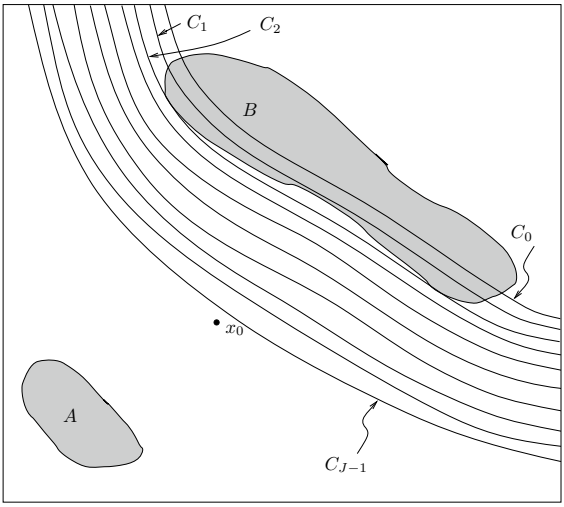

Fig. 16.1 Splitting based on $[\bar{V}(x) - c] \vee 0$

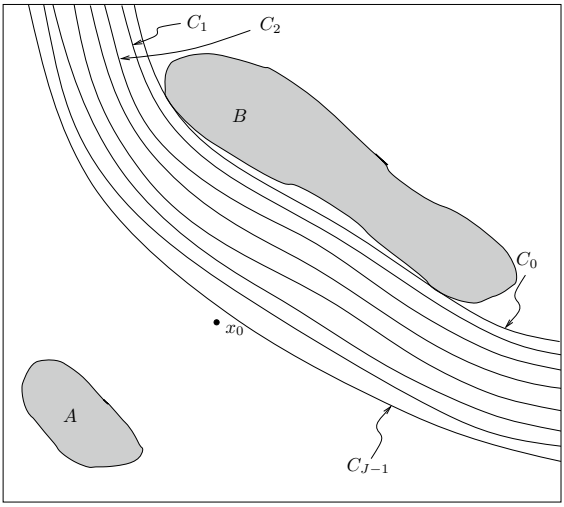

Fig. 16.2 Splitting based on $\bar{V}(x) \vee 0$

the common value of $\bar{U}(x)$ for all x such that $\sigma(x) = k$ $\bar{U}(x)$ if $\sigma(x) = k$ (i.e., $x \in C_k \backslash C_{k-1}$), and note that with this notation,

$$e^{-\bar{U}(x)} = e^{-\bar{U}_{\sigma(x)}} = e^{-\bar{U}_k} = R^{-k} \text{ for all } x \in C_k \backslash C_{k-1}, k = J, J-1, \ldots, 0. \quad (16.2)$$

With ordinary splitting, weights are assigned to particles so that a unit mass associated with a single particle starting at any location is partitioned evenly among the descendants at each splitting. Thus (16.2) are natural weightings for the given splitting rates, in that the fraction of the mass associated with each such descendant after k thresholds are crossed is R^{-k}. The issue is more subtle with RESTART, since particles reaching B can be the result of more splitting events than the number of intervening thresholds (due to multiple reentries of a particle into a splitting threshold). Nonetheless, owing to the killing, R^{-k} is still the correct weight to apply, as will be shown in the proof of unbiasedness.

In the standard version of the RESTART algorithm, splitting is fairly simple. Every time a particle enters a splitting threshold C_j, the deterministic number $R-1$ of offspring are generated, so that including the parent, R particles result. Also, particles are destroyed when they exit the splitting threshold in which they are born, and thus each particle has an integer attached to it to record this splitting threshold. These are referred to as the support thresholds of the particles. While in general, one could allow the number of particles to be random and thereby accommodate arbitrary $R \in (0, \infty)$, there seems to be little practical benefit in doing so, and we restrict to the case in which R is an integer. In this chapter we will consider multilevel splitting, which accounts for the fact that the particles can jump more than one level in a single step. In such a case, different support thresholds must be assigned to the offspring. In order to analyze this mathematically, it is convenient to use the following notation. Let $\mathbb{S}$ be the set of elements $q \in \mathbb{N}_0^\infty$ such that $q_j = 0$ for all sufficiently large j. Vectors $q \in \mathbb{S}$ will be referred to as splitting vectors.

Consider a particle in a multilevel splitting algorithm that moves from $C_j \backslash C_{j-1}$ to $C_k \backslash C_{k-1}, k < j$, in a given time step. Then splitting will occur, and all offspring as well as the original parent particle will be located in $C_k \backslash C_{k-1}$. The support threshold of each new particle will be an element of $\{k, \ldots, j-1\}$, and numbers of offspring and their support thresholds will be independent of all past data except through the values of j and k. It follows that the splitting of a particle is equivalent to assigning to each particle that splits a vector $q \in \mathbb{S}$. The number of *new* particles will be equal to $\sum_{l=0}^{\infty} q_l$, and precisely q_l of the new particles will be given support threshold l. Given that each particle generates $R-1$ descendants upon moving from $C_{j+1} \backslash C_j$ to $C_j \backslash C_{j-1}$, it is clear that when moving from $C_j \backslash C_{j-1}$ to $C_k \backslash C_{k-1}$, we should use the splitting vector $q(j,k)$ defined by $q_l(j,k) \doteq 0$ if either $l \geq j$ or $l < k$, and splitting vectors

$$q_l(j,k) \doteq (R-1)R^{j-l-1} \text{ if } k \leq j-1 \text{ and } k \leq l \leq j-1. \tag{16.3}$$

Note that $\sum_{l \in \mathbb{N}_0} q_l(j,k) = R^{j-k} - 1$, and so including the original particle, exactly R^{j-k} particles are produced. We take $q_l(j,k) = 0$ for all l if $j \leq k$.

16.2 Formulation of the Algorithm

To define the algorithm, we assume the following condition. Conditions that imply the finiteness of M are given in Proposition 15.19.

Condition 16.1 *M is almost surely finite.*

Following the standard logic of acceleration methods generally, the hope is that with a well-chosen importance function $U(x)$, the variance of the estimator is made lower than that of standard Monte Carlo by building in information regarding the underlying process and the event of interest.

In order to analyze the algorithm, we will need some recursive formulas. Observe that if we examine a generic particle at some time after the algorithm has started, then it will be in a set of the form $C_j \backslash C_{j-1}$ and have a killing threshold in $\{j, \ldots, J\}$. For a Markov property to hold, if we imagine starting the process with an initial condition in $C_j \backslash C_{j-1}$, $j < J$, then the support threshold is part of the state variable, and so we must also assign a distribution to the support threshold that is consistent with the dynamics prior to entering $C_j \backslash C_{j-1}$. The distribution of the support threshold of the initial particle will be denoted by $\mathscr{I}$, and will be referred to as the *initializing distribution*. The correct form for such initializing distributions will be identified later on.

The estimator of (16.1), rewritten for a general initial condition x_0 (i.e., one not necessarily in $C_J \backslash C_{J-1}$), is

$$\sum_{i=0}^{\infty} \int_{\mathbb{R}^d} 1_B(x) e^{\bar{U}(x)-\bar{U}(x_0)} \bar{\delta}_{Z_i}(dx).$$

We recall that particles are killed the step after entering A or B, and so contribute to the sum for at most one time index. The weighting term $e^{\bar{U}(x)-\bar{U}(x_0)}$ is important when the number of thresholds crossed before reaching B depends on where the particle is located in B, as in Fig. 16.1.

The splitting thresholds, splitting rates, and splitting vectors of the algorithm will be defined using importance functions U and initialization distributions $\mathscr{I}$ as described previously. In Theorem 16.3 it will be shown that the algorithm is unbiased when the initializing distributions have a prescribed form that will be identified below. The algorithm, with the dependence on these quantities suppressed in the notation, can be written in pseudocode as follows.

RESTART/DPR Algorithm

```
Variables:
  i current time
  N_i number of particles at time i
  Z_{i,j} position of j^{th} particle at time i
  C_{i,j} current threshold of j^{th} particle at time i
  L_{i,j} support threshold of j^{th} particle at time i
  γ (at termination) an estimator of P_{x_0}{X_M ∈ B}
  j, k, l counting variables
  Y_{i,j} free variables
Initialization Step:
  N_0 = 1, Z_{0,1} = x_0, C_{0,1} = σ(x_0), γ = 0, i = 0
  generate a random variable L with distribution 𝒥
  L_{0,1} = L
Main Algorithm:
  while N_i ≠ 0
    N_{i+1} = 0
    for j = 1, ..., N_i
      Test to see if the particle is not killed due
      to stopping:
      if Z_{i,j} ∉ A ∪ B
        generate a random variable Y_{i,j} with
        distribution P{Y_{i,j} ∈ dy} = P{X_{i+1} ∈ dy|X_i = Z_{i,j}}

        Test to see if the particle is not killed
        due to threshold:
        if σ(Y_{i,j}) ≤ L_{i,j}
          N_{i+1} = N_{i+1} + 1
          Z_{i+1,N_{i+1}} = Y_{i,j}
          C_{i+1,N_{i+1}} = σ(Y_{i,j})
          L_{i+1,N_{i+1}} = L_{i,j}
        end

        Test to see if particle should be branched:
        if σ(Y_{i,j}) < C_{i,j}
          for k = 1, ..., J
            for l = 1, ..., q_k(C_{i,j}, σ(Y_{i,j}))
              N_{i+1} = N_{i+1} + 1
              Z_{i+1,N_{i+1}} = Y_{i,j}
              C_{i+1,N_{i+1}} = σ(Y_{i,j})
              L_{i+1,N_{i+1}} = k
            end
          end
        end
      end
    Test to see if the particle is stopped:
      if Z_{i,j} ∈ A ∪ B
        γ = γ + 1_B(Z_{i,j}) e^{Ū(Z_{i,j})}
      end
    end
    i = i + 1
  end
  γ = e^{-Ū(x_0)} γ
```

Remark 16.2 The pseudocode just given presents a "parallel" version of the algorithm, in that all particles for a given threshold are split and then simulated either to the next threshold, the stopping criteria, or killing. Alternatively, one can implement a "sequential" version in which a particle is simulated until it either reaches the stopping criteria or is killed, recording where appropriate the number of additional particles that remain to be simulated for each threshold. After the current particle has been simulated to termination, the algorithm reverts to the highest threshold for which particles remain to be simulated, and starts a new particle.

Note that the output of the algorithm is indeed equal to the desired quantity

$$\gamma = e^{-\bar{U}(x_0)} \sum_{i=0}^{\infty} \int_{\mathbb{R}^d} 1_B(y) e^{\bar{U}(y)} \bar{\delta}_{Z_i}(dy). \tag{16.4}$$

An algorithm resulting from an importance function U, the collection of splitting vectors $q(j,k)$, and an initializing distribution $\mathscr{I}$ will be said to be unbiased if

$$E_{x_0}\left[\gamma\right] = P_{x_0}\{X_M \in B\}.$$

Recall that the splitting rates R are defined in terms the level $\Delta > 0$ that was used to partition U through $R = e^{\Delta}$. Define the vector

$$q_l \doteq \begin{cases} 1, & l = J, \\ (R-1)R^{J-l-1}, & 0 \le l \le J-1. \end{cases} \tag{16.5}$$

In the setting of ordinary splitting, where particles are branched only when they enter a threshold for the first time, q_l is the number of descendants that would be born in threshold l if all particles descending from a single particle in threshold J were to make it to l. We then define probability distributions λ_k on $\{k, \ldots, J\}$ by initializing distribution for splitting

$$\lambda_k(l) \doteq q_l / R^{J-k} = \begin{cases} R^{k-J}, & l = J, \\ (R-1)R^{k-l-1}, & k \le l \le J-1. \end{cases} \tag{16.6}$$

We extend the definition of λ_k to $\{0, \ldots, J\}$ by setting $\lambda_k(l) = 0$ for $l = 0, \ldots, k-1$.

Theorem 16.3 *Fix $x_0 \in (A \cup B)^c$ and suppose that $X_0 = x_0$. Let U be an importance function, and assume Condition 16.1. If the initializing distribution is $\mathscr{I} = \lambda_{\sigma(x_0)}$ and the splitting vectors $q(j,k)$ are as in (16.3), then the resulting splitting algorithm is unbiased.*

The proof of Theorem 16.3 relies on the following lemma. Recall that as in the pseudocode, support threshold of particle m at time i

$$L_{i,m} \doteq \text{ support threshold of particle } m \text{ at time } i,$$

and that $\bar{\delta}_{Z_i} \doteq \sum_{m=1}^{N_i} \delta_{Z_{i,m}}$.

Lemma 16.4 *Assume Condition 16.1 and let h be a nonnegative function on $\mathbb{R}^d$. Let $\bar{h}(x) = h(x)e^{\bar{U}(x)}$ and let $i \in \mathbb{N}_0$ be given. For a splitting scheme with $\mathscr{I} = \lambda_{\sigma(x_0)}$ and $q(j,k)$ as in (16.3),*

$$e^{-\bar{U}(x_0)} E_{x_0}\left[\sum_{m=1}^{N_i} \bar{h}(Z_{i,m}) 1_{\{L_{i,m}=l\}}\right] = E_{x_0}\left[h(X_i)\lambda_{\sigma(X_i)}(l) 1_{\{M\geq i\}}\right],\ l = 0, 1, \ldots, J,$$

and

$$e^{-\bar{U}(x_0)} E_{x_0}\left[\int_{\mathbb{R}^d} \bar{h}(y)\bar{\delta}_{Z_i}(dy)\right] = E_{x_0}\left[h(X_i)1_{\{M\geq i\}}\right].$$

Proof We assume that $M > 0$, since otherwise, the lemma is trivial. The second result is obtained by summing the first one over l. Recall that $Z_{i,j}$ records the location of particle number j at time i. We will prove the first display by induction on i. The result holds for $i = 0$, since in this case there is only a single particle with support threshold distribution $\lambda_{\sigma(x_0)}(l)$, and thus

$$\begin{aligned} e^{-\bar{U}(x_0)} E_{x_0}\left[\bar{h}(Z_{0,1}) 1_{\{L_{0,1}=l\}}\right] &= e^{-\bar{U}(x_0)}\bar{h}(x_0)\lambda_{\sigma(x_0)}(l) \\ &= E_{x_0}\left[h(X_0)\lambda_{\sigma(X_0)}(l)1_{\{M\geq 0\}}\right]. \end{aligned}$$

Suppose the result has been proved up to some time i^*. We then claim that

$$\begin{aligned} &e^{-\bar{U}(x_0)} E_{x_0}\left[\sum_{j=1}^{N_{i^*+1}} \bar{h}(Z_{i^*+1,j}) 1_{\{L_{i^*+1,j}=l\}}\right] \\ &\quad = e^{-\bar{U}(x_0)} E_{x_0}\left[e^{\bar{U}_{\sigma(Z_{0,1})}-\bar{U}_{\sigma(Z_{1,1})}} E_{Z_{1,1}}\left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m}) 1_{\{L_{i^*,m}=l\}}\right]\right]. \end{aligned} \tag{16.7}$$

Thus we currently have one particle located at x_0, whose support threshold lies in $\{\sigma(x_0), \ldots, J\}$. To prove the claim, we will compute the above expectation by conditioning on $Y_{0,1}$ as it appears in the pseudocode, which, we recall, has the distribution $P\{Y_{0,1} \in dy\} = P\{X_1 \in dy | X_0 = Z_{0,1}\}$. It suffices to show that for all $y \in \mathbb{R}^d$,

$$\begin{aligned} &E_{x_0}\left[\sum_{j=1}^{N_{i^*+1}} \bar{h}(Z_{i^*+1,j}) 1_{\{L_{i^*+1,j}=l\}} \,\middle|\, Y_{0,1} = y\right] \\ &\quad = e^{\bar{U}_{\sigma(x_0)}-\bar{U}_{\sigma(y)}} E_y\left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m}) 1_{\{L_{i^*,m}=l\}}\right]. \end{aligned} \tag{16.8}$$

Decomposing according to the support threshold, which has initializing distribution $\lambda_{\sigma(x_0)}(\cdot)$, we have

$$E_{x_0}\left[\sum_{j=1}^{N_{i^*+1}} \bar{h}(Z_{i^*+1,j})1_{\{L_{i^*+1,j}=l\}} \,\middle|\, Y_{0,1}=y\right]$$
$$= \sum_{k=\sigma(x_0)}^{J} \lambda_{\sigma(x_0)}(k) E_{x_0,y,k}\left[\sum_{m=1}^{N_{i^*+1}} \bar{h}(Z_{i^*+1,m})1_{\{L_{i^*+1,m}=l\}}\right], \tag{16.9}$$

where $E_{x_0,y,k}$ denotes expected value given $Z_{0,1}=x_0$, $Y_{0,1}=y$, and $L_{0,1}=k$. Similarly, $E_{x,k}$ will denote the expected value given $Z_{0,1}=x$ and $L_{0,1}=k$. Note that by the Markov property,

$$E_{x_0,y,k}\left[\sum_{m=1}^{N_{i^*+1}} \bar{h}(Z_{i^*+1,m})1_{\{L_{i^*+1,m}=l\}}\right]$$
$$= E_{x_0,y,k}\left[\sum_{r=1}^{N_1} E_{Z_{1,r},L_{1,r}}\left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m})1_{\{L_{i^*,m}=l\}}\right]\right].$$

Thus the expression in (16.9) can be written as

$$\sum_{k=\sigma(x_0)}^{J} \lambda_{\sigma(x_0)}(k) E_{x_0,y,k}\left[\sum_{r=1}^{N_1} E_{Z_{1,r},L_{1,r}}\left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m})1_{\{L_{i^*,m}=l\}}\right]\right]. \tag{16.10}$$

We now consider the expression (16.10) for the three cases $\sigma(y)=\sigma(x_0)$, $\sigma(y)>\sigma(x_0)$, and $\sigma(y)<\sigma(x_0)$, and show that in each of these cases, the expression equals the right side of (16.8).

Consider the first case $\sigma(y)=\sigma(x_0)$. In this case, neither killing nor branching occurs, and so we have $N_1=1$, $Z_{1,1}=y$, and $L_{1,1}=L_{0,1}=k$. Thus (16.10) can be written as

$$\sum_{k=\sigma(y)}^{J} \lambda_{\sigma(y)}(k) E_{y,k}\left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m})1_{\{L_{i^*,m}=l\}}\right] = E_y\left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m})1_{\{L_{i^*,m}=l\}}\right],$$

which equals the right side of (16.8), since $e^{\bar{U}_{\sigma(x_0)}-\bar{U}_{\sigma(y)}}=1$.

Consider next the second case $\sigma(y)>\sigma(x_0)$. In this case, the particle has moved to a threshold with higher index, and branching does not occur. Recall that the particle is killed if and only if $k<\sigma(y)$, since this means that the particle exited its support threshold. Thus for $\sigma(x_0)\le k<\sigma(y)$, since $N_1=0$,

$$E_{x_0,y,k}\left[\sum_{r=1}^{N_1} E_{Z_{1,r},L_{1,r}}\left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m})1_{\{L_{i^*,m}=l\}}\right]\right]=0.$$

Also, if $k \geq \sigma(y)$, then $N_1 = 1$ and $L_{1,1} = L_{0,1} = k$. Since $\lambda_{\sigma(x_0)}(k)/\lambda_{\sigma(y)}(k) = R^{\sigma(x_0)-\sigma(y)} = e^{\bar{U}_{\sigma(x_0)}-\bar{U}_{\sigma(y)}}$, (16.10) in this case can be written as

$$\begin{aligned}
&\sum_{k=\sigma(y)}^{J} \lambda_{\sigma(x_0)}(k) E_{y,k} \left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m}) 1_{\{L_{i^*,m}=l\}} \right] \\
&= \sum_{k=\sigma(y)}^{J} \frac{\lambda_{\sigma(x_0)}(k)}{\lambda_{\sigma(y)}(k)} \lambda_{\sigma(y)}(k) E_{y,k} \left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m}) 1_{\{L_{i^*,m}=l\}} \right] \\
&= e^{\bar{U}_{\sigma(x_0)}-\bar{U}_{\sigma(y)}} \sum_{k=\sigma(y)}^{J} \lambda_{\sigma(y)}(k) E_{y,k} \left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m}) 1_{\{L_{i^*,m}=l\}} \right],
\end{aligned}$$

which once more equals the right side of (16.8).

Finally, consider the case $\sigma(y) < \sigma(x_0)$. Here there is the possibility that new particles are created (i.e., $N_1 > 1$), though in all cases we have $Z_{1,r} = y$. When new particles are created, the associated thresholds are determined according to the measure $q_l(j,k)$, and so using (16.3) and the definition (16.6), (16.10) takes the form

$$\begin{aligned}
&\sum_{k=\sigma(x_0)}^{J} \lambda_{\sigma(x_0)}(k) \left[\sum_{j=\sigma(y)}^{\sigma(x_0)-1} q_j(\sigma(x_0), \sigma(y)) E_{y,j} \left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m}) 1_{\{L_{i^*,m}=l\}} \right] \right. \\
&\left. + E_{y,k} \left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m}) 1_{\{L_{i^*,m}=l\}} \right] \right].
\end{aligned}$$

Since the sum $\sum_{j=\sigma(y)}^{\sigma(x_0)-1}$ has no k dependence, using $\sum_{k=\sigma(x_0)}^{J} \lambda_{\sigma(x_0)}(k) = 1$, that for $j \in \{\sigma(y), \ldots, \sigma(x_0) - 1\}$,

$$q_j(\sigma(x_0), \sigma(y)) = e^{\bar{U}_{\sigma(x_0)}-\bar{U}_{\sigma(y)}} \lambda_{\sigma(y)}(j),$$

and that for $k \geq \sigma(x_0)$,

$$\lambda_{\sigma(x_0)}(k) = e^{\bar{U}_{\sigma(x_0)}-\bar{U}_{\sigma(y)}} \lambda_{\sigma(y)}(k),$$

this quantity can be written as

$$e^{\bar{U}_{\sigma(x_0)}-\bar{U}_{\sigma(y)}} \sum_{k=\sigma(y)}^{J} \lambda_{\sigma(y)}(k) E_{y,k} \left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m}) 1_{\{L_{i^*,m}=l\}} \right].$$

Thus in this case as well, (16.10) equals the right side of (16.8). This completes the proof of (16.8), and hence we have proved the claim in (16.7).

Thus from the induction hypothesis and (16.7), we have that

$$
\begin{aligned}
& e^{-\bar{U}(x_0)} E_{x_0}\left[\sum_{j=1}^{N_{i^*+1}} \bar{h}(Z_{i^*+1,j}) 1_{\{L_{i^*+1,j}=l\}}\right] \\
&= e^{-\bar{U}(x_0)} E_{x_0}\left[e^{\bar{U}_{\sigma(Z_{0,1})}-\bar{U}_{\sigma(Z_{1,1})}} E_{Z_{1,1}}\left[\sum_{m=1}^{N_{i^*}} \bar{h}(Z_{i^*,m}) 1_{\{L_{i^*,m}=l\}}\right]\right] \\
&= e^{-\bar{U}(x_0)} E_{x_0}\left[e^{\bar{U}_{\sigma(Z_{0,1})}} E_{Z_{1,1}}\left[h(X_{i^*})\lambda_{\sigma(X_{i^*})}(l) 1_{\{M\geq i^*\}}\right]\right] \\
&= E_{x_0}\left[E_{Z_{1,1}}\left[h(X_{i^*})\lambda_{\sigma(X_{i^*})}(l) 1_{\{M\geq i^*\}}\right]\right] \\
&= E_{x_0}\left[E_{X_1}\left[h(X_{i^*})\lambda_{\sigma(X_{i^*})}(l) 1_{\{M\geq i^*\}}\right]\right] \\
&= E_{x_0}\left[h(X_{i^*+1})\lambda_{\sigma(X_{i^*+1})}(l) 1_{\{M\geq i^*+1\}}\right],
\end{aligned}
$$

where the third equality uses the fact that $Z_{1,1}$ and X_1 have the same distribution, and the last equality uses the Markov property of $\{X_i\}$. This completes the induction step, and thus the lemma follows. □

Proof (*of Theorem* 16.3) Since by Condition 16.1, $M < \infty$ a.s., we have

$$
\begin{aligned}
E_{x_0}\left[1_B(X_M)\right] &= E_{x_0}\left[\sum_{i=0}^{\infty} 1_B(X_i) 1_{\{M=i\}}\right] \\
&= \sum_{i=0}^{\infty} E_{x_0}\left[1_B(X_i) 1_{\{M=i\}}\right] \\
&= \sum_{i=0}^{\infty} e^{-\bar{U}(x_0)} E_{x_0}\left[\int_{\mathbb{R}^d} 1_B(x) e^{\bar{U}(x)} \bar{\delta}_{Z_i}(dx)\right] \\
&= E_{x_0}\left[\sum_{i=0}^{\infty} e^{-\bar{U}(x_0)} \int_{\mathbb{R}^d} 1_B(x) e^{\bar{U}(x)} \bar{\delta}_{Z_i}(dx)\right],
\end{aligned}
$$

where the second and fourth equalities use Tonelli's theorem, and the third uses Lemma 16.4 applied to $h = 1_B$ and the observation that $1_B(X_i) 1_{\{M=i\}} = 1_B(X_i) 1_{\{M\geq i\}}$. The result now follows on observing that the term on the last line equals $E_{x_0}[\gamma]$. □

16.3 Performance Measures

Recall that to derive a recurrence equation, we had to consider initializing distributions of the form $\mathscr{I} = \lambda_{\sigma(x_0)}$. In actual numerical implementation, it is always the case that $x_0 \in C_J$, which implies that all mass will be placed on $l = J$.

The performance of the algorithm depends on two factors: the second moment (and hence variance) of the estimator and the computational cost of each

simulation. To avoid discussion of any issues relating to the specific way the algorithm is implemented in practice, the computational cost is defined to be

$$w = \sum_{i=0}^{\infty} N_i, \tag{16.11}$$

where $N_i = \int_{\mathbb{R}^d} \bar{\delta}_{Z_i}(dx)$. Thus w is the sum of the lifetimes of all the particles simulated in the algorithm. In this section, formulas for both the second moment and computational cost are derived in terms of only the importance function and the underlying process. Throughout it is assumed that U is an importance function and Condition 16.1 is satisfied.

We begin by characterizing the mean of w.

Theorem 16.5 *Assume Condition 16.1. Then*

$$E_{x_0}[w] = e^{\bar{U}(x_0)} E_{x_0}\left[\sum_{i=0}^{M} e^{-\bar{U}(X_i)}\right].$$

Proof With the third equality due to Lemma 16.4 applied to $h(x) = e^{\bar{U}(x_0)-\bar{U}(x)}$, an application of Tonelli's theorem gives

$$\begin{aligned}
E_{x_0}[w] &= E_{x_0}\left[\sum_{i=0}^{\infty} \int_{\mathbb{R}^d} \bar{\delta}_{Z_i}(dx)\right] \\
&= \sum_{i=0}^{\infty} E_{x_0}\left[\int_{\mathbb{R}^d} \bar{\delta}_{Z_i}(dx)\right] \\
&= \sum_{i=0}^{\infty} e^{\bar{U}(x_0)} E_{x_0}\left[e^{-\bar{U}(X_i)} 1_{\{M \geq i\}}\right] \\
&= e^{\bar{U}(x_0)} E_{x_0}\left[\sum_{i=0}^{\infty} e^{-\bar{U}(X_i)} 1_{\{M \geq i\}}\right] \\
&= e^{\bar{U}(x_0)} E_{x_0}\left[\sum_{i=0}^{M} e^{-\bar{U}(X_i)}\right].
\end{aligned}$$

□

Next note that the following bounds hold for all U, all $0 \leq k \leq l < j \leq J, 0 \leq k \leq m < j \leq J$. Since $q_l(j,k)$ as defined in (16.3) equals $[R^{j-l} - R^{j-l-1}]$, it follows that

$$\begin{aligned}
(q_l(j,k))^2 - q_l(j,k) &\leq [R^{j-l} - R^{j-l-1}]^2, \\
q_l(j,k) q_m(j,k) &= [R^{j-l} - R^{j-l-1}][R^{j-m} - R^{j-m-1}].
\end{aligned} \tag{16.12}$$

We can now give bounds for the second moment of the splitting estimator. Let $\mathfrak{S}(\bar{U})$ denote $E_{x_0}[\gamma^2]$ when $\bar{U}$ is used to design the splitting scheme.

Theorem 16.6 *Assume Condition 16.1. Then for all $x_0 \in (A \cup B)^c$,*

$$\mathfrak{S}(\bar{U}) \le e^{-\bar{U}(x_0)} E_{x_0}\left[\sum_{i=1}^{M} e^{\bar{U}(X_{i-1})}\left[P_{X_i}\{X_M \in B\}\right]^2\right] + e^{-\bar{U}(x_0)} E_{x_0}\left[e^{\bar{U}(X_M)} 1_B(X_M)\right]. \tag{16.13}$$

Proof Recall that M is the first entry time of the set $A \cup B \subset D$. First consider the case in which there is $T < \infty$ such that $M \le T$ P_{x_0}-a.s. Let $W(x) \doteq e^{\bar{U}(x)} E_x[\gamma^2]$ with γ as in (16.4), and let $s(x, j; k)$, $k = 0, \ldots,$ denote iid sequences of random variables with the same distribution as γ, conditioned on $Z_{0,1} = x$ and $L_{0,1} = j$. Note that since the maximum possible number of particles is bounded, these random variables are bounded.

The proof is based on finding a recurrence equation for W. If $x_0 \notin A \cup B$, then there are two contributions to γ depending on the killing and/or splitting that takes place over the next time step. The first is due to future contributions if the particle stays within the support threshold, and the second occurs if new particles are generated $[\sigma(X_1) < \sigma(X_0)]$. To account for thresholds of both the existing particles and those that might be generated, let $Q_l(j,k)$ random vector defined in terms of $q_l(j,k)$ be random vectors defined by $Q_l(j,k) = q_l(j,k)$ for $j > l$ (i.e., these components are deterministic), and such that $Q_l(j,k)$ equals 1 for exactly one value of $j \le l \le J$ and 0 for remaining values, with the index chosen according to the initializing distribution λ_j. Recall that $q_l(j,k) = 0$ for all l if $k \ge j$. To abbreviate notation temporarily, let $\sigma_i = \sigma(Z_{i,1})$, $i = 0, 1$. Then

$$\begin{aligned} W(x_0) &= e^{\bar{U}(x_0)} E_{x_0}\left[\left(1_{\{L_{0,1} \ge \sigma_1\}} e^{\bar{U}(Z_{1,1}) - \bar{U}(Z_{0,1})} s(Z_{1,1}, L_{0,1}; 0) \right.\right. \\ &\qquad \left.\left. + 1_{\{\sigma_0 > \sigma_1\}}\left(\sum_{j=\sigma_1}^{\sigma_0 - 1} \sum_{m=1}^{q_j(\sigma_0,\sigma_1)} e^{\bar{U}(Z_{1,1}) - \bar{U}(Z_{0,1})} s(Z_{1,1}, j; m)\right)\right)^2\right] \\ &= e^{\bar{U}(x_0)} E_{x_0}\left[\left(\sum_{j=0}^{J} \sum_{m=1}^{Q_j(\sigma_0,\sigma_1)} e^{\bar{U}(Z_{1,1}) - \bar{U}(Z_{0,1})} s(Z_{1,1}, j; m)\right)^2\right]. \end{aligned}$$

We now use the following facts: $L_{0,1}$ has distribution $\lambda_{\sigma(X_0)}$; $Z_{1,1}$ has the same distribution (conditioned on $Z_{0,1} = X_0 = x_0$) as X_1; by the definition of $Q_l(j,k)$, for all j, k, l [see also (16.6) and (16.5)],

$$E_{x_0} Q_l(j,k) e^{\bar{U}_k - \bar{U}_j} = \lambda_k(l); \tag{16.14}$$

and that the future evolution of the algorithm is independent of the $Q_l(j,k)$. Now let σ_i denote $\sigma(X_i)$, $i = 0, 1$. Together with the expression just given for $W(x_0)$, these give

$$W(x_0) = e^{-\bar{U}(x_0)} E_{x_0}\left[\sum_{j,k=0}^{J} e^{2\bar{U}(X_1)} Q_j(\sigma_0,\sigma_1) Q_k(\sigma_0,\sigma_1) E_{X_1,j}[\gamma] E_{X_1,k}[\gamma]\right] \tag{16.15}$$

$$+ e^{-\bar{U}(x_0)} E_{x_0}\left[\sum_{j=0}^{J} e^{2\bar{U}(X_1)} Q_j(\sigma_0,\sigma_1)\left(E_{X_1,j}\left[\gamma^2\right] - \left(E_{X_1,j}[\gamma]\right)^2\right)\right].$$

We examine the various terms separately. Using (16.14) and $W(x) \doteq e^{\bar{U}(x)} E_x[\gamma^2]$,

$$\begin{aligned} & e^{-\bar{U}(x_0)} E_{x_0}\left[\sum_{j=0}^{J} e^{2\bar{U}(X_1)} Q_j(\sigma_0,\sigma_1) E_{X_1,j}\left[\gamma^2\right]\right] \\ &= E_{x_0}\left[e^{\bar{U}(X_1)} \sum_{j=0}^{J} e^{\bar{U}(X_1)-\bar{U}(X_0)} Q_j(\sigma_0,\sigma_1) E_{X_1,j}\left[\gamma^2\right]\right] \\ &= E_{x_0}\left[e^{\bar{U}(X_1)} \sum_{j=0}^{J} \lambda_{\sigma_1}(j) E_{X_1,j}\left[\gamma^2\right]\right] \\ &= E_{x_0}\left[W\left(X_1\right)\right]. \end{aligned} \tag{16.16}$$

If (16.16) is subtracted from the right side of (16.15), the remaining quantity is

$$\begin{aligned} & e^{\bar{U}(x_0)} E_{x_0}\left[\sum_{j=0}^{J}\sum_{l=0}^{J} e^{2\bar{U}(X_1)-2\bar{U}(X_0)} Q_j(\sigma_0,\sigma_1) Q_l(\sigma_0,\sigma_1) E_{X_1,j}\left[\gamma\right] E_{X_1,l}\left[\gamma\right]\right] \\ & - e^{\bar{U}(x_0)} E_{x_0}\left[\sum_{j=0}^{J} e^{2\bar{U}(X_1)-2\bar{U}(X_0)} Q_j(\sigma_0,\sigma_1)\left(E_{X_1,j}\left[\gamma\right]\right)^2\right]. \end{aligned} \tag{16.17}$$

The terms with both l and j at or above $\sigma(X_0)$ contribute nothing to this expression. Indeed, $Q_j(\sigma_0,\sigma_1)$ is equal to 1 for exactly one j and to 0 for all remaining j that are greater than $\sigma(X_0) - 1$. Hence the corresponding terms in the double and single sums cancel. Also, this is the only possibility when $\sigma(X_1) \geq \sigma(X_0)$, and so we restrict to $\sigma(X_1) < \sigma(X_0)$, and use that $Q_j(\sigma_0,\sigma_1) = 0$ for $j < \sigma(X_1)$ when this is the case. Dropping terms that contribute nothing, we decompose the double sum as

$$\sum_{j=\sigma(X_1)}^{\sigma(X_0)-1}\sum_{l=\sigma(X_1)}^{\sigma(X_0)-1} + 2\sum_{j=\sigma(X_0)}^{J}\sum_{l=\sigma(X_1)}^{\sigma(X_0)-1}.$$

Using (16.14), we get the following upper bound for expression (16.17):

$$\begin{aligned}
&e^{\bar{U}(x_0)}E_{x_0}\left[1_{\{\sigma(X_1)<\sigma(X_0)\}}\left(\sum_{j=\sigma(X_1)}^{\sigma(X_0)-1}\lambda_{\sigma(X_1)}(j)E_{X_1,j}[\gamma]\right)^2\right]\\
&+2e^{\bar{U}(x_0)}E_{x_0}\left[1_{\{\sigma(X_1)<\sigma(X_0)\}}\left(\sum_{j=\sigma(X_0)}^{J}\lambda_{\sigma(X_1)}(j)E_{X_1,j}[\gamma]\right)\right.\\
&\qquad\left.\times\left(\sum_{l=\sigma(X_1)}^{\sigma(X_0)-1}\lambda_{\sigma(X_1)}(l)E_{X_1,l}[\gamma]\right)\right]\\
&\le e^{\bar{U}(x_0)}E_{x_0}\left[1_{\{\sigma(X_1)<\sigma(X_0)\}}\left(\sum_{j=\sigma(X_1)}^{J}\lambda_{\sigma(X_1)}(j)E_{X_1,j}[\gamma]\right)^2\right]. \qquad (16.18)
\end{aligned}$$

We now combine (16.15), (16.16), (16.18), and Theorem 16.3 to get that for $x_0 \notin A \cup B$,

$$W(x_0) \le e^{\bar{U}(x_0)}E_{x_0}\left[1_{\{\sigma(X_1)<\sigma(X_0)\}}\left(E_{X_1}[1_B(X_M)]\right)^2\right] + E_{x_0}[W(X_1)].$$

Since all functions involved are bounded and nonnegative, it follows that the sequence

$$\Sigma_i \doteq W(X_{i\wedge M}) + \sum_{j=1}^{i\wedge M}\left\{e^{\bar{U}(X_{j-1})}\left(E_{X_j}[1_B(X_M)]\right)^2\right\}$$

defined for $i \in \{0,\dots,T\}$ is a submartingale. Thus, using $W(X_{T\wedge M}) = e^{\bar{U}(X_M)}$ $1_B(X_M)$ and that $1_B(X_k) = 0$ if $k < M$, we have

$$\begin{aligned}
e^{\bar{U}(x_0)}\mathfrak{S}(\bar{U}) &= W(x_0)\\
&= \Sigma_0\\
&\le E_{x_0}[\Sigma_T]\\
&= E_{x_0}\left[\sum_{i=1}^{M}e^{\bar{U}(X_{i-1})}\left(E_{X_i}[1_B(X_M)]\right)^2\right] + E_{x_0}\left[e^{\bar{U}(X_M)}1_B(X_M)\right],
\end{aligned}$$

which is the same as (16.13).

We next remove the restriction that $M \le T$ for some constant $T < \infty$. We add time as a state variable [i.e., work with the process (X_i, i)], and consider the analogous

estimation problem in which the stopping set is $(A \cup B) \times \{T\}$ (i.e., we stop if either X_i enters $A \cup B$ or $i = T$). Then γ_T defined in an analogous manner is an unbiased estimator of $E_{x_0}\left[1_B(X_M)1_{\{M\le T\}}\right]$, and by the previous result for bounded stopping times,

$$e^{\bar{U}(x_0)}E_{x_0}[(\gamma_T)^2] \tag{16.19}$$
$$\le E_{x_0}\left[\sum_{i=1}^{M\wedge T} e^{\bar{U}(X_{i-1})}\left(E_{X_i}\left[1_B(X_M)1_{\{M\le T\}}\right]\right)^2\right] + E_{x_0}\left[e^{\bar{U}(X_M)}1_B(X_M)1_{\{M\le T\}}\right].$$

Also note that

$$\gamma_T = e^{-\bar{U}(x_0)}\sum_{i=0}^{T}\int_{\mathbb{R}^d} e^{\bar{U}(y)}1_B(y)\bar{\delta}_{Z_i}(dy)$$

and $\gamma_T \uparrow \gamma$ a.s. By the monotone convergence theorem,

$$E_{x_0}\left[(\gamma_T)^2\right] \to \mathfrak{S}(\bar{U}).$$

Using this in (16.19), the nonnegativity of 1_B, and the monotone convergence theorem a second time gives (16.13) without the restriction on M. □

The following result gives a lower bound on the second moment of the estimator, complementing the upper bound in Theorem 16.6.

Theorem 16.7 *Assume Condition 16.1. Then*

$$\mathfrak{S}(\bar{U}) \ge e^{-\bar{U}(x_0)}E_{x_0}\left[e^{\bar{U}(X_M)}1_B(X_M)\right].$$

Proof From the nonnegativity of 1_B, (16.15), and (16.16), it follows that $W(x_0) \ge E_{x_0}[W(X_1)]$. From the Markov property of $\{X_i\}$, it follows that $\Sigma_i \doteq W(X_{i\wedge M})$ is a supermartingale, and in particular,

$$E_{x_0}[W(X_{M\wedge i})] \le W(x_0).$$

The definition $W \doteq e^{\bar{U}(x)}E_x[\gamma^2]$ and its nonnegativity then give

$$E_{x_0}\left[W(X_M)1_{\{M\le i\}}\right] \le E_{x_0}[W(X_{M\wedge i})] \le e^{\bar{U}(x_0)}\mathfrak{S}(\bar{U}).$$

Since $W(x) = e^{\bar{U}(x)}1_B(x)$ for $x \in A \cup B$, the last display gives

$$E_{x_0}\left[e^{\bar{U}(X_M)}1_B(X_M)1_{\{M\le i\}}\right] \le e^{\bar{U}(x_0)}\mathfrak{S}(\bar{U}).$$

The result now follows on sending $i \to \infty$ and using the monotone convergence theorem. □

16.4 Design and Asymptotic Analysis of Splitting Schemes

Thus far we have considered only the problem of estimating a single probability of the form (16.1). Now we shall turn to the problem of estimating a sequence of such probabilities

$$P_{x_0}\{X^n_{M^n} \in B\}, \tag{16.20}$$

$n \in \mathbb{N}$, where $M^n \doteq \inf\{i : X^n_i \in A \cup B\}$ and $\{X^n_i\}_{i\in\mathbb{N}_0}$ is a Markov chain for each n that satisfies a large deviation principle as $n \to \infty$ (see Condition 16.9). We recall that by assumption, A is open, B is closed, and $(A \cup B)^c$ is bounded. With $\mathfrak{S}^n(\bar{U})$ denoting the second moment $E_{x_0}\left[(\gamma^n)^2\right]$, the asymptotic performance will be evaluated using the following measure of work-normalized error:

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{\mathfrak{S}^n(\bar{U}) E_{x_0}[w^n]}{\left[P_{x_0}\{X^n_{M^n} \in B\}\right]^2}, \tag{16.21}$$

where γ^n is the splitting-based estimator for (16.20), and w^n is its computational cost, which was defined in the nonasymptotic setting in (16.11). (Such a weighted performance measure is not needed for importance sampling, since the cost per sample is essentially independent of the subsolution.)

Suppose that $-(1/n)\log P_{x_0}\{X^n_{M^n} \in B\} \to V(x_0)$ as $n \to \infty$. Jensen's inequality as discussed in Chap. 14 shows that the best possible value of (16.21) is zero, and this occurs only when the work grows subexponentially and the second moment $\mathfrak{S}^n(\bar{U})$ decays at rate $2V(x_0)$. Bounds on the asymptotic behavior of the work-normalized error will be obtained using Theorems 16.5–16.7 and are stated in Theorem 16.15, Corollary 16.16, and Theorem 16.18.

Remark 16.8 As in Chap. 15, the theoretical bounds on performance are given for the case of a fixed initial condition x_0. However, all results are easily generalized to the case of varying initial conditions x_n that converge to x_0 as $n \to \infty$. This generalization is useful for systems with discrete state spaces, such as queueing networks.

The theory presented in this section will require some fairly standard assumptions on the stability and large deviation behavior of $\left\{X^n_i\right\}$, and also some regularity properties on A and B that are qualitatively similar to assumptions made in Chap. 15 (e.g., Condition 15.9). For example, we will want to know that $\tau^n \doteq M^n/n$ can essentially be taken as bounded, in the sense that there is some $T < \infty$ such that the event $\tau^n > T$ is unimportant as far as the large deviation asymptotics are concerned. This is an important qualitative assumption, and it is related to stability properties of the law of large numbers limit processes obtained when $n \to \infty$.

We define continuous time stochastic processes as usual by setting $X^n(t) = X^n_i$ for $t = i/n$ and by piecewise linear interpolation for $t \in [i/n, (i+1)/n)$. Throughout, we assume that $x_0 \in (\bar{A} \cup B)^c$. The following condition will be needed to establish

a limit for the second moment; it is not needed if one wants to give just upper bounds on the second moment.

Condition 16.9 *For every $T \in (0, \infty)$, the sequence $\{X^n\}_{n\in\mathbb{N}}$ satisfies a large deviation principle on $\mathscr{C}\left([0, T] : \mathbb{R}^d\right)$ that is uniform with respect to the initial condition in compacts sets. The rate function is of the form*

$$I_T(\phi) \doteq \int_0^T L(\phi(s), \dot{\phi}(s))ds$$

if $\phi \in \mathscr{C}\left([0, T] : \mathbb{R}^d\right)$ is absolutely continuous with $\phi(0) = x_0$ and ∞ otherwise, where L is a nonnegative measurable function.

As remarked above, the conditions we use beyond the LDP can be partitioned into "stability" and "controllability" type conditions. We give two conditions that will be sufficient (but not necessary) for what follows. Moreover, the sufficient conditions we give will by themselves cover many interesting problems. The stability condition (Condition 16.10) will imply that the algorithm is practical in that the tails of the hitting times are controlled, and also that the escape time problem can be approximated using estimates over finite time intervals. The condition we refer to as "controllability" (Condition 16.11) is needed to establish limits rather than just bounds, and is analogous to the additional conditions that would have been required in Chap. 15 as noted in Remark 15.16.

We will assume the following condition, which is the same as Condition 15.9 in Chap. 15.

Condition 16.10 *There exist $c > 0$, $T_0 \in (0, \infty)$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $T < \infty$, and $x \in D$,*

$$P_x\{\tau^n > T\} \leq \exp\{-cn(T - T_0)\}.$$

Note that Condition 16.10 implies that

$$\limsup_{T\to\infty} \limsup_{n\to\infty} \sup_{x\in D} \frac{1}{n} \log P_x\{\tau^n > T\} = -\infty. \tag{16.22}$$

Condition 16.10 would not hold if there were two attractors for the zero-cost trajectories, $A \cup B$ contains one of the attractors but not the other, and the process starts in the domain of attraction of the stable point that is not in $A \cup B$.

Condition 16.11 *Suppose we are given absolutely continuous ϕ satisfying $\phi(0) = x_0 \notin \bar{A} \cup B$, $\phi(t) \notin A \cup B^\circ$ for $t \in [0, T)$, and $\phi(T) \in B$ for some $T < \infty$. Then given $\gamma > 0$, there exist absolutely continuous ϕ^*, $T^* < \infty$, and $\tau^* < T^*$ such that $\phi^*(0) = x_0$, $\phi^*(t) \notin \bar{A} \cup B$ for $t \in [0, \tau^*)$, $\phi^*(t) \in B^\circ$ for $t \in (\tau^*, T^*]$, and such that*

$$\int_0^{T^*} L(\phi^*(r), \dot{\phi}^*(r))dr \le \int_0^T L(\phi(r), \dot{\phi}(r))dr + \gamma, \ \ \|\phi(T) - \phi^*(\tau^*)\| \le \gamma.$$

One can consider this a controllability-type condition. It says that given a trajectory ϕ that enters $\bar{A} \cup B$ but not $A \cup B^\circ$ and finally enters B° at T, one can find a trajectory with almost the same cost that avoids $\bar{A} \cup B$ till τ^*, at which time it enters B° near $\phi(T)$. One can establish more concrete conditions that imply this condition, such as assuming that $L(x, \beta)$ is continuous, bounded on each compact subset of $\mathbb{R}^d \times \mathbb{R}^d$, and assuming regularity properties for the boundaries of A and B.

We next give a definition of subsolution appropriate to this problem, but phrased directly in terms of the calculus of variations problem. The definition via calculus of variations is somewhat more to the point of what is required and is used in the proofs. For $y \in (A \cup B^\circ)^c$ and $T \in (0, \infty)$, define

$$K_{y,T} \doteq \left\{\phi \in \mathscr{AC}([0, T] : \mathbb{R}^d) : \phi(0) = y; \phi(s) \notin A \cup B^\circ, s \in (0, T), \phi(T) \in B\right\}. \tag{16.23}$$

Definition 16.12 A continuous function $\bar{V} : \mathbb{R}^d \to \mathbb{R}$ is a subsolution if it is bounded from below,

$$\bar{V}(y) \le \inf_{\phi \in K_{y,T}, T < \infty} \left[\int_0^T L(\phi(s), \dot{\phi}(s))ds + \bar{V}(\phi(T))\right] \tag{16.24}$$

for all $y \in (A \cup B^\circ)^c$, and $\bar{V}(z) \le 0$ for $z \in B$.

Remark 16.13 (*Relations between notions of subsolution I*) Suppose that $\bar{V}$ is a subsolution in the sense of Definitions 14.4 or 14.5 that is bounded from below, and to simplify the discussion, assume also that $\mathbb{H}(x, \alpha)$ is continuous. Then we claim that $\bar{V}$ is a subsolution in the sense of Definition 16.12. Consider the case of Definition 14.4, and for $y \in (A \cup B^\circ)^c$, suppose $\phi \in K_{y,T}$. Since $\bar{V}$ is continuously differentiable, the definition $\mathbb{H}(x, \alpha) \doteq \inf_{\beta \in \mathbb{R}^d} [\langle \alpha, \beta \rangle + L(x, \beta)]$ implies

$$\left\langle D\bar{V}(\phi(s)), \dot{\phi}(s) \right\rangle + L(\phi(s), \dot{\phi}(s)) \ge 0$$

for a.e. $s \in [0, T]$. Integrating gives

$$\bar{V}(y) \le \left[\int_0^T L(\phi(s), \dot{\phi}(s))ds + \bar{V}(\phi(T))\right].$$

Since $\phi \in K_{y,T}$ is arbitrary, and $\bar{V}(z) \le 0$ for $z \in B$ is part of Definition 14.4, the claim follows. For the case of piecewise classical subsolutions it is enough to note that the mollification (14.16) produces a classical-sense subsolution $\bar{V}^\delta$, and the claim follows by taking the limit $\delta \downarrow 0$. Note that one does not need to use the mollified subsolution for the design of the splitting scheme, but can instead use the potentially

nonsmooth limit. It is also worth noting that the continuity of $\mathbb{H}$ is not used in any essential way, so that the analogous claim holds for problems involving discontinuous statistics, such as queueing networks.

Remark 16.14 (*Relations between notions of subsolution II)* The set of trajectories $K_{y,T}$ in (16.23) differs from $C_{y,T}$ introduced in Chap. 14 in that trajectories are excluded from B for $s \in (0, T)$ for $C_{y,T}$, but only from B° for $K_{y,T}$. The reason for the difference is the slightly different way in which the subsolution property is used in the cases of importance sampling and splitting. However, under the conditions we assume, to obtain a limit as in Theorem 16.15 one could use $C_{y,T}$ in Definition 16.12 instead. Indeed, since $C_{y,T} \subset K_{y,T}$, one has only to check that the defining inequality (16.24) holds for $\phi \in K_{y,T}$ for all T if it holds for all $\phi \in C_{y,S}$, $y \in (A \cup B^\circ)^c$, and $S \in (0, \infty)$. But this is easy to check under Condition 16.11. Note also that the two definitions are equivalent without reference to Condition 16.11 if $\bar{V}$ is constant on B, simply because $L \geq 0$.

Note that $\bar{V}$ is never greater than the solution to the calculus of variations problem, which is defined by $V(z) = 0$ for $z \in B$, $V(z) = \infty$ for $z \in A$, and

$$V(x) = \inf_{\phi \in K_{x,T}; T < \infty} \left[\int_0^T L(\phi(s), \dot{\phi}(s)) ds \right] \tag{16.25}$$

for $x \in (A \cup B)^c$. Given a subsolution, the corresponding splitting scheme is defined as follows. Thresholds are defined in terms of the levels

$$\bar{V}(x_0), \bar{V}(x_0) - (\log R)/n, \bar{V}(x_0) - (2 \log R)/n, \ldots .$$

Let J^n be the smallest number such that $\bar{V}(x) \geq \bar{V}(x_0) - (J^n \log R)/n$ for all $x \in D$, so that there are no more than J^n thresholds. Then we define $C^n_{J^n} \doteq D$, $C^n_{-1} \doteq \emptyset$, and

$$C^n_{J^n - j} \doteq \left\{ x : \bar{V}(x) \leq \bar{V}(x_0) - (j \log R)/n \right\}, \quad j = 1, \ldots, J^n. \tag{16.26}$$

Recall $R = e^{\Delta}$ and define $\Delta^n \doteq \Delta/n$. Also define a sequence $\{\bar{U}^n\}$ according to

$$\bar{U}^n(x) \doteq (J^n \log R)/n - (j \log R)/n \text{ for } x \in C^n_{J^n - j} \backslash C^n_{J^n - j - 1}, \quad j = 0, 1, \ldots, J^n.$$

Note that whenever $y_n \to y$, $\bar{U}^n(y_n) - \bar{U}^n(x_0) \to \bar{V}(y) \wedge \bar{V}(x_0) - \bar{V}(x_0)$. Note also that if $\bar{V}$ is a subsolution in the sense of Definition 16.12, then so is $\bar{V}(\cdot) \wedge \bar{V}(x_0)$. To simplify notation, we assume without loss that $\bar{V}(x) \leq \bar{V}(x_0)$ for all $x \in D$, and therefore for $x \in D$,

$$\left| (\bar{U}^n(x_0) - \bar{U}^n(x)) - (\bar{V}(x_0) - \bar{V}(x)) \right| \leq \frac{\log R}{n}. \tag{16.27}$$

In particular, $\bar{U}^n(x_0) - \bar{U}^n(x) \to \bar{V}(x_0) - \bar{V}(x)$ for all $x \in D$. We now apply Theorems 16.6 and 16.7, with Δ replaced by Δ^n and $\bar{U}$ replaced by $n\bar{U}^n$, to the

Markov chain $\{X_i^n\}$. Following the same notation as in Chaps. 14 and 15, we denote the second moment of the estimator $E_{x_0}[(\gamma^n)^2]$ by $\mathfrak{S}^n(\bar{V})$. The corresponding initializing distribution λ_k^n is defined as $\lambda_k^n(l) = q_l^n / R^{J^n - k}$, with q_l^n defined by (16.5) but with J replaced by J^n. Theorems 16.6 and 16.7 then say that

$$e^{-n\bar{U}^n(x_0)} E_{x_0}\left[\sum_{i=1}^{M^n} e^{n\bar{U}^n(X_{i-1}^n)} \left(E_{X_i^n}\left[1_B(X_{M^n}^n)\right]\right)^2\right] + e^{-n\bar{U}^n(x_0)} E_{x_0}\left[e^{n\bar{U}^n(X_{M^n}^n)} 1_B(X_{M^n}^n)\right] \geq \mathfrak{S}^n(\bar{V}) \tag{16.28}$$

and

$$\mathfrak{S}^n(\bar{V}) \geq e^{-n\bar{U}^n(x_0)} E_{x_0}\left[e^{n\bar{U}^n(X_{M^n}^n)} 1_B(X_{M^n}^n)\right],$$

where $\gamma^n = e^{-n\bar{U}^n(x_0)} \sum_{i=0}^{\infty} \int_{\mathbb{R}^d} 1_B(y) e^{-n\bar{U}^n(y)} \bar{\delta}_{Z_i^n}(dy)$ is the estimator of (16.20) based on the importance function $\bar{U}^n$.

Theorem 16.15 below describes the asymptotic performance of the splitting scheme based on importance functions $\{\bar{U}^n\}$. As a consequence of this result, in Corollary 16.16, we will see that the decay rate

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathfrak{S}^n(\bar{V})$$

is bounded below by $V(x_0) + \bar{V}(x_0)$, where $V(x_0)$ is defined in (16.25), which is the same as the decay rate for an importance sampling scheme based on the same subsolution if it is sufficiently regular (see Theorem 15.10). We recall from (14.5) that when the large deviation limit holds, the decay rate of any unbiased splitting scheme is bounded above by $2V(x_0)$. In particular, if $\bar{V}(x_0) = V(x_0)$, we get the best possible decay rate $2V(x_0)$. Finally, in Theorem 16.18 below, we will show that

$$\lim_{n\to\infty} \frac{1}{n} \log E_{x_0}\left[w^n\right] = 0.$$

Thus the work associated with such a scheme grows subexponentially, and consequently, the decay rate of the work-normalized error is zero, which is the best possible rate.

It is easily checked that if $\bar{V}$ is not a subsolution, then at points where the subsolution property fails, the branching is supercritical, and hence in this case there exists $y \in D$ such that if $y_n \to y$, then

$$\liminf_{n\to\infty} \frac{1}{n} \log E_{y_n}\left[w^n\right] > 0.$$

It follows that importance functions that are not obtained from subsolutions should not be used to design schemes, since it is possible that the computational costs of such schemes will grow exponentially.

Recall that $x_0 \notin \bar{A} \cup B$, $\bar{V}$ is a subsolution as in Definition 16.12, and that as noted previously, we can assume $\bar{V}(x) \le \bar{V}(x_0)$ for all $x \in D$. Recall also the definition of $K_{y,T}$ in (16.23).

Theorem 16.15 *Assume Conditions 16.9–16.11. Then for $x_0 \notin \bar{A} \cup B$,*

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathfrak{S}^n(\bar{V}) = \inf_{\phi \in K_{x_0,T}, T<\infty} \left[\int_0^T L(\phi(r), \dot{\phi}(r))dr + \bar{V}(\phi(0)) - \bar{V}(\phi(T)) \right]. \tag{16.29}$$

Proof We first consider the lower bound

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathfrak{S}^n(\bar{V}) \ge \inf_{\phi \in K_{x_0,T}, T<\infty} \left[\int_0^T L(\phi(r), \dot{\phi}(r))dr + \bar{V}(\phi(0)) - \bar{V}(\phi(T)) \right], \tag{16.30}$$

which is based on (16.28). While there are two terms in (16.28), the second term can be treated in a similar manner as the first term, and so we focus on the first. This term is

$$e^{-n\bar{U}^n(x_0)} E_{x_0} \left[\sum_{i=1}^{M^n} e^{n\bar{U}^n(X^n_{i-1})} 1_B \left(X^{1,i,n}_{M^{1,i,n}} \right) 1_B \left(X^{2,i,n}_{M^{2,i,n}} \right) \right], \tag{16.31}$$

where $\{X^{k,i,n}_j\}_{j\ge i}$, $k = 1, 2$, are (conditionally) independent copies of $\{X^n_j\}_{j\ge i}$ that start at X^n_i at $j = i$, and $M^{k,i,n}$ are the corresponding escape times.

We claim that instead of the large deviation asymptotics of (16.31), it suffices to consider the large deviation asymptotics of

$$e^{-n\bar{U}^n(x_0)} E_{x_0} \left[\sum_{i=1}^{\lfloor nT \rfloor} 1_{\{M^n \ge i\}} e^{n\bar{U}^n(X^n_{i-1})} 1_B \left(X^{1,i,n}_{M^{1,i,n}\wedge\lfloor nT \rfloor} \right) 1_B \left(X^{2,i,n}_{M^{2,i,n}\wedge\lfloor nT \rfloor} \right) \right] \tag{16.32}$$

for some fixed and finite T. Assuming the claim, observe that there are no more than order-n terms in the expected value, and it suffices to obtain the desired bound on each of these terms. Let i_n index such a term. In obtaining a bound, we can assume without loss that i_n/n will converge to some limit $t \in [0, T]$, and to simplify notation, we write i for i_n. We first show that

$$\liminf_{n\to\infty} -\frac{1}{n} \log e^{-n\bar{U}^n(x_0)} E_{x_0} \left[e^{n\bar{U}^n(X^n_{i-1})} 1_{\{M^n \ge i\}} 1_B \left(X^{1,i,n}_{M^{1,i,n}\wedge\lfloor nT \rfloor} \right) 1_B \left(X^{2,i,n}_{M^{2,i,n}\wedge\lfloor nT \rfloor} \right) \right] \ge \inf \left[\int_0^s L(\phi(r), \dot{\phi}(r))dr + \bar{V}(x_0) - \bar{V}(\phi(s)) \right], \tag{16.33}$$

where the infimum is over all absolutely continuous ϕ such that $\phi(0) = x_0$ and $\phi(r) \notin A \cup B^\circ$ for $r \in (0, s)$ and $\phi(s) \in B$ for some $s \geq t$. Let $\hat{X}^n(t)$ be the continuous time trajectory that interpolates X_j^n up till i, and thereafter is a two-component process that interpolates $X_j^{1,i,n}$ and $X_j^{2,i,n}$ up until $\lfloor nT \rfloor$. It is straightforward, using the Markov property and the uniformity of the large deviation estimates with respect to initial conditions that is assumed in Condition 16.9, to check that $\{\hat{X}^n\}$ satisfies a large deviation property, and that the rate function (with obvious notation for a trajectory η that branches at time t into η^1 and η^2) is

$$\int_0^t L(\eta(r), \dot{\eta}(r))dr + \int_t^T L(\eta^1(r), \dot{\eta}^1(r))dr + \int_t^T L(\eta^2(r), \dot{\eta}^2(r))dr.$$

Since $\bar{V}$ is continuous and B is closed, we obtain the lower bound

$$\inf \left[\int_0^t L(\eta(r), \dot{\eta}(r))dr + \int_t^T L(\eta^1(r), \dot{\eta}^1(r))dr + \int_t^T L(\eta^2(r), \dot{\eta}^2(r))dr + \bar{V}(x_0) - \bar{V}(\eta(t)) \right]$$

for the left side of (16.33), where the infimum is over all η such that $\eta(0) = x_0$ and $\eta(r) \notin A \cup B^\circ$ for $r \in (0, t]$ and $\eta^k, k = 1, 2$ such that $\eta^k(t) = \eta(t)$ and $\eta^k(r) \notin A \cup B^\circ$ for $r \in [t, s^k]$, $s^k \in [t, T]$, $\eta^k(s^k) \in B$. Without loss we can assume that the cost is zero after s^k and that $\eta^1 = \eta^2$ (which we relabel as η, and s^k as s). By the subsolution property,

$$\bar{V}(\eta(t)) \leq \int_t^s L(\eta(r), \dot{\eta}(r))dr + \bar{V}(\eta(s)),$$

which gives (16.33).

We now prove the claim. It remains to show that (16.31) has the same large deviation asymptotics as (16.32). To justify bounding the other random times by $\lfloor nT \rfloor$, we need to show that

$$\limsup_{T\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log e^{-n\bar{U}^n(x_0)} E_{x_0} \left[\sum_{i=1}^{M^n \wedge \lfloor nT \rfloor} e^{n\bar{U}^n(X_{i-1}^n)} \left(1_{\{\tau^{1,i,n} \geq nT\}} + 1_{\{\tau^{2,i,n} \geq nT\}} \right) \right] = -\infty. \tag{16.34}$$

However, using (16.27), the expected value is bounded above by

$$2Re^{n2\|\bar{V}\|_\infty} \sum_{i=1}^{\lfloor nT \rfloor} P_{x_0} \left\{ \tau^{1,i,n} \geq nT \right\},$$

and thus (16.34), and therefore the claim, follows from Condition 16.10 [see (16.22)].

We now prove the upper bound

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathfrak{S}^n(\bar{V}) \leq \inf_{\phi\in K_{x_0,T},T<\infty}\left[\int_0^T L(\phi(r),\dot{\phi}(r))dr + \bar{V}(\phi(0)) - \bar{V}(\phi(T))\right]. \tag{16.35}$$

Fix $\varepsilon \in (0,1)$, let Γ denote the right-hand side of (16.35), and using Condition 16.11 choose an absolutely continuous ϕ, T and $\tau \in (0,T)$ such that $\phi(0) = x_0, \phi(\tau) \in B$, $\phi(t) \notin \bar{A} \cup B$ for all $t \in (0,\tau)$, $\phi(t) \in B^\circ$ for all $t \in (\tau, T]$, and

$$\int_0^T L(\phi(r),\dot{\phi}(r))dr + \bar{V}(\phi(0)) - \bar{V}(\phi(\tau)) \leq \Gamma + \varepsilon. \tag{16.36}$$

Let $\delta > 0$ satisfy $\|\phi(r) - x\| \geq \delta$ if $r \in [0, \tau-\delta)$ and $x \in \bar{A} \cup B$ and $\|\phi(r) - x\| \geq \delta$ if $r \in [\tau+\delta, T]$ and $x \notin B$. Let also $\Xi^n \doteq \{\sup_{0\leq t\leq T} \|X^n(t) - \phi(t)\| < \delta\}$. Then using the large deviation lower bound for the third inequality and (16.36) for the fourth, we obtain

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{n}\log \mathfrak{S}^n(\bar{V}) \\
&\geq \liminf_{n\to\infty} \frac{1}{n}\log e^{-n\bar{U}^n(x_0)} E_{x_0}\left[e^{n\bar{U}^n(X^n_{M^n})} 1_B(X^n_{M^n})\right] \\
&\geq \liminf_{n\to\infty} \frac{1}{n}\log e^{-n\bar{U}^n(x_0)} E_{x_0}\left[e^{n\bar{U}^n(X^n_{M^n})} 1_B(X^n_{M^n}) 1_{\Xi^n}\right] \\
&\geq \inf_{y\in C_\delta} \bar{V}(y) - \bar{V}(x_0) + \liminf_{n\to\infty} \frac{1}{n}\log P_{x_0}\{\Xi^n\} \\
&\geq \inf_{y\in C_\delta} \bar{V}(y) - \bar{V}(\phi(\tau)) - \Gamma - \varepsilon,
\end{aligned}$$

where $C_\delta \doteq \{y : \|y - \phi(t)\| < \delta \text{ for some } t \in [\tau-\delta, \tau+\delta]\}$. Since $\bar{V}$ and ϕ are continuous, we have $\inf_{y\in C_\delta} \bar{V}(y) - \bar{V}(\phi(\tau)) \to 0$ as $\delta \to 0$. Since $\varepsilon > 0$ and $\delta > 0$ are arbitrary, this proves the upper bound in (16.35). □

Corollary 16.16 *Under same conditions and with the same notation as in Theorem 16.15,*

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathfrak{S}^n(\bar{V}) \geq V(x_0) + \bar{V}(x_0).$$

In particular, if $\bar{V}(x_0) = V(x_0)$, then

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathfrak{S}^n(\bar{V}) = 2V(x_0).$$

Proof Recall the set $K_{x_0,T}$ introduced in (16.23) and consider any $\phi \in K_{x_0,T}$ with $T \in (0, \infty)$. Since $\bar{V}$ is a subsolution, it follows that $\bar{V}(z) \le 0$ for $z \in B$, and so

$$\int_0^T L(\phi(r), \dot{\phi}(r))dr + \bar{V}(\phi(0)) - \bar{V}(\phi(T)) \ge \int_0^T L(\phi(r), \dot{\phi}(r))dr + \bar{V}(x_0).$$

Taking the infimum over all $\phi \in K_{x_0,T}$ and $T \in (0, \infty)$, we have from (16.25) and (16.29) that

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathfrak{S}^n(\bar{V}) \ge V(x_0) + \bar{V}(x_0),$$

proving the first statement in the corollary. On the other hand, as was argued previously,

$$\limsup_{n\to\infty} -\frac{1}{n} \log \mathfrak{S}^n(\bar{V}) \le 2V(x_0).$$

The second statement in the corollary follows. □

Remark 16.17 An examination of the proof shows that the greatest contribution to the second moment of the estimator is from the correlation of particles that make it to B and whose last common ancestor is located in one of the thresholds close to B.

Finally, we now show that the work associated with the splitting scheme based on the sequence $\{\bar{U}^n\}$ grows subexponentially.

Theorem 16.18 *Under same conditions and with the same notation as in Theorem 16.15,*

$$\lim_{n\to\infty} \frac{1}{n} \log E_{x_0}\left[w^n\right] = 0,$$

where w^n is defined by the right side of (16.11) replacing N_i by $N_i^n \doteq \int_D \bar{\delta}_{Z_i^n}(dx)$.

Proof We know from Theorem 16.5 that

$$E_{x_0}\left[w^n\right] = e^{n\bar{U}^n(x_0)} E_{x_0}\left[\sum_{i=0}^{M^n} e^{-n\bar{U}^n(X_i^n)}\right].$$

Exactly as in the proof of (16.30), it follows that the large deviation asymptotics of $E_{x_0}[w^n]$ are the same as those of

$$e^{n\bar{U}^n(x_0)} E_{x_0}\left[\sum_{i=0}^{M^n \wedge \lfloor nT \rfloor} e^{-n\bar{U}^n(X_i^n)}\right]$$

for some sufficiently large but finite T. The convergence $\bar{U}^n(y) - \bar{U}^n(x_0) \to \bar{V}(y) - \bar{V}(x_0)$ and the same line of argument as in the proof of (16.30) show that

$$\limsup_{n\to\infty} \frac{1}{n} \log e^{n\bar{U}^n(x_0)} E_{x_0}\left[\sum_{i=0}^{M^n \wedge \lfloor nT \rfloor} e^{-n\bar{U}^n(X_i^n)}\right]$$
$$\leq -\inf_{\phi \in K_{x_0,T}, T<\infty}\left[\int_0^T L(\phi(s), \dot{\phi}(s))ds - (\bar{V}(x_0) - \bar{V}(\phi(T)))\right].$$

By the subsolution property (Definition 16.12), the quantity to be minimized is always nonnegative, and so the upper bound follows. Since $E_{x_0}[w^n] \geq 1$ for all n, the lower bound is automatic, which completes the proof. □

Remark 16.19 Although the subsolution property implies a type of stability as asserted in Theorem 16.18, it could allow for polynomial growth of the number of particles. If in practice one observes that a large number of particles make it to B in the course of simulating a single sample, then one can consider the use of a *strict* subsolution, i.e., a function $\bar{V}$ that satisfies the boundary conditions and

$$\bar{V}(y) \leq \inf_{\phi \in K_{y,T}, T<\infty}\left[\int_0^T [L(\phi(s), \dot{\phi}(s)) - \varepsilon]ds + \bar{V}(\phi(T))\right]$$

for some $\varepsilon > 0$. Because the value of $\bar{V}(x_0)$ is lowered slightly, there will be a slight increase in the second moment of the estimator. However, the strict inequality provides stronger control, and indeed, the expected number of particles and moments of the number of particles are bounded uniformly in n. See [76] for further discussion and examples. If phrased in terms of $\mathbb{H}$ as in Remark 16.14, the strict subsolution property means that $\mathbb{H}(x, D\bar{V}(x)) \geq \varepsilon$ for $x \in (A \cup B^\circ)^c$.

16.5 Splitting for Finite-Time Problems

By adding time as a state variable, finite-time problems such as those discussed in the context of importance sampling in Sect. 15.2 can also be be put into the RESTART framework. Thus the process $\{X_i^n\}$ is replaced by $\{(X_i^n, t_i^n)\}$, where

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n), \quad X_0^n = x_0, \quad t_{i+1}^n = t_i^n + \frac{1}{n}, \quad t_0^n = 0.$$

Consider, for example, the estimation of $P_{x_0}\{X^n(T) \in B\}$. For this problem, it is assumed that the rare set $B \subset \mathbb{R}^d$ does not contain the terminal values $\phi(T)$ of zero-cost trajectories that start at x_0, and the typical behavior is $\{X^n(T) \in A\}$ with $A = B^c$. (Note that if we wish to continue the reduction to the time-independent case, then with the state space $\mathbb{R}^{d+1}$, we would call the rare set $B \times \{T\} \subset \mathbb{R}^{d+1}$ and the typical set $A \times \{T\}$.) The definition of subsolution becomes the following, with $\bar{K}_{y,t,T}$ the set of absolutely continuous trajectories ϕ with $\phi(t) = y$ and $\phi(T) \in B$.

Definition 16.20 A continuous function $\bar{V} : \mathbb{R}^d \times [0, T] \to \mathbb{R}$ is a subsolution if it is bounded from below and

$$\bar{V}(y, t) \leq \inf_{\phi \in \bar{K}_{y,t,T}} \left[\int_t^T L(\phi(s), \dot{\phi}(s)) ds + \bar{V}(\phi(T), T) \right]$$

for all $(y, t) \in \mathbb{R}^d \times [0, T)$, and $\bar{V}(z, T) \leq 0$ for $z \in B$.

One can add a "bounding" set as in Remark 15.4, which does not change the requirement for $\bar{V}$ to be a subsolution, except that $A \times \{T\}$ now also includes the points $D^c \times [0, T)$, and we restrict in the definition to $(y, t) \in D \times [0, T)$. As in Remark 16.13, a classical or piecewise classical subsolution in the sense of Definitions 14.1 and 14.2 is a subsolution in the sense of Definition 16.20.

A related problem of interest is to estimate the probability of escaping any time during the time interval, i.e., $P_{x_0}\{X^n(t) \in B \text{ for some } t \in [0, T]\}$. In this case, one should replace B in the time-independent setting by $B \times [0, T]$, and A by $A \times \{T\}$. The definition of subsolution is then the following.

Definition 16.21 A continuous function $\bar{V} : \mathbb{R}^d \times [0, T] \to \mathbb{R}$ is a subsolution if it is bounded from below and

$$\bar{V}(y, t) \leq \inf_{\phi \in \bar{K}_{y,t,s}, t \leq s \leq T} \left[\int_t^s L(\phi(s), \dot{\phi}(s)) ds + \bar{V}(\phi(s), s) \right]$$

for all $(y, t) \in \mathbb{R}^d \times [0, T)$ and $\bar{V}(z, t) \leq 0$ for $z \in B$ and $t \in [0, T]$.

As an elementary time-dependent example, we consider the case in which the $\{v_i(x)\}$ are $N(0, 1)$ (and thus independent of x), so that $H(\alpha) = \alpha^2/2$ and $L(\beta) = \beta^2/2$. With $B = [1, \infty)$ and $T = 1$,

$$\bar{V}(x, t) = -x + \frac{1}{2}t + \frac{1}{2}$$

is a subsolution with the optimal value at $(0, 0)$. Splitting thresholds as well as the start of a simulation with splitting rate $R = 3$ are depicted in Fig. 16.3.

16.5.1 Subsolutions for Analysis of Metastability

Suppose that x^* has the property that all zero-cost trajectories are attracted to x^* in the sense that for all $x_0 \in \mathbb{R}^d$, the properties $I_S(\phi) = 0$ for all S and $\phi(0) = x_0$ imply that $\phi(S) \to x^*$ as $S \to \infty$. Consider the issue of estimating $P_{x^*}\{X^n(T) \in B\}$. Assume for simplicity that B^c is bounded and define

$$W(x, y) = \inf \left[I_S(\phi) : \phi(0) = x, \phi(S) = y, S < \infty \right].$$

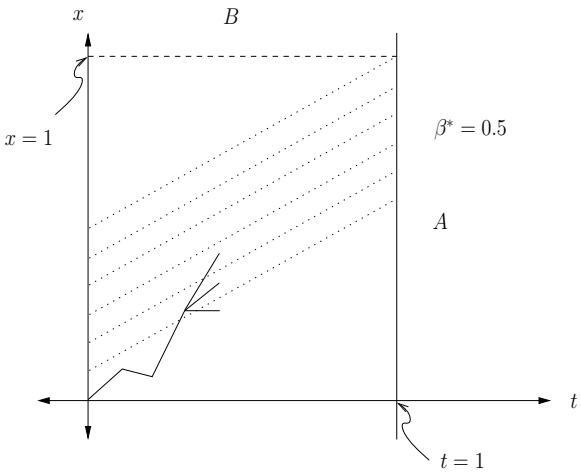

Fig. 16.3 Splitting thresholds for time-dependent problem

Then $W(x, y)$ is the Freidlin–Wentzell quasipotential [140] relative to the starting point x. Suppose again for simplicity of presentation that $W(x^*, \cdot)$ is continuous. In this context, a particularly convenient subsolution is that of the form

$$\bar{V}(y, t) = \bar{V}(y) = -W(x^*, y) + c.$$

Here $c \in \mathbb{R}$ is the largest value such that the boundary condition $-W(x^*, y) + c \leq 0$ holds for all $y \in B$.

To see that $\bar{V}(y)$ is a subsolution, we note that $W(x, z)$ satisfies the dynamic programming equation

$$W(x, z) = \inf_{y \in \mathbb{R}^d} [W(x, y) + W(y, z)],$$

from which it follows that for all $y \in \mathbb{R}^d$ and $c \in \mathbb{R}$,

$$-W(x^*, y) + c \leq -W(x^*, z) + c + W(y, z).$$

The definition of $W(y, z)$ then gives [for all $\phi \in \bar{K}_{y,t,T}$ and with $z = \phi(T)$] that

$$\bar{V}(y) \leq \bar{V}(\phi(T)) + \int_0^T L(\phi(s), \dot{\phi}(s))ds,$$

and therefore $\bar{V}$ is a subsolution. One can show that under appropriate conditions, as $T \to \infty$,

$$\inf_{\phi \in \bar{K}_{x^*,t,T}: \phi(T) \in B} \left[\int_0^T L(\phi(s), \dot{\phi}(s))ds \right]$$

converges to $\bar{V}(x^*)$, and hence $\bar{V}(y)$ is a potentially useful subsolution for studying escape to B from a neighborhood of the attractor x^* at the end of a long time interval.

With regard to the problem of estimating $P_{x^*}\{X^n(t) \in B \text{ for some } t \in [0, T]\}$, $\bar{V}(y)$ again provides a (time-independent) subsolution with a nearly optimal value at x^* when T is large. The argument is similar to the case of $P_{x^*}\{X^n(T) \in B\}$ and is hence omitted.

16.6 Notes

Particle splitting methods originate with [166], and are further developed in [30]. A review of their application to rare-event problems appears in [145], as well as [223]. The RESTART algorithm, which is the focus of this chapter, was first presented in [241].

The main source for this chapter is [77], which uses a more general formulation and also phrases the assumptions to explicitly include queueing networks and expected values. Just as with importance sampling, some incorrect uses of large deviation asymptotics for the design of splitting schemes have been proposed, and a discussion on these issues can be found in [149].

Chapter 17
Examples of Subsolutions and Their Application

In this chapter we present examples to illustrate the importance sampling and splitting techniques developed in Chaps. 14, 15, and 16. There are many different types of problems one might consider, and the interested reader can find additional examples in the references [76, 77, 101, 103, 105, 110, 112, 113, 116, 117]. As mentioned in Chaps. 14 and 16, an important distinction is that in the case of importance sampling, we use a smooth classical-sense subsolution, while in the case of splitting, we use a continuous but not necessarily smooth weak-sense solution. For many of the examples presented, the construction of subsolutions can be carried out in arbitrary dimension. However, for higher-dimensional problems, the construction can be algebraically complicated, and for this reason we present lower-dimensional examples and send the reader to the references for the general case. It also should be repeated that the structure of each *subsolution* that we construct is usually much simpler than that of the corresponding *solution*.

The purpose of each section is to illustrate how different features are accommodated in the construction. Section 17.1 describes the construction for expected values rather than probabilities. We will also give, at least for this first example, calculations analogous to those of Remark 16.13, to show that any classical or piecewise classical subsolution suitable for importance sampling is also suitable for splitting. This argument is broadly applicable, and can be adapted to cover all the other examples we present. Section 17.2 considers level crossings and more generally hitting probabilities, and Sect. 17.3 considers a functional that depends on the entire trajectory of the process. Section 17.4 contains the only example we present for the important class of problems with discontinuous statistics, which includes various problems from queueing theory, and Sect. 17.5 considers the construction of subsolutions that are appropriate for probabilities that fall into the regime of moderate deviations. Section 17.6 compares importance sampling and splitting for estimating probabilities that a process escapes from a neighborhood of a metastable point, and concludes (based only on empirical evidence) that splitting is preferred when the time interval of interest is large.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_17

In each section we essentially follow the following format: problem statement; formulation of the PDE and boundary/terminal conditions; identification of functions from which a subsolution can be assembled, which we call component functions; construction of subsolutions; and finally, numerical examples.

17.1 Estimating an Expected Value

As noted in Sect. 14.5.1, various risk-sensitive functionals can be approximated using the same methods as used to estimate rare events. In this section we consider such a problem when the underlying process model is a random walk, and numerically test the resulting importance sampling and splitting algorithms.

17.1.1 Problem Statement

We consider an example of the estimation problem discussed in Sect. 14.5.1. Suppose $\{v_i\}_{i=1}^{\infty}$ are iid $\mathbb{R}^d$-valued random variables, and let

$$X_i^n = \frac{1}{n}\sum_{j=1}^{i} v_j, \;\; X_0^n = 0.$$

We are interested in unbiased Monte Carlo approximation of $E[e^{-nF(X_n^n)}]$, where $F : \mathbb{R}^d \to \mathbb{R}$ is a continuous function. This is a finite-time problem in the form considered in Sect. 14.3, generalized to the setting of a general functional. It follows from Jensen's inequality that one can restrict in the variational expression (14.22) for the large deviation limit to constant-velocity paths. Therefore, if F satisfies the condition required of h in Theorem 1.18, then

$$\eta \doteq \lim_{n\to\infty} -\frac{1}{n}\log E[e^{-nF(X_n^n)}] = \inf_{\beta\in\mathbb{R}^d}[L(\beta) + F(\beta)]\,. \tag{17.1}$$

Here $L(\beta)$ is the Legendre–Fenchel transform of $H(\alpha)$, the log moment-generating function of v_i. For the right-hand side of (17.1) to be finite, one should impose growth conditions on F. For example, if the $\{v_i\}$ are $N(0, 1)$, so that $H(\alpha) = \alpha^2/2$, and if $F(\beta) = -a\beta^2$, then the right-hand side of (17.1) is $-\infty$ if $a > 1/2$ and finite if $a \le 1/2$.

17.1.2 Associated PDE

As discussed in Sects. 14.2 and 14.3, the Hamilton–Jacobi–Bellman (HJB) equation that is used to construct subsolutions for this problem is

$$V_t(x,t) + \mathbb{H}(DV(x,t)) = 0 \tag{17.2}$$

for $(x,t) \in \mathbb{R}^d \times [0,1)$, where $\mathbb{H}(p) = -H(-p)$, together with the terminal condition

$$V(x,1) = F(x) \text{ for } x \in \mathbb{R}^d. \tag{17.3}$$

17.1.3 Component Functions

As in Sect. 14.3, a natural collection of functions to use in building a subsolution are those of the form

$$\bar{V}(x,t;\beta,c) \doteq -\langle \alpha, (x-\beta)\rangle - H(\alpha)[1-t] + c, \quad \beta \in \mathbb{R}^d, c \in \mathbb{R},$$

where $\alpha = DL(\beta)$ is the point that is dual or conjugate to β in the Legendre–Fenchel transform relating $L(\beta)$ and $H(\alpha)$. Using this duality, one can check as in Sect. 14.5.1 that these are in fact *solutions* to the HJB equation (17.2) [though they may not satisfy the terminal condition]. Although here we parametrize by β, one could also use α. For the Gaussian case considered later, we have the very simple relation $\alpha = \beta$, and so the component functions (when $d = 1$) are $\bar{V}(x,t;\beta,c) \doteq -\beta(x-\beta) - \beta^2[1-t]/2 + c$.

17.1.4 Subsolutions

Convex functions and classical subsolutions. The construction of a subsolution with the maximum possible value (namely η) at $(0,0)$ requires knowing the minimizers in the definition of η in (17.1). For example, if F is convex, then there is a unique minimizer β^*, and one might conjecture that $\bar{V}(x,t;\beta^*, F(\beta^*))$ is such a subsolution, where $c = F(\beta^*)$ is the largest value of c that allows the terminal inequality $\bar{V}(x,1;\beta^*,c) \le F(x)$ to be satisfied. When F is twice continuously differentiable, this can be seen as follows. Using the optimality of β^*, we have that

$$0 = DL(\beta^*) + DF(\beta^*) = \alpha^* + DF(\beta^*),$$

with α^* dual to β^*. By Taylor's theorem, convexity, and the last display, we have

$$\begin{aligned} F(x) - F(\beta^*) &= \left\langle (x-\beta^*), DF(\beta^*)\right\rangle + \frac{1}{2}\left\langle (x-\beta^*), D^2F(\bar{x})(x-\beta^*)\right\rangle \\ &\ge \left\langle (x-\beta^*), DF(\beta^*)\right\rangle \\ &= -\left\langle (x-\beta^*), \alpha^*\right\rangle, \end{aligned}$$

where $\bar{x}$ is some point between x and β^*. Therefore,

$$F(x) \geq -\left\langle (x-\beta^*), \alpha^* \right\rangle + F(\beta^*) = \bar{V}(x, 1; \beta^*, F(\beta^*)),$$

showing that $\bar{V}(x, t; \beta^*, F(\beta^*))$ is a subsolution to both the PDE and terminal condition. Since

$$\bar{V}(0, 0; \beta^*, F(\beta^*)) = \left\langle \beta^*, \alpha^* \right\rangle - H(\alpha^*) + F(\beta^*) = L(\beta^*) + F(\beta^*) = \eta,$$

it has the optimal value at $(0, 0)$. Although the derivation just given assumes that F is smooth, it can be generalized to arbitrary convex functions by approximation via smooth convex functions.

Piecewise classical subsolution. Suppose that F is the minimum of a finite collection of convex functions F_k, $k = 1, \ldots, K$ and β_k^* are the corresponding minimizers in (17.1) (with F replaced by F_k). If $\bar{V}^{(k)}(x, t) \doteq \bar{V}(x, t; \beta_k^*, F_k(\beta_k^*))$, $k = 1, \ldots, K$, then the preceding calculations show that $\bar{V}(x, t) \doteq \wedge_{k=1}^K \bar{V}^{(k)}(x, t)$ is a piecewise classical subsolution to (17.2) and (17.3), in the sense analogous to Definition 14.2 but for the terminal cost F. In particular, we have that $\bar{V}(x, 1) = \wedge_{k=1}^K \bar{V}^{(k)}(x, 1) \leq \wedge_{k=1}^K F_k(x) = F(x)$. Then as in Chap. 14, one can use the mollification of this function given by (14.16) and with implementations as described in Sect. 14.4 to construct associated importance sampling schemes. Note that for each k,

$$\bar{V}^{(k)}(0, 0) = L(\beta_k^*) + F_k(\beta_k^*) = \inf_{\beta \in \mathbb{R}^d} [L(\beta) + F_k(\beta)],$$

and therefore

$$\bar{V}(0, 0) = \wedge_{k=1}^K \inf_{\beta \in \mathbb{R}^d} [L(\beta) + F_k(\beta)] = \inf_{\beta \in \mathbb{R}^d} [L(\beta) + F(\beta)] = \eta.$$

Therefore, up to mollification, the optimal value at $(0, 0)$ is achieved. (We also remind the reader that as discussed in Theorem 15.14, one can let the mollification parameter $\delta > 0$ tend to 0 as $n \to \infty$ to obtain asymptotic optimality.)

Subsolution for splitting. While a smooth subsolution is needed for importance sampling, less regularity is needed for splitting, and the analogue of Definition 16.12 that is appropriate for the present setting is to require

$$\bar{V}(y, t) \leq \inf_{\phi : \phi(t) = y} \left[\int_t^1 L(\dot{\phi}(r)) dr + \bar{V}(\phi(1), 1) \right] \tag{17.4}$$

for all $y \in \mathbb{R}^d$, $0 \leq t \leq 1$, and $\bar{V}(x, 1) \leq F(x)$ for $x \in \mathbb{R}^d$. A standard verification argument shows that a piecewise classical subsolution is a subsolution in the sense of (17.4) (together with the boundary condition). Indeed, let absolutely continuous ϕ satisfy $\phi(t) = y$ and let k satisfy $\bar{V}(\phi(1), 1) = \bar{V}^{(k)}(\phi(1), 1)$. Then since $\bar{V}^{(k)}$ is a classical-sense solution to (17.2), for all $\beta \in \mathbb{R}^d$, $z \in \mathbb{R}^d$ and $r \in (0, 1)$,

$$0 = \bar{V}_t^{(k)}(z,r) + \mathbb{H}(D\bar{V}^{(k)}(z,r)) \le \bar{V}_t^{(k)}(z,r) + \langle D\bar{V}^{(k)}(z,r), \beta\rangle + L(\beta).$$

Replacing (z,β) by $(\phi(r), \dot{\phi}(r))$ and integrating from t to 1 gives

$$\bar{V}^{(k)}(y,t) \le \bar{V}^{(k)}(\phi(1),1) + \int_t^1 L(\dot{\phi}(r))dr.$$

Now use that $\bar{V}(y,t) \le \bar{V}^{(k)}(y,t)$, $\bar{V}^{(k)}(\phi(1),1) = \bar{V}(\phi(1),1)$, and that ϕ is arbitrary to get (17.4).

When F is not the minimum of a finite collection of convex functions, one should not use component functions that correspond to the solution to the HJB equation with an affine terminal condition, since they do not give a convenient approximation to F from below. One alternative is to consider solutions to the HJB equation with concave terminal conditions. In the special case of Gaussian $\{v_i\}$, one might consider concave quadratic functions.

17.1.5 Example

As an example, we consider $\{v_i\}$ with distribution $N(0,1)$ and $F = F_1 \wedge F_2 \wedge F_3$ with

$$F_1(x) = (ax + a + 1)^+, \quad F_2(x) = 1, \quad F_3(x) = (bx - b - 1)^-,$$

where $a = 3/2$ and $b = 4$. See Fig. 17.1.

Each F_k is convex, with F_1 giving the minimum on $(-\infty, 0]$, F_2 on $[0,1]$, and F_3 on $[1,\infty)$. One can check that with these choices, $\alpha_1^* = \beta_1^* = -3/2$, $\alpha_2^* = \beta_2^* = 0$, and $\alpha_3^* = \beta_3^* = 5/4$, and that the corresponding classical subsolutions are

$$\bar{V}^{(1)}(x,t) = \frac{3}{2}x + \frac{9}{8}t + \frac{11}{8}, \quad \bar{V}^{(2)}(x,t) = 1, \quad \bar{V}^{(3)}(x,t) = -\frac{5}{4}x + \frac{25}{32}t + \frac{25}{32}.$$

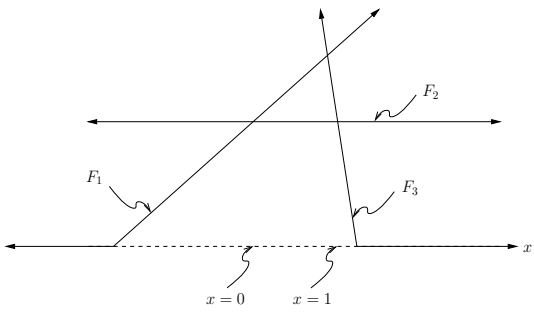

Fig. 17.1 F as the minimum of three convex functions

Table 17.1 Importance sampling estimation of an expected value

	$n = 10$	$n = 20$	$n = 30$
Theoretical value	1.03×10^{-4}	1.87×10^{-8}	5.63×10^{-12}
Estimate	1.02×10^{-4}	1.90×10^{-8}	5.55×10^{-12}
Standard error	0.01×10^{-4}	0.03×10^{-8}	0.09×10^{-12}
95% CI	$[1.00, 1.05] \times 10^{-4}$	$[1.84, 1.95] \times 10^{-8}$	$[5.38, 5.72] \times 10^{-12}$
LD estimate	4.05×10^{-4}	1.64×10^{-7}	6.63×10^{-11}

Table 17.2 Splitting estimation of an expected value

	$n = 10$	$n = 20$	$n = 30$
Theoretical value	1.03×10^{-4}	1.87×10^{-8}	5.63×10^{-12}
Estimate	1.00×10^{-4}	1.64×10^{-8}	5.05×10^{-12}
Standard error	0.04×10^{-4}	0.18×10^{-8}	0.80×10^{-12}
Average # particles	1.5	1.4	1.5
SD # particles	0.03	0.04	0.05
Max # particles	177	208	297
Average # steps	13.1	27.0	42.4
SD # steps	0.14	0.41	0.84
Max # steps	914	2498	4211
95% CI	$[0.91, 1.08] \times 10^{-4}$	$[1.28, 2.00] \times 10^{-8}$	$[3.45, 6.65] \times 10^{-12}$

The large deviation decay rate is $\eta = 25/32$, and the dominant contribution to the integral $E[e^{-nF(X_n^\eta)}]$ is from points in the neighborhood of $x = 5/4$.

For importance sampling, we use a mollification parameter $\delta = 0.1$, and each estimate is based on $L = 20{,}000$ simulations. The theoretical value is computed using the cumulative distribution function of $N(0, 1)$. The simulation results for importance sampling are presented in Table 17.1. In the table, "CI" stands for "confidence interval," and "LD estimate" stands for "large deviation estimate," by which we mean $e^{-n\eta}$.

We next present numerical results for splitting, where no mollification is needed. Splitting thresholds are defined according to (16.26) based on the piecewise classical subsolution $\bar{V}$, and the splitting scheme is implemented as described in Sect. 16.2. The splitting rate is $R = 5$, and we take $\Delta = \log R/n$. Each estimate is based on 20,000 simulations. In Table 17.2, "# particles" is the number of particles that reach discrete time n, and "# steps" is the total number of transitions of all particles that occur during a simulation. "SD" stands for "standard deviation."

17.2 Hitting Probabilities and Level Crossing

17.2.1 Problem Statement

In this section we present examples of importance sampling and splitting for computing hitting probabilities, which include the level crossing probabilities of Sect. 14.5.3 as a special case. Let $\{v_i\}_{i=1}^{\infty}$ be iid random variables taking values in $\mathbb{R}^d$ with common distribution θ, and for $n \in \mathbb{N}$ let $S_n = \sum_{i=1}^{n} v_i$, with the convention that $S_0 = 0$. Given a closed set $B \subset \mathbb{R}^d$ equal to the closure of its interior and $z \in (0, \infty)$, define

$$T_z \doteq \inf\{n \geq 0 : S_n \in zB\}.$$

We are interested in estimating $P\{T_z < \infty\}$, and we assume that $\{T_z < \infty\}$ is a rare event for large z in the sense that

$$\lim_{z \to \infty} -\frac{1}{z} \log P\{T_z < \infty\} = \eta \tag{17.5}$$

for some $\eta > 0$.

17.2.2 Associated PDE

Let

$$H(\alpha) = \log \int_{\mathbb{R}^d} e^{\langle \alpha, y \rangle} \theta(dy)$$

be the log-moment generating function and as usual let L be the Legendre–Fenchel transform of H and $\mathbb{H}(p) = -H(-p)$. The HJB equation that is used to construct subsolutions for this problem is $\mathbb{H}(DV(x)) = 0$ for $x \in \mathbb{R}^d \backslash B$, together with the boundary condition $V(x) = 0$ for $x \in B$.

17.2.3 Component Functions

First suppose that B is a closed convex set (not containing 0) such that $m \doteq \int_{\mathbb{R}^d} y\theta(dy) \notin B$, and that B is the closure of its interior. Then under various regularity conditions on the rate function (e.g., if L is superlinear and finite in a neighborhood of the origin [65]), (17.5) holds with

$$\eta = \inf\left[TL(\beta) : T\beta \in B, T \in (0, \infty)\right]. \tag{17.6}$$

If $H(\alpha) < \infty$ for $\alpha \in \mathbb{R}^d$, then L is superlinear, and in this case, convexity of B implies that there are unique β^* and T^* such that $\eta = T^* L(\beta^*)$ and $T^*\beta^* \in B$. If α^* is dual to β^*, then we claim that $\bar{V}(x; \beta^*, T^*)$ is a subsolution to the PDE and boundary condition, where

$$\bar{V}(x; \beta, T) \doteq -\langle \alpha, (x - T\beta) \rangle, \quad \beta \in \mathbb{R}^d, T \in (0, \infty)$$

and again $\alpha = DL(\beta)$ is the point dual to β. We argue this for the case that B has a smooth boundary of the form $\{y : f(y) = 0\}$, $B = \{y : f(y) \le 0\}$, $Df(y) \neq 0$ for $y \in \partial B$, and f is convex and continuously differentiable, leaving the more general case to the reader.

For B of this form, we can use Lagrange multipliers with Lagrangian $TL(\beta) + \lambda f(T\beta)$ to argue that there exist $T^* \in (0, \infty)$, $\beta^* \in \mathbb{R}^d$ and $\lambda^* \in \mathbb{R}$ that solve

$$\begin{aligned} L(\beta^*) + \lambda^* \left\langle \beta^*, Df(T^*\beta^*) \right\rangle &= 0, \\ T^* DL(\beta^*) + T^* \lambda^* Df(T^*\beta^*) &= 0 \\ f(T^*\beta^*) &= 0, \end{aligned}$$

and for which the (T^*, β^*) component can be identified as the unique minimizer in (17.6). Convexity of f and $0 \notin B$ implies $\langle \beta^*, Df(T^*\beta^*) \rangle < 0$, so that $L(\beta^*) \ge 0$ gives $\lambda^* \ge 0$. Since by convex duality $DL(\beta^*)$ is equal to α^*, the first two equations in the last display give the third equality in the following:

$$\begin{aligned} H(\alpha^*) &= \left\langle \alpha^*, \beta^* \right\rangle - L(\beta^*) \\ &= \left\langle DL(\beta^*), \beta^* \right\rangle - L(\beta^*) \\ &= -\left\langle \lambda^* Df(T^*\beta^*), \beta^* \right\rangle + \lambda^* \left\langle \beta^*, Df(T^*\beta^*) \right\rangle \\ &= 0. \end{aligned}$$

Thus $\bar{V}(x; \beta^*, T^*)$ solves the HJB equation. Moreover, for $x \in B$, convexity implies $\langle Df(T^*\beta^*), (x - T^*\beta^*) \rangle \le 0$, and therefore for such x,

$$\begin{aligned} \bar{V}(x; \beta^*, T^*) &= -\left\langle DL(\beta^*), (x - T^*\beta^*) \right\rangle \\ &= \lambda^* \left\langle Df(T^*\beta^*), (x - T^*\beta^*) \right\rangle \\ &\le 0. \end{aligned} \tag{17.7}$$

Lastly, we note that

$$\bar{V}(0; \beta^*, T^*) = \left\langle DL(\beta^*), T^*\beta^* \right\rangle = -T^*\lambda^* \left\langle Df(T^*\beta^*), \beta^* \right\rangle = T^* L(\beta^*) = \eta,$$

and therefore $\bar{V}(0; \beta^*, T^*)$ achieves the maximum possible value.

17.2.4 Subsolutions

Suppose that B can be written as the union of K convex and closed sets B_k. Assume also that with $m \doteq \int_{\mathbb{R}^d} y\theta(dy)$ we have $m \notin B_k$ for $k = 1, \ldots, K$, that the limit on the left side of (17.5) exists with B replaced by B_k, and that there are unique (T_k^*, β_k^*) that minimize in $\inf [TL(\beta) : T\beta \in B_k, T \in (0, \infty)]$. Let $\bar{V}^{(k)}(x) \doteq \bar{V}(x; \beta_k^*, T_k^*)$, $\bar{V}(x) = \wedge_{k=1}^K \bar{V}^{(k)}(x)$ and $\eta = \bar{V}(0)$. Then (17.5) holds, and moreover, $\bar{V}(x)$ is a piecewise classical subsolution to the HJB equation and boundary condition. The verification of the last claim follows the same line of argument as in the previous section, but is included here for completeness. It is automatic from the definition that each $\bar{V}^{(k)}(x)$ is a classical-sense solution to $\mathbb{H}(DV(x)) = 0$, and so only the boundary condition needs to be checked. However, if $x \in B_k$, then by (17.7),

$$\bar{V}(x) \le \bar{V}(x; \beta_k^*, T_k^*) \le 0,$$

and since k is arbitrary, the boundary inequality holds. Hence the mollification of this subsolution as described in Sect. 14.4 is appropriate for the design of importance sampling schemes. For $\bar{V}(x)$ to be used in the design of splitting schemes, the required properties are the just established boundary inequality $\bar{V}(x) \le 0$ for $x \in \cup_{k=1}^K B_k$, and also

$$\bar{V}(y) \le \int_0^T L(\phi(r), \dot{\phi}(r))dr + \bar{V}(x) \tag{17.8}$$

whenever absolutely continuous ϕ satisfies $\phi(0) = y, \phi(T) = x$ and $\phi(t) \notin B^\circ$ for $t \in [0, T)$, with $T \in (0, \infty)$ (see Definition 16.12). Let k satisfy $\bar{V}^{(k)}(x) = \bar{V}(x)$. Using $0 \le \bar{V}^{(k)}(z) + \langle D\bar{V}^{(k)}(z), \beta\rangle + L(\beta)$ for $z, \beta \in \mathbb{R}^d$ and integrating with $(z, \beta) = (\phi(r), \dot{\phi}(r))$ gives

$$\bar{V}^{(k)}(y) \le \int_0^T L(\phi(r), \dot{\phi}(r))dr + \bar{V}(x).$$

Then (17.8) follows, since $\bar{V}^{(k)}(y) \ge \bar{V}(y)$.

17.2.5 Examples

We present two examples. The first is a two-dimensional Gaussian system that satisfies all the needed conditions for the theoretical results developed in previous chapters to apply. The second example, studied in [150], has been used to illustrate the problems with importance sampling estimators that do not allow state feedback. Since the second example involves exponential random variables, $H(\alpha)$ is not finite for all α, and the theory of Chap. 4 does not apply, and in particular one does not have a standard LDP on path space. A theory on path space has been developed [185], but

the rate function will no longer be finite only on absolutely continuous paths. This is due to the relatively heavy tail of the exponential distribution, and the rate can even be finite on paths with jumps. Nonetheless, for the problem of level crossing one can rigorously justify the importance sampling scheme that is appropriate when $H(\alpha)$ is finite for all α.

For these level crossing problems one might be concerned that simulation times are finite in an appropriate sense, since the problem is of interest over a potentially unbounded time interval. However, the importance sampling change of measure based on a subsolution leads to a process that hits the target set B with probability one, and in fact, one has finite moments for the expected hitting time.

17.2.5.1 Two-Dimensional Gaussian Example

Let

$$v_i = \begin{pmatrix} v_i^1 \\ v_i^2 \end{pmatrix} = \begin{bmatrix} 1 & 1.2 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} Z_i^1 \\ Z_i^2 \end{pmatrix} - \begin{pmatrix} 0.5 \\ 0.4 \end{pmatrix},$$

where $\{Z_i^1\}_{i=1}^{\infty}$ and $\{Z_i^2\}_{i=1}^{\infty}$ are iid $N(0, 1)$, and let

$$B \doteq \{(x_1, x_2) : x_1 \geq 1 \text{ or } x_2 \geq 1.2\}.$$

If

$$\mu \doteq -\begin{pmatrix} 0.5 \\ 0.4 \end{pmatrix}, \quad \Sigma \doteq \begin{bmatrix} 2.44 & 1.2 \\ 1.2 & 1 \end{bmatrix},$$

then the log moment-generating function for v_i is

$$H(\alpha) = \langle \mu, \alpha \rangle + \frac{1}{2} \langle \alpha, \Sigma \alpha \rangle .$$

Suppose that $\eta_i > 0$ satisfy $H((\eta_1, 0)^T) = 0$ and $H((0, \eta_2)^T) = 0$, so that $\eta_1 = 1/2.44$ and $\eta_2 = 0.8$. Let $\alpha_1^* \doteq (\eta_1, 0)$ and $\alpha_2^* \doteq (0, \eta_2)$. Then

$$B = \cup_{i=1}^{2} \{x : \langle x, \alpha_i^* \rangle \geq \eta_i\}. \tag{17.9}$$

We always have $(\beta_k^* T_k^*)_k = 1$, and thus

$$\bar{V}^{(k)}(x) = -\langle \alpha_k^*, x \rangle + \eta_k, \ k = 1, 2.$$

Using (17.9), it is easily checked that $\bar{V} \doteq \bar{V}^{(1)} \wedge \bar{V}^{(2)}$ satisfies the boundary condition $\bar{V}(x) \leq 0$ for $x \in B$ and that $\bar{V}(0) = \eta$, and in fact, $\bar{V}$ is the minimal cost for the corresponding optimal control problem.

To implement importance sampling with mollification parameter $\delta > 0$ for this problem we use the randomized implementation of Sect. 14.4. The process under the

Table 17.3 Importance sampling for level crossing with Gaussian distributions

	$z = 10$	$z = 20$	$z = 30$
Theoretical value	1.14×10^{-2}	1.90×10^{-4}	3.15×10^{-6}
Estimate	1.15×10^{-2}	1.89×10^{-4}	3.15×10^{-6}
Standard error	2.38×10^{-5}	3.99×10^{-7}	6.66×10^{-9}
95% CI	$[1.14, 1.15] \times 10^{-2}$	$[1.88, 1.90] \times 10^{-4}$	$[3.14, 3.17] \times 10^{-6}$

Table 17.4 First splitting for level crossing with Gaussian distributions

	$z = 10$	$z = 20$	$z = 30$
Theoretical value	1.14×10^{-2}	1.90×10^{-4}	3.15×10^{-6}
Estimate	1.16×10^{-2}	1.83×10^{-4}	3.08×10^{-6}
Standard error	3.87×10^{-4}	8.75×10^{-6}	2.02×10^{-7}
Average # steps	32.70	67.42	102.28
Max # steps	336	1520	4953
95% CI	$[1.09, 1.24] \times 10^{-2}$	$[1.66, 2.00] \times 10^{-4}$	$[2.68, 3.47] \times 10^{-6}$
# level crossed	1419	712	476

importance sampling change of measure based on the subsolution is

$$\bar{S}_n = \sum_{i=1}^{n} \bar{v}_i,$$

where the conditional distribution of $\bar{v}_i$ given $\bar{S}_{i-1}$ is

$$\rho_1^\delta\left(\frac{\bar{S}_{i-1}}{z}\right) P_1(dy) + \rho_2^\delta\left(\frac{\bar{S}_{i-1}}{z}\right) P_2(dy),$$

and where $P_k(dy)$ is the distribution of $N(\mu + \Sigma\alpha_k^*, \Sigma)$.

For values of $z = 10, 20, 30$, estimates are based on 20,000 simulations and use $\delta = 0.1$. What we call "theoretical" values are obtained by running 1,000,000 simulations. The results are presented in Table 17.3.

Since $S_n \to (-\infty, -\infty)$ a.s., for the splitting algorithm to terminate we must kill particles if they go too far in the $(-\infty, -\infty)$ direction, which will lead to a small bias (which could itself be estimated using a large deviation calculation). Here we use the terminal set $(-\infty, -1] \times (-\infty, -1]$. The splitting rate is $R = 5$, and we take $\Delta = \log R/z$. The subsolution used to define splitting thresholds is $\bar{V}^{(1)} \wedge \bar{V}^{(2)}$. We conduct 20,000 simulations for each estimate. The theoretical values are obtained by running 1,000,000 simulations of importance sampling. These results are presented in Table 17.4. Here "# level crossed" means the number of simulations in which the level is crossed successfully.

Table 17.5 Second splitting for level crossing with Gaussian distributions

	$z = 10$	$z = 20$	$z = 30$
Theoretical value	1.14×10^{-2}	1.90×10^{-4}	3.15×10^{-6}
Estimate	1.16×10^{-2}	1.85×10^{-4}	3.08×10^{-6}
Standard error	3.90×10^{-4}	8.83×10^{-6}	1.91×10^{-7}
Average # steps	131.02	265.94	400.66
Max # steps	413	1536	3434
95% CI	$[1.09, 1.24] \times 10^{-2}$	$[1.68, 2.03] \times 10^{-4}$	$[2.71, 3.46] \times 10^{-6}$
# level crossed	1418	728	488

Table 17.5 considers the terminal set $(-\infty, -5] \times (-\infty, -5]$. The simulation time is approximately three to four times greater.

17.2.5.2 Two-Dimensional Exponential Example

Let

$$v_i = (v_i^1, v_i^2) = (Z_i^1 - 2, Z_i^2 - 3),$$

where $\{Z_i^1\}_{i=1}^{\infty}$ and $\{Z_i^2\}_{i=1}^{\infty}$ are iid and mutually independent exponentially distributed random variables with mean 1, and let $B = \{x = (x_1, x_2) : x_1 \geq 1 \text{ or } x_2 \geq 1\}$. The log moment-generating functions H_j for v_1^j, $j = 1, 2$, are

$$H_1(\alpha) = -2\alpha - \log(1 - \alpha), \;\; H_2(\alpha) = -3\alpha - \log(1 - \alpha), \;\; \alpha \in (-\infty, 1).$$

For $i = 1, 2$ let η_i be the unique positive solution of $H_i(\alpha) = 0$. Then it can be checked that the limit in (17.5) exists and

$$\eta \doteq \lim_{z \to \infty} -\frac{1}{z} \log P\{T_z < \infty\} = \eta_1 \wedge \eta_2.$$

Let $\alpha_1^* \doteq (\eta_1, 0)$ and $\alpha_2^* \doteq (0, \eta_2)$. Then

$$B = \cup_{i=1}^{2} \{x : \langle x, \alpha_i^* \rangle \geq \eta_i\}.$$

A numerical approximation yields that $\eta_1 \approx 0.8$ and $\eta_2 \approx 0.94$.

We always have $(\beta_k^* T_k^*)_k = 1$, and thus

$$\bar{V}^{(k)}(x) = -\langle \alpha_k^*, x \rangle + \eta_k, \;\; k = 1, 2.$$

As before, $\bar{V} \doteq \bar{V}^{(1)} \wedge \bar{V}^{(2)}$ satisfies the boundary condition $\bar{V}(x) = 0$ for $x \in B$ and $\bar{V}(0) = \eta$, and $\bar{V}$ is the solution. To implement importance sampling with mol-

Table 17.6 Importance sampling for level crossing with exponential distributions

	$z = 10$	$z = 20$	$z = 30$
Theoretical value	7.53×10^{-4}	2.48×10^{-8}	8.47×10^{-12}
Estimate	7.44×10^{-5}	2.46×10^{-8}	8.51×10^{-12}
Standard error	0.08×10^{-5}	0.02×10^{-8}	0.08×10^{-12}
95% CI	$[7.28, 7.60] \times 10^{-5}$	$[2.41, 2.51] \times 10^{-8}$	$[8.35, 8.67] \times 10^{-12}$

lification parameter $\delta > 0$ for this problem we use the randomized implementation of Sect. 14.4. In particular, if the index $k = 1$ is selected then, $\bar{\nu}_i$ will have the distribution of $(\hat{Z}_1 - 2, Z_2 - 3)$, where $\hat{Z}_1$ is exponential with mean $1/(1 - \alpha_1^*)$ and Z_2 (independent of $\hat{Z}_1$) is exponential with mean 1. If $k = 2$ is selected, then $\bar{\nu}_i$ has the distribution of $(Z_1 - 2, \hat{Z}_2 - 3)$, with Z_1 exponential with mean 1 and $\hat{Z}_2$ (independent of Z_1) exponential with mean $1/(1 - \alpha_2^*)$.

We run 20,000 simulations for values of $z = 10, 20, 30$ and use $\delta = 0.1$. The theoretical values are exact up to numerical precision, and they can be obtained because the dynamics in the two dimensions are independent (under the original distribution). Simulation results for importance sampling are presented in Table 17.6.

Splitting can in principle also be applied to this problem. However, the need to truncate the state space causes some difficulties not present in the Gaussian case. Owing to the relatively heavy tail of the exponential distribution, there is a significant probability of reaching the crossing level even when if one starts far away. Therefore, the terminated trajectories will lead to significant errors unless they are terminated only when they are very far from the crossing boundary. This leads to many paths that must be simulated for very long times. In comparison, with importance sampling all simulations reach the crossing level with probability one, and typically in short time, and with no truncation error. For these reasons, importance sampling appears better suited in this context.

17.3 Path-Dependent Functional

In this section we consider the estimation of expected values that depend on the path of a random walk. Problems of this type in a particular form were introduced in Sect. 14.5.4, where the relevant subsolutions were characterized in terms of a coupled pair of PDEs, and the specific form of the estimator was made precise. In this section we explain how one can construct a subsolution in terms of affine functions, and illustrate the performance of the resulting schemes.

17.3.1 Problem Formulation

Let $\{v_i\}_{i=1}^{\infty}$ be iid $\mathbb{R}$-valued random variables with common distribution θ such that $E[v_i] = 0$. Let H and L be the log moment-generating function and its Legendre–Fenchel transform, and let $X_i^n = \frac{1}{n}\sum_{j=1}^{i} v_j$. We are interested in estimating

$$E\left[e^{-nF(X_n^n)}1_{\{\max_{0\le i\le n} X_i^n \ge h\}}\right]$$

for some $h \in (0,\infty)$ and a suitable function F. Under conditions on F and θ, we have

$$\lim_{n\to\infty} -\frac{1}{n}\log E\left[e^{-nF(X_n^n)}1_{\{\max_{0\le i\le n} X_i^n \ge h\}}\right] = \eta,$$

where η is characterized by

$$\eta \doteq \inf\left[\int_0^1 L(\dot{\phi}(t))dt + F(\phi(1)) : \phi(0) = 0, \max_{s\in[0,1]} \phi(s) \ge h\right]. \tag{17.10}$$

17.3.2 Subsolutions

Suppose we consider the particular problem

$$P\left\{\max_{0\le i\le n} X_i^n \ge h, X_n^n \le l\right\}$$

for some $0 < l < h$. Thus $F(x) = \infty 1_{(l,\infty)}(x)$. The relevant PDEs and boundary/terminal conditions for a similar problem were derived in Sect. 14.5.4 and stated in (14.26)–(14.29). For the present problem, they take the form

$$\bar{V}_t(1,x,t) + \mathbb{H}(D\bar{V}(1,x,t)) \ge 0, \quad x \in \mathbb{R}, t \in (0,1),$$

$$\bar{V}(1,x,1) \le 0, \quad x \le l, \tag{17.11}$$

and

$$\bar{V}_t(0,x,t) + \mathbb{H}(D\bar{V}(0,x,t)) \ge 0, \quad x \in \mathbb{R}, t \in (0,1),$$

$$\bar{V}(0,x,t) \le \bar{V}(1,x,t), \quad x \ge h, t \in (0,1). \tag{17.12}$$

As in the previous examples, a subsolution can be identified once we know the solution to the large deviation variational problem. Specifically, using convexity and Jensen's inequality, the decay rate η in (17.10) can be equivalently written as

$$\eta = \inf\left[\rho_0 L\left(\frac{h}{\rho_0}\right) + \rho_1 L\left(\frac{l-h}{\rho_1}\right) : \rho_i \geq 0, i = 0, 1, \rho_0 + \rho_1 = 1\right].$$

Since the mean under θ is zero, we have $L(\beta) = 0$ if and only if $\beta = 0$, and thus the infimum is achieved at $\rho_0^*, \rho_1^* \in (0, 1)$. Let $\beta_0^* = h/\rho_0^*$, $\beta_1^* = (l-h)/\rho_1^*$, and let α_0^*, α_1^* be their convex conjugates. Then ρ_0^* is the length of time, after starting at 0 at $t = 0$, that the velocity β_0^* is applied to push the trajectory over h, and ρ_1^* and β_1^* are the subsequent time and velocity needed to optimally push the trajectory below l at time 1. It can be checked that optimality implies $\alpha_1^* < 0$, $\alpha_0^* > 0$, and $H(\alpha_0^*) = H(\alpha_1^*)$. Using the interpretation of $\bar{V}(1, x, t)$ as the solution to the finite-time problem with the terminal condition $\bar{V}(1, x, 1) \leq 0$ for $x \leq l$, a subsolution as in Sect. 17.1 is given by

$$\begin{aligned}\bar{V}(1, x, t) &= -\alpha_1^*(x - \beta_1^*) - H(\alpha_1^*)[1 - t] + c_1 \\ &= -\alpha_1^*(x - l) - H(\alpha_1^*)[1 - t]\end{aligned}$$

when c_1 is chosen to be the largest value that satisfies (17.11). For times prior to exceeding h, we use a subsolution of the form

$$\bar{V}(0, x, t) = -\alpha_0^*(x - \beta_0^*) - H(\alpha_1^*)[1 - t] + c_0.$$

For the boundary condition (17.12) to be satisfied, we need

$$-\alpha_0^*(h - \beta_0^*) + c_0 \leq -\alpha_1^*(h - l),$$

and to obtain the largest value for $\bar{V}(0, 0, 0)$ we take $c_0 = h[\alpha_0^* - \alpha_1^*] - \alpha_0^*\beta_0^* + \alpha_1^* l$. In particular, the two functions agree when $x = h$. Since $\rho_0^* + \rho_1^* = 1$, $\rho_0^*\beta_0^* + \rho_1^*\beta_1^* = l$, and $H(\alpha_0^*) = H(\alpha_1^*)$, it follows that

$$\begin{aligned}\bar{V}(0, 0, 0) &= \alpha_0^*\beta_0^* - H(\alpha_1^*) + h[\alpha_0^* - \alpha_1^*] - \alpha_0^*\beta_0^* + \alpha_1^* l \\ &= -H(\alpha_1^*) + \rho_0^*\beta_0^*[\alpha_0^* - \alpha_1^*] + \alpha_1^*\left(\rho_0^*\beta_0^* + \rho_1^*\beta_1^*\right) \\ &= \rho_0^*[\alpha_0^*\beta_0^* - H(\alpha_0^*)] + \rho_1^*[\alpha_1^*\beta_1^* - H(\alpha_1^*)],\end{aligned}$$

which equals the optimal asymptotic decay rate η. Note that this is a classical-sense subsolution to the pair of PDEs and boundary/terminal condition, and therefore no mollification is needed.

17.3.3 Example

For a numerical example we take $v_i \sim N(0, 1)$, $h = 1$, and $l = 0.8$. This results in the subsolution

$$\bar{V}(1,x,t) = \frac{6}{5}x - \frac{6}{5}(.8) - \frac{1}{2}(1-t)\left(\frac{6}{5}\right)^2$$

and

$$\bar{V}(0,x,t) = -\frac{6}{5}x + \frac{6}{5}(1.2) - \frac{1}{2}(1-t)\left(\frac{6}{5}\right)^2.$$

Mollification is not needed, since the subsolution is in the classical sense for the pair of PDEs. The controlled process using the importance sampling change of measure based on this subsolution takes the form

$$Y_i^n = \frac{1}{n}\sum_{j=1}^{i} w_j^n,$$

where

$$w_i^n \sim \begin{cases} N(-6/5, 1) & \text{if } Z_i^n = 1, \\ N(6/5, 1) & \text{otherwise,} \end{cases}$$

and $Z_i^n = 1$ if $\max_{0\le j\le i} Y_j^n \ge h$ and $Z_i^n = 0$ otherwise. The corresponding importance sampling estimator is

$$1_{\{\max_{0\le i\le n} Y_i^n \ge 1, Y_n^n \le .8\}} \prod_{j=1}^{n} e^{w_j^n \bar{V}_x(Z_j^n, Y_j^n, j/n) + (6/5)^2}.$$

The results presented in Table 17.7 are based on 20,000 simulations, and we use $n = 10, 20, 30$. The theoretical value is an estimate based on one billion simulations of the importance sampling scheme.

The results for splitting are given in Table 17.8. The splitting rate is $R = 5$, and we take $\Delta = \log R/n$. No mollification is needed, and we conduct 20,000 simulations for each estimate. Here "# particles" is the number of particles that reach the terminal time.

Table 17.7 Importance sampling estimation of a path-dependent functional

	$n = 10$	$n = 20$	$n = 30$
Theoretical value	1.68×10^{-5}	9.66×10^{-9}	6.09×10^{-12}
Estimate	1.71×10^{-5}	9.63×10^{-9}	6.37×10^{-12}
Standard error	0.04×10^{-5}	0.27×10^{-9}	0.19×10^{-12}
95% CI	$[1.62, 1.79] \times 10^{-5}$	$[9.10, 10.2] \times 10^{-9}$	$[6.00, 6.75] \times 10^{-12}$

Table 17.8 Splitting estimation of a path-dependent functional

	$n = 10$	$n = 20$	$n = 30$
Theoretical value	1.68×10^{-5}	9.66×10^{-9}	6.09×10^{-12}
Estimate	2.06×10^{-5}	10.2×10^{-9}	5.04×10^{-12}
Standard error	0.25×10^{-5}	2.06×10^{-9}	0.99×10^{-12}
Average # particles	1.2	1.2	1.2
SD # particles	0.01	0.02	0.02
Max # particles	53	153	152
Average # steps	12.5	27.1	40.8
SD # steps	0.10	0.40	0.62
Max # steps	445	4308	2401
95% CI	$[1.58, 2.54] \times 10^{-5}$	$[6.15, 14.25] \times 10^{-9}$	$[3.10, 6.98] \times 10^{-12}$

17.4 Serve-the-Longest Queue

Subsolutions that induce asymptotically optimal importance sampling and splitting schemes can be constructed for many queuing systems and networks, and references are given in Sect. 17.7. This is due to the fact that the dynamics of these models are "sectionally" homogeneous, and as a consequence, the subsolutions are constructed in terms of a finite number of affine functions. We will illustrate this using the weighted serve-the-longest queue model of Sect. 13.2. There are, however, some subtleties, which arise from the fact that while sectionally homogeneous, the dynamics also have certain discontinuous behaviors. The key issue then is the proper definition of the PDE that characterizes large deviation rates as a function of starting position and time, and which is also used to characterize subsolutions. We review the construction and use of subsolutions for the two-dimensional problem in this section, and refer the reader to [105] for the general case. One can construct subsolutions for either the true continuous time model or, when the event of interest does not depend on time, in terms of the embedded discrete time Markov process, with both constructions leading to the same algorithm.

17.4.1 Problem Formulation

The weighed serve-the-longest queue was described in detail in Sect. 13.2, and we adopt the notation of that section, specialized to the two-dimensional setting. The system state at time t is the vector of queue lengths and is denoted by $Q(t) \doteq (Q_1(t), Q_2(t))$. Also, $\mathscr{V} \doteq \{\pm e_1, \pm e_2\}$, and for $v \in \mathscr{V}$, $r(x; v)$ denotes the jump intensity of the process Q from state x to state $x + v$. For $x = (x_1, x_2) \in \mathbb{R}^2_+$ and $x \neq 0$, let

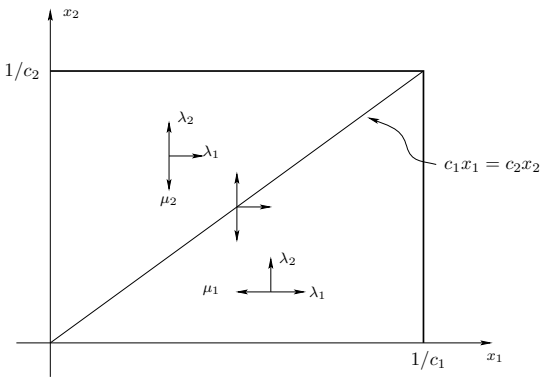

Fig. 17.2 System dynamics for $d = 2$

$$\pi(x) \doteq \left\{ i \in \{1,2\} : c_i x_i = \max_{j \in \{1,2\}} c_j x_j \right\}.$$

Then

$$r(x;v) = \begin{cases} \lambda_i & \text{if } v = e_i, i = 1,2, \\ \mu_2 & \text{if } v = -e_2, 2 \in \pi(x), \\ \mu_1 & \text{if } v = -e_1, \{1\} = \pi(x), \\ 0 & \text{otherwise.} \end{cases}$$

Thus queue 2 is given priority when there are ties. For $x = 0$, there is no service, and the jump intensities are

$$r(0;v) = \begin{cases} \lambda_i, & \text{if } v = e_i \text{ and } i = 1,2, \\ 0, & \text{otherwise.} \end{cases}$$

See Fig. 17.2.

For the system to be stable, we assume

$$\sum_{j=1}^{2} \frac{\lambda_j}{\mu_j} < 1. \tag{17.13}$$

The quantity of interest is the buffer overflow probability

$$p_n \doteq P\{Q_i(t) \text{ reaches } n/c_i \text{ for } i = 1 \text{ or } 2 \text{ before } Q(t) = 0 \,|\, Q(0) = q_0\},$$

where n is large and $q_0 = (0,1)$ or $(1,0)$. Let

$$G \doteq \{x = (x_1, x_2) \in \mathbb{R}^2_+ : c_i x_i < 1, i = 1,2\}.$$

and let G^c denote the complement of G relative to $\mathbb{R}^2_+$. The probability p_n is then conveniently phrased in terms of the scaled process $X^n(t) \doteq Q(nt)/n$ as

$$p_n \doteq P\left\{X^n(t) \text{ reaches } G^c \text{ before } X^n(t) = 0 \,\middle|\, X^n(0) = q_0/n\right\}.$$

17.4.2 Associated Rate Function

For $i = 1, 2$ and for $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$ define

$$H^{\{i\}}(\alpha) \doteq \mu_i(e^{-\alpha_i} - 1) + \sum_{j=1}^{2} \lambda_j(e^{\alpha_j} - 1).$$

We recall the definition of the rate function for the WSLQ model from Sect. 13.26 [see (13.6)], specialized to $d = 2$, and for absolutely continuous trajectories $\phi : [0, T] \to \mathbb{R}^2_+$. For $i = 1, 2$, let $L^{\{i\}}$ be the Legendre–Fenchel transform of $H^{\{i\}}$. We also define $L^{\{1,2\}}$ to be the Legendre–Fenchel transform of $\max_{i\in\{1,2\}} H^{\{i\}}$, that is,

$$L^{\{1,2\}}(\beta) \doteq \sup_{\alpha\in\mathbb{R}^2} \left[\langle \alpha, \beta\rangle - \max_{i\in\{1,2\}} H^{\{i\}}(\alpha)\right]$$

for each $\beta \in \mathbb{R}^2$. Because the system is stable, we have $L^{\pi(0)}(0) = 0$ (see Lemma 13.4), and since for an absolutely continuous ϕ the Lebesgue measure of the set $\{t \in [0, T] : \phi(t) = 0, \dot{\phi}(t) \neq 0\}$ is zero, this is the only value of $L^{\pi(0)}(\cdot)$ that matters. As shown in Chap. 13, the rate function takes the form

$$I_{\phi(0)}(\phi) \doteq \int_0^T L^{\pi(\phi(t))}(\dot{\phi}(t))dt.$$

Thus $L^{\{1,2\}}(\beta)$ is the correct local rate function when ϕ is on the boundary set $\{(x_1, x_2) : \pi(x) = \{1, 2\}\}$. Since $\max_{i\in\{1,2\}} H^{\{i\}}(\alpha)$ is a proper convex function, it is the Legendre–Fenchel transform of $L^{\{1,2\}}(\beta)$. It is important to note that this very simple relationship, in which the local rate function at x is simply the Legendre–Fenchel transformation of the pointwise maximum of the $H^{\{i\}}$ for which $i \in \pi(x)$, in general does not hold for processes with such "discontinuous statistics." In the current setting, it is due to the "stability about the interface" property of the WSLQ discipline policy discussed at length in Sect. 13.2.

17.4.3 Adaptations Needed for the WSLQ Model

In Chaps. 14–16, we considered the problem of hitting a set B before A. A standing assumption was that all trajectories with zero cost would eventually enter the open

set A. This setup is appropriate for many types of processes, since, for example, with a diffusion process in dimension greater than one, it does not make sense to consider the question of reaching some set B before hitting a single point. However, for some systems, and in particular for queueing systems, such a question can make perfect sense.

Indeed, for the WSLQ system, we replace A by $\{0\}$, and we also have initial conditions of the form $X^n(0) \in \{(1/n, 0), (0, 1/n)\}$, which converge to 0 as $n \to \infty$. Since the event of interest is qualitatively different, it follows that the formulation of the variational problem and related definitions of subsolution are slightly different, as is the proof of the performance bounds. The variational problem becomes

$$V(x) = \inf \left[\int_0^T L^{\pi(\phi(t))}(\dot{\phi}(t))dt : \phi(0) = x, \phi(T) \in B, T < \infty \right], \qquad (17.14)$$

where $B = G^c$. Although one might expect, as in Chap. 15, that there is also the requirement that $\phi(t) \notin \{0\}$ for $t \in (0, T)$, it is easy to check that if a trajectory passes through $\{0\}$ on the way to B, then one can find a trajectory with starting point x and nearly the same cost that avoids $\{0\}$. The formulation (17.14) is preferred, because it gives the correct value for the statement of rate of decay for the second moment at the point $x = 0$, which, as observed, is the limit of the starting points that are natural for this problem.

A second adaptation is needed owing to the discontinuous statistical behavior of the process. As we will see, the construction of subsolutions will be based on important roots associated with the Hamiltonians $H^{\{i\}}(\alpha)$, $i = 1, 2$, and $H^{\{1,2\}}(\alpha) \doteq H^{\{1\}}(\alpha) \vee H^{\{2\}}(\alpha)$. We recall that as used in Chap. 15, the classical-sense subsolution property for $\bar{V}$ with respect to the PDE (but not necessarily the boundary conditions) means that $\mathbb{H}(x, D\bar{V}(x)) \geq 0$ for x in the domain of interest. However, Theorem 15.1 in that chapter assumes conditions that imply, among other properties, that $\mathbb{H}(x, \alpha)$ is jointly continuous. As a consequence, this performance bound on importance sampling does not directly cover the WSLQ model, owing to the discontinuity of the dynamics in a neighborhood of $\{(x_1, x_2) : \pi(x) = \{1, 2\}\}$.

For an analysis of the second moment that gives the (not tight) lower bound of $2\bar{V}(x)$ on the decay rate of the second moment, we refer to [105]. The reader interested in the tighter bound $V(x) + \bar{V}(x)$, which is analogous to that stated in Theorem 15.10, can adapt the arguments of Chap. 15 using the definition of subsolution given in the next subsection. The situation with regard to splitting is simpler. Since the performance analysis uses only certain large deviation properties of the process model, one can continue to use the notion of subsolution introduced in Definition 16.12, except as with (17.14), trajectories do not have to avoid $A = \{0\}$ before reaching B.

17.4.4 Characterization of Subsolutions

Let

$$\mathbb{H}^{\{1\}}(\alpha) \doteq -H^{\{1\}}(-\alpha), \quad \mathbb{H}^{\{2\}}(\alpha) \doteq -H^{\{2\}}(-\alpha), \quad \mathbb{H}^{\{1,2\}}(\alpha) \doteq -H^{\{1,2\}}(-\alpha),$$

and let $\bar{V}$ be continuously differentiable on an open neighborhood of G. Each Hamiltonian in the last display is appropriate to that part of the domain where its dual is the local rate function. To be precise, given a continuously differentiable function $\bar{V} : \mathbb{R}^2 \to \mathbb{R}$, the analogue of Definition 14.4 is

$$\mathbb{H}^{\Pi}(\nabla \bar{V}(x)) \geq 0 \text{ for } x \in \{x \in \mathbb{R}^2_+ \backslash \{0\} : \pi(x) = \Pi\}$$

and $\bar{V}(x) \leq 0$ for $x \in B$, and the definition corresponding to Definition 14.5 is analogous.

17.4.5 Component Functions

Ignoring for the moment the boundary condition, we will construct a subsolution as the minimum of three affine functions, one for each region. The critical component function will be the one corresponding to $\mathbb{H}^{\{1,2\}}$, since it must bridge the regions that correspond to $\pi(x) = \{1\}$ and $\pi(x) = \{2\}$.

In the situation considered here, with just two regions separated by one boundary, it will be easy to construct the needed affine functions. We refer to [105] for the more involved general case. We first claim that there exist vectors $\alpha^{\{1\}}$, $\alpha^{\{2\}}$ such that

$$\mathbb{H}^{\{1\}}(\alpha^{\{1\}}) = 0 \text{ with } \alpha_2^{\{1\}} = 0 \text{ and } \alpha_1^{\{1\}} < 0,$$
$$\mathbb{H}^{\{2\}}(\alpha^{\{2\}}) = 0 \text{ with } \alpha_1^{\{2\}} = 0 \text{ and } \alpha_2^{\{2\}} < 0.$$

The existence of $\alpha^{\{1\}}$ follows from the following: $\mathbb{H}^{\{1\}}(0) = 0$; $D_\alpha \mathbb{H}^{\{1\}}(0) = D_\alpha H^{\{1\}}(0) = (\lambda_1 - \mu_1, \lambda_2)$; the stability condition (17.13) then implies that the 1-component of $D_\alpha \mathbb{H}^{\{1\}}(0)$ is negative; and that $\mathbb{H}^{\{1\}}((z, 0)) \to -\infty$ as $z \to -\infty$. These facts imply the existence of $z < 0$ such that $\mathbb{H}^{\{1\}}((z, 0)) = 0$, and the strict concavity of $\mathbb{H}^{\{1\}}$ implies that the solution is unique. A similar argument gives the existence (and uniqueness) of $\alpha^{\{2\}}$.

Next we observe that owing to (17.13), there exists $z \in (0, \min_{i \in \{1,2\}} \mu_i)$ such that

$$\sum_{j \in \{1,2\}} \frac{\lambda_j}{\mu_j - z} = 1, \tag{17.15}$$

and define

$$\alpha_i^{\{1,2\}} \doteq \log\left(1 - \frac{z}{\mu_i}\right) \text{ for } i = 1, 2.$$

Since $z \in (0, \min_{i\in\{1,2\}} \mu_i)$, it follows that $\alpha_i^{\{1,2\}} < 0$ for $i = 1, 2$. We claim that $\mathbb{H}^{\{i\}}(\alpha^{\{1,2\}}) = 0$ for $i = 1, 2$, and thus $\mathbb{H}^{\{1,2\}}(\alpha^{\{1,2\}}) = 0$. We prove just the case $i = 1$, since the case $i = 2$ follows via an analogous argument. Using the explicit expression for $\mathbb{H}^{\{1\}}$ and (17.15), we obtain

$$\begin{aligned}
\mathbb{H}^{\{1\}}(\alpha^{\{1,2\}}) &= -\mu_1(e^{\alpha_1^{\{1,2\}}} - 1) - \lambda_1(e^{-\alpha_1^{\{1,2\}}} - 1) - \lambda_2(e^{-\alpha_2^{\{1,2\}}} - 1) \\
&= -\mu_1\left[\left(1 - \frac{z}{\mu_1}\right) - 1\right] - \lambda_1\left[\left(1 - \frac{z}{\mu_1}\right)^{-1} - 1\right] \\
&\quad - \lambda_2\left[\left(1 - \frac{z}{\mu_2}\right)^{-1} - 1\right] \\
&= z - \lambda_1\left[\frac{z}{\mu_1 - z}\right] - \lambda_2\left[\frac{z}{\mu_2 - z}\right] \\
&= 0.
\end{aligned}$$

The component functions that will be used to construct a piecewise classical subsolution are then

$$\bar{V}^{\Pi}(x) = \langle x, \alpha^{\Pi}\rangle + b_{\Pi}, \quad \Pi = \{1\}, \{2\}, \{1, 2\},$$

where the constants b_{Π} will be chosen to satisfy boundary conditions.

17.4.6 Subsolutions

Define the vector

$$\bar{c} \doteq (1/c_1, 1/c_2)$$

and define

$$\eta \doteq \min\left\{\langle -\alpha^{\Pi}, \bar{c}\rangle : \Pi = \{1\}, \{2\}, \{1, 2\}\right\}. \tag{17.16}$$

Then by a verification argument, η can be shown to be the value of the calculus of variations problem that is associated with the large deviation decay rate of $\{p_n\}$ given by (17.14) with $x = 0$, with the minimizing Π identifying the region traveled by the optimal trajectory (where $\Pi = \{1, 2\}$ corresponds to traveling along the interface). Suppose the constants b_{Π} are defined by $b_{\Pi} \doteq \langle -\alpha^{\Pi}, \bar{c}\rangle$ for $\Pi = \{1\}, \{2\}, \{1, 2\}$. Then for all $x \in \mathbb{R}^+$, we have the boundary values

$$\bar{V}^{\{1\}}((1/c_1, x)) = \langle \alpha^{\{1\}}, \bar{c}\rangle + b_{\{1\}} = 0, \quad \bar{V}^{\{2\}}((x, 1/c_2)) = \langle \alpha^{\{2\}}, \bar{c}\rangle + b_{\{2\}} = 0,$$

and similarly $\bar{V}^{\{1,2\}}((1/c_1, 1/c_2)) = 0$. Also, it is automatic from the definition that if Π is the minimizer in (17.16), then $\bar{V}^{\Pi}(0) \le \bar{V}^{\Theta}(0)$ for $\Theta \neq \Pi$.

We can now construct piecewise classical subsolutions to be used for importance sampling as well as the subsolutions that will be used for splitting. The construction is divided into cases according to the minimizer in (17.16), and by symmetry, it is enough to consider just the cases in which that minimizer is either $\Pi = \{1\}$ or $\Pi = \{1, 2\}$.

The case $\Pi = \{1\}$ is the unique minimizer. In this case, the minimizing trajectory in (17.14) that starts at 0 stays in the region where $\pi(x) = \{1\}$, so that c_1 times the first component is larger than c_2 times the second. In particular, this trajectory will exit at a point of the form $(1/c_1, x_2)$, with $x_2 < 1/c_2$. This suggests the use of $\bar{V}^{\{1\}}(x)$ in this region, but does not resolve the issue of what to use in the regions where either $\pi(x) = \{2\}$ or $\pi(x) = \{1, 2\}$. In order that the correct affine function minimize in these regions, we lower the functions $\bar{V}^{\{2\}}$ and $\bar{V}^{\{1,2\}}$ so that their values at 0 agree with $\bar{V}^{\{1\}}(0) = \eta$ (since $\Pi = \{1\}$ is the unique minimizer, it follows that $\bar{V}^{\{2\}}(0) > \eta$ and $\bar{V}^{\{1,2\}}(0) > \eta$). Specifically, we let

$$r_2 \doteq \eta / \bar{V}^{\{2\}}(0) \in (0, 1) \text{ and } r_{1,2} \doteq \eta / \bar{V}^{\{1,2\}}(0) \in (0, 1).$$

Owing to the concavity of $\mathbb{H}^{\{2\}}$, $r_2 \bar{V}^{\{2\}}(x)$ satisfies

$$\mathbb{H}^{\{2\}}(D[r_2 \bar{V}^{\{2\}}](x)) \ge r_2 \mathbb{H}^{\{2\}}(\alpha^{\{2\}}) + (1 - r_2)\mathbb{H}^{\{2\}}(0) = 0,$$

and likewise $\mathbb{H}^{\{2\}}(D[r_{1,2}\bar{V}^{\{1,2\}}](x)) \ge 0$, and therefore the subsolution property is preserved if $\bar{V}^{\{2\}}$ [resp. $\bar{V}^{\{1,2\}}$] is replaced by $r_2 \bar{V}^{\{2\}}$ [resp. $r_{1,2}\bar{V}^{\{1,2\}}$].

There is a last small adjustment needed for importance sampling. If the mollification of Sect. 14.2 were applied to the current piecewise classical subsolution to produce $\bar{V}^{\delta}(x)$, then gradients in any open neighborhood of $\{x \in G : \pi(x) = \{1, 2\}\}$ would be a convex combination $\rho_1 \alpha^{\{1\}} + \rho_2 r_2 \alpha^{\{2\}}$, and thus the subsolution property may no longer hold in this region. To deal with this, we will lower $r_{1,2}\bar{V}^{\{1,2\}}(x)$ slightly so that it is the smallest of the three functions in some open neighborhood of this set. Thus the piecewise classical subsolution that is used for importance sampling is

$$\bar{V}_\kappa(x) = \bar{V}^{\{1\}}(x) \wedge [r_2 \bar{V}^{\{2\}}(x)] \wedge [r_{1,2}\bar{V}^{\{1,2\}}(x) - \kappa],$$

where $\kappa \in (0, 1)$. While it is still true that the gradient of $\bar{V}_\kappa$ is a convex combination of $\alpha^{\{1\}}$, $r_2\alpha^{\{2\}}$, and $r_{1,2}\alpha^{\{1,2\}}$, it is easy to check (see [105, Sect. 7.2]) that because $\mathbb{H}^{\{1\}}(r_{1,2}\alpha^{\{1,2\}}) \wedge \mathbb{H}^{\{2\}}(r_{1,2}\alpha^{\{1,2\}}) \ge 0$, in fact $\mathbb{H}(x, D\bar{V}_\kappa^\delta(x)) \ge -Ke^{-\kappa/\delta}$ for all $x \in G$, where $K \in [0, \infty)$ can be made explicit in terms of the system parameters. This "loss" of the subsolution property can be made negligible by choosing κ appropriately in terms of δ. Since $\bar{V}_\kappa^\delta(x)$ satisfies $\bar{V}_\kappa^\delta(x) \le 0$ for $x \in \partial G$, this function can be used for either the randomized or ordinary implementation. One can go further, and show that with $\kappa = \kappa_n$ and $\delta = \delta_n$, the resulting schemes are asymptotically optimal so

long as $\kappa_n \to 0$, $\kappa_n/\delta_n \to \infty$ and $n\delta_n \to \infty$ as $n \to \infty$ (see [105, Proposition 8.1] and Theorem 15.14).

Using a verification argument, one can show that $\bar{V}_\kappa$ is a suitable subsolution for splitting in the sense of the appropriate analogue of Definition 16.12 (see Sect. 17.2) for each $\kappa > 0$. Sending $\kappa \to 0$ shows that

$$\bar{V}_0(x) = \bar{V}^{\{1\}}(x) \wedge [r_2 \bar{V}^{\{2\}}(x)]$$

is a suitable subsolution for the design of splitting schemes, with the optimal value at $x = 0$.

The case $\Pi = \{1, 2\}$ is a minimizer. An analogous argument in this case shows that

$$\bar{V}_\kappa(x) = [r_1 \bar{V}^{\{1\}}(x)] \wedge [r_2 \bar{V}^{\{2\}}(x)] \wedge [\bar{V}^{\{1,2\}}(x) - \kappa]$$

is a suitable piecewise classical subsolution for importance sampling when $\kappa \in (0, 1)$, where $r_1 = \eta/\bar{V}^{\{1\}}(0) \in (0, 1)$, and that

$$\bar{V}_0(x) = [r_1 \bar{V}^{\{1\}}(x)] \wedge [r_2 \bar{V}^{\{2\}}(x)]$$

is a suitable subsolution for the design of splitting schemes. The piecewise classical subsolution is mollified when used for importance sampling just as in the case that $\Pi = \{1\}$ is the minimizer.

17.4.7 Example

We consider the problem with data

$$\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 3, \mu_2 = 4, c_1 = 1/2, c_2 = 1.$$

Then the relevant roots are

$$\begin{aligned}
\alpha^{\{1\}} &= (\log(1/3), 0) \approx (-1.0986, 0), \\
\alpha^{\{2\}} &= (0, \log(1/2)) \approx (0, -0.6931), \\
\alpha^{\{1,2\}} &= (\log([1 + \sqrt{2}]/3), \log([2 + \sqrt{2}]/4])) \approx (-0.2172, -0.1583).
\end{aligned}$$

Since $\bar{c} = (2, 1)$, we see that the minimizer in (17.16) is approximately 0.5927, and it occurs for $\Pi = \{1, 2\}$. Hence the minimizing trajectory is along the interface.

The queueing model is a continuous time pure jump Markov process, and in general, one should use the likelihood ratio appropriate for such processes (e.g., Theorem [161, Theorem III.3.24] and Appendix D.3). However, since we consider here an exit probability, we can instead work with the embedded discrete time Markov

chain, whose transition probabilities are given by renormalizing the rates, so that the probability of a transition from state q to $q + v/n$ is given by

$$r(x, v) \Big/ \sum_{w \in \mathcal{V}} r(x, w).$$

Whereas in the continuous time setting one would interpret the duality formula

$$\mathbb{H}(x, D\bar{V}(x)) = \inf \left[\left\langle D\bar{V}(x), \beta \right\rangle + \sum_{w \in \mathscr{V}} r(x, w) \ell \left(\bar{r}(x, w)/r(x, w) \right) \right]$$

as indicating that the process used for importance sampling is a continuous time Markov process with jump rates (see Sect. 14.5.2)

$$\bar{r}(x, v) = r(x, v) e^{-\langle v, D\bar{V}(x) \rangle},$$

for the embedded chain, we use transition probabilities $\bar{r}(x, v)/\bar{R}(x)$ and the importance sampling estimator

$$1_{\{Y^n_{N^n} \neq 0\}} \prod_{j=0}^{N^n - 1} \frac{r(Y^n_j/n, w_{j+1})/R(Y^n_j/n)}{\bar{r}(Y^n_j/n, w_{j+1})/\bar{R}(Y^n_j/n)}.$$

Here $R(x) \doteq \sum_{v \in \mathcal{V}} r(x, v)$, $\bar{R}(x) \doteq \sum_{v \in \mathscr{V}} \bar{r}(x, v)$, $w_{j+1} = n(Y^n_{j+1} - Y^n_j)$,

$$N^n \doteq \inf\{k \in \mathbb{N} : Y^n_j \notin G \text{ or } Y^n_j = 0\},$$

and $\{Y^n_j\}$ is a Markov chain with these transition probabilities and initial condition $Y^n_0 = q_0$. We note that an entirely analogous development of the large deviation and importance sampling theory is possible that works directly with the embedded chain, and following that route would produce the same expressions.

For importance sampling we use the ordinary mollification with $\delta_n = 1/5 \log n$ and $\kappa_n = -\delta_n \log \delta_n$. We ran 20,000 simulations for values of $n = 20, 50, 80$, and results are presented in Table 17.9. For the system, the exact values are obtained by solving the linear system that arises from a first step analysis. To compute the probability of hitting B before $(0, 0)$ after leaving $(0, 0)$, we compute the probability of returning to $(0, 0)$ before reaching B, after starting at $(1, 0)$ or $(0, 1)$ with probabilities 2/3 and 1/3, respectively.

Table 17.10 gives the corresponding results for splitting. The splitting rate is $R = 5$, and we take $\Delta = \log R/n$. The subsolution used to define splitting thresholds is

$$\bar{V}_0(x) = [r_1 \bar{V}^{\{1\}}(x)] \wedge [r_2 \bar{V}^{\{2\}}(x)].$$

20,000 simulations are used for each estimate.

Table 17.9 Importance sampling for serve-the-longest queue

	$n = 20$	$n = 50$	$n = 80$
Theoretical value	1.90×10^{-5}	4.36×10^{-13}	8.31×10^{-21}
Estimate	1.96×10^{-5}	4.20×10^{-13}	7.75×10^{-21}
Standard error	3.81×10^{-7}	1.87×10^{-14}	5.86×10^{-22}
95% CI	$[1.88, 2.03] \times 10^{-5}$	$[3.84, 4.57] \times 10^{-13}$	$[6.60, 8.90] \times 10^{-21}$

Table 17.10 Splitting for serve-the-longest queue

	$n = 20$	$n = 50$	$n = 80$
Theoretical value	1.90×10^{-5}	4.36×10^{-13}	8.31×10^{-21}
Estimate	1.75×10^{-5}	5.07×10^{-13}	9.81×10^{-21}
Standard error	1.23×10^{-6}	6.21×10^{-14}	1.59×10^{-21}
Average # steps	109.60	464.11	842.64
Max # steps	20650	160622	692367
95% CI	$[1.51, 1.99] \times 10^{-5}$	$[3.86, 6.29] \times 10^{-13}$	$[6.70, 12.93] \times 10^{-21}$
# success	604	193	94

17.5 Jump Markov Processes with Moderate Deviation Scaling

Recall from Chaps. 5, 9 and 10 that a moderate deviation principle provides approximations for events that fall somewhere between a central limit approximation and the full large deviation approximation. As such, they can be used for the development of splitting and importance sampling schemes for events in this regime. The construction uses the same ideas and methods as in the large deviation setting, but is in some sense simpler, since the rate functions are usually quadratic approximations of the large deviation rate in a neighborhood of its minimizers. For more on this and other aspects of the use of the moderate deviation approximation, as well as a discussion of different ways the schemes can be implemented, we refer to [101].

In this section we present an example of a discrete state model in continuous time. We should remark that Theorem 15.1 does not cover this problem, since (as with the development of the moderate deviation theory itself) tightness of processes requires a different treatment from that under the large deviation scaling (see Chap. 5 and Sect. 13.3). We refer to [101] for a proper theoretical basis of the importance sampling scheme. The case of splitting is covered by Chap. 15, since it requires only that certain large deviation bounds be valid.

17.5.1 Problem Formulation

The vehicle that is used to illustrate the use of moderate deviation approximations in algorithm design is the empirical measure process of a collection of finite state Markov chains with mean field interaction. We let $n \in \mathbb{N}$ denote the number of chains and $\mathscr{X} \doteq \{1, 2, 3\}$ the state space of each chain. Then the state space for the empirical measure is a subset of the two-dimensional unit simplex $\mathscr{S} \doteq \{x \in \mathbb{R}^3_+ : \langle x, (1, 1, 1)\rangle = 1\}$. In particular, the state space for each fixed n is finite, but the number of states increases with n. The precise model is as follows.

Suppose that $\{Z^n_i\}_{i=1,\dots,n}$ is a collection of continuous time jump processes with right continuous paths that have limits from the left and with values in $\mathscr{X}$. The associated empirical measure process is given by $X^n(t) \doteq \frac{1}{n}\sum_{i=1}^n \delta_{Z^n_i(t)}$. Then X^n is a stochastic process that takes values in $\mathscr{S}_n \doteq \{x \in \mathscr{S} : x_i = j_i/n, j_i \in \mathbb{N}_0, i = 1, 2, 3\}$. We assume that conditioned on $\{X^n(s), s \in [0, t]\}$, waiting times until the next jump of each of the particles are independent, with rate of jump of a particle from state i to state j given by

$$r_{ij}(x) \doteq 2 - x_j,$$

where $x = X^n(t)$. This implies that the transition rates for the empirical distribution process X^n are given by

$$R^n\left(x, x + \frac{1}{n}(e_j - e_i)\right) \doteq n x_i (2 - x_j),$$

where $x \in \mathscr{S}_n$ and e_i is the unit vector in the ith direction in $\mathbb{R}^3$. The jump rate $R^n(x, y)$ for all other $x, y \in \mathscr{S}_n$ is zero. In the notation of Sect. 13.3 and using $x_1 + x_2 + x_3 = 1$, we have $\beta(x) = (2 - x_1, 2 - x_2, 2 - x_3)$, and therefore if $X^n(0) = x_n$, where $x_n \in \mathscr{S}_n$ satisfies $x_n \to x_0$ as $n \to \infty$, then X^n converges in probability in $\mathscr{D}([0, 1] : \mathscr{S})$ to the solution X^0 of the ODE

$$\frac{d}{dt}X^0(t) = 6\left(\frac{1}{3}(1, 1, 1) - X^0(t)\right), \quad X^0(0) = x_0.$$

Let $\{\varkappa(n)\}$ be a sequence of positive real numbers such that $\varkappa(n) \to 0$ and $\varkappa(n)n \to \infty$ as $n \to \infty$. A moderate deviation principle for X^n gives the asymptotics of probabilities associated with $Y^n(\cdot) = \sqrt{\varkappa(n)n}(X^n(\cdot) - X^0(\cdot)) \in \mathbb{R}^3$. In particular, it follows from Theorem 13.18 (see Remark 13.20 for this form of the rate function) that if $\sqrt{\varkappa(n)n}(x_n - x_0) \to 0$, then Y^n satisfies a Laplace principle on $\mathscr{D}([0, T] : \mathbb{R}^3)$ with scaling function $\varkappa(n)$ and rate function I_M given by

$$I_M(\phi) \doteq \inf\left[\frac{1}{2}\int_0^T \|u(t)\|^2\, dt\right],$$

where the infimum is over all $u \in L^2([0,T]:\mathbb{R}^3)$ such that

$$\phi(t) = -6\int_0^t \phi(s)ds + \int_0^t A^{1/2}(X^0(s))u(s)ds,\ t \in [0,T],$$

and for $x \in \mathscr{S}$,

$$A_{ij}(x) = \begin{cases} -2(x_i + x_j - x_i x_j), & \text{for } i \neq j, \\ 2(1+x_i^2), & \text{for } i = j. \end{cases}$$

Since Sect. 13.3.2 uses the scaling sequence $\varepsilon > 0$, we should make the identification $\varepsilon = 1/n$, and $\varkappa(\varepsilon)$ there is $\varkappa(n)$ here. For this model we will use the moderate deviation principle as the basis for constructing Monte Carlo schemes for estimating probabilities of the form

$$p_n = P\{X_1^n(T) - X_1^0(T) \notin (-c,c)\}$$

for $c \in (0,1)$.

17.5.2 Associated PDE

Since the moderate deviation approximation requires centering on the minimizers of the rate for the corresponding large deviation approximation (i.e., the LLN limit), the event $\{X_1^n(T) - X_1^0(T) \notin (-c,c)\}$ should be reformulated in terms of Y^n as $\{Y_1^n(T) \notin (-c\sqrt{\varkappa(n)n}, c\sqrt{\varkappa(n)n})\}$. Thus when a subsolution is constructed, the boundary condition will depend on n. Since there are many ways in which one could embed $\{X^n\}_{n\in\mathbb{N}}$ into a moderate deviation approximation (that is to say, on the particular scaling sequence $\varkappa(n)$ used), one might be concerned that the approximation itself depends on the embedding. However, as discussed in [101], when the algorithm is written in terms of an importance sampling scheme or splitting scheme for the original process $\{X^n\}_{n\in\mathbb{N}}$, there is no dependence on the embedding.

With the rate function as defined in the last subsection, the moderate deviation approximation leads to an HJB equation for which subsolutions $\bar{V}(y,t)$ are characterized by

$$\bar{V}_t(y,t) \geq \frac{1}{2}\left\langle D_y\bar{V}(y,t), A(X^0(t))D_y\bar{V}(y,t)\right\rangle + 6\left\langle D_y\bar{V}(y,t), y\right\rangle$$

for $y \in \mathbb{R}^3$ and $t \in [0,T]$, and

$$\bar{V}(y,T) \leq 0 \text{ for } y_1 \in (-b,b)^c.$$

Note that for a fixed n, we will have $b = c\sqrt{\varkappa(n)n}$. Hence the subsolution depends on n, but as remarked previously, when interpreted as inducing a scheme for estimating $P\{X_1^n(T) - X_1^0(T) \notin (-c, c)\}$, this dependence vanishes.

17.5.3 Component Functions

As in previous examples (e.g., as in Sect. 17.1), the component functions will be obtained as solutions to the corresponding HJB equation together with a terminal condition that allows for explicit expressions. In the present setting, we will use an affine terminal condition, in which case the PDE and terminal condition are then a special case of those that characterize the value function of the linear/quadratic regulator. To be precise, suppose we denote the solution to

$$\bar{V}_t(y,t) = \frac{1}{2}\left\langle D_y\bar{V}(y,t), A(X^0(t))D_y\bar{V}(y,t)\right\rangle + 6\left\langle D_y\bar{V}(y,t), y\right\rangle$$

for $y \in \mathbb{R}^3$ and $t \in [0,T]$ and

$$\bar{V}(y,T) = \langle w, y\rangle + c$$

by $\bar{V}(y,t;w,c)$. Then by simply computing the derivatives, one can verify that

$$\bar{V}(y,t;w,c) = \left\langle e^{\Omega(t,T)^T}w, y\right\rangle - \frac{1}{2}\left\langle w, \left(\int_t^T e^{\Omega(s,T)}A\left(X^0(s)\right)e^{\Omega(s,T)^T}ds\right)w\right\rangle + c$$

satisfies the PDE and terminal condition (see [101, Theorem 4.1]), where $\Omega(s,t)$ takes values in the set of 3×3 real matrices and is the solution to

$$\frac{d}{dt}\exp\{\Omega(s,t)\} = -6\exp\{\Omega(s,t)\}, s < t, \quad \Omega(s,s) = 0.$$

17.5.4 Subsolutions

Subsolutions will be constructed for the Y^n process but then applied, after a change of variable, to the X^n process. For importance sampling, we will use the mollification of two subsolutions $\bar{V}(y,t;w_1,c_1)$ and $\bar{V}(y,t;w_2,c_2)$, and for (near) asymptotic optimality (under the moderate deviation scaling), we will need that

$$\bar{V}(0,0;w_1,c_1) \wedge \bar{V}(0,0;w_2,c_2) = \inf[I_M(\phi) : \phi(0) = 0, \phi_1(T) \notin (-b,b)].$$

One can show using Lagrange multipliers and the solutions of the form $\bar{V}(y,t;w,c)$ that the minimal cost from $(0,0)$ to (z,T) for arbitrary z satisfies

$$\begin{aligned} C(z) &= \inf\left[I_M(\phi) : \phi(0) = 0, \phi_1(T) = z\right] \\ &= \frac{1}{2}\left\langle z, \left(\int_0^T e^{\Omega(s,T)} A\left(X^0(s)\right) e^{\Omega(s,T)^T} ds\right)^{-1} z\right\rangle, \end{aligned}$$

[101] where $\int_0^T e^{\Omega(s,T)} A\left(X^0(s)\right) e^{\Omega(s,T)^T} ds$ is calculated using numerical approximation to the corresponding ODEs. Therefore,

$$\inf\left[I_M(\phi) : \phi(0) = 0, \phi_1(T) \notin (-b,b)\right] = \inf\left[C(z) : z_1 \in \{-b,b\}\right],$$

and it turns out (for the example) that the minimum of $C(y)$ over z of the form (b,z_2,z_3) is at $(b,0,0)$, and likewise, that over $(-b,z_2,z_3)$ is at $(-b,0,0)$. We also have the representation

$$\bar{V}(0,0;w,g) = \inf_{z\in\mathbb{R}^3}\left[C(z) + \langle w,z\rangle + g\right].$$

We can then use Lagrange multipliers again to maximize over w and g subject to the constraints $\langle w,(b,0,0)\rangle + g \le 0$ and $\langle w,(-b,0,0)\rangle + g \le 0$ to find (w_1,g_1) and (w_2,g_2) such that

$$\begin{aligned} \bar{V}(0,0;w_1,g_1) \wedge \bar{V}(0,0;w_2,g_2) &= C((b,0,0)) \wedge C((-b,0,0)) \\ &= \inf\left[I_M(\phi) : \phi(0) = 0, \phi_1(T) \notin (-b,b)\right]. \end{aligned}$$

17.5.5 Example

We take $T = 1$, and find using numerical evaluation of $\int_0^1 e^{\Omega(t,1)} A(X^0(t)) e^{\Omega(t,1)^T} dt$ and $e^{\Omega(t,1)}$ that

$$w_1 = (-5.39,0,0)b, \quad g_1 = 5.39b^2$$

and

$$w_2 = (5.39,0,0)b, \quad g_2 = 5.39b^2.$$

We use the ordinary implementation of the mollification of $\bar{V}(y,t;w_1,g_1)$ and $\bar{V}(y,t;w_2,g_2)$ as described in Sect. 14.2.

Remark 17.1 Besides the "standard" scheme defined by this mollification, the paper [100] also develops a "corrected" scheme that more carefully accounts for the fact that increments of the true model may not be very close to their Gaussian approximations. For some problems, the corrected scheme outperforms (at times significantly) the

standard scheme when n is small. For the present example, there was little difference in performance, and for this reason we discuss only for the standard scheme.

Although our convention is to use Y^n for the process simulated for purposes of importance sampling, here we have used Y^n already for the moderate deviation approximation. We therefore (knowing that it conflicts with notation in the first three parts of the book) use $\bar{Y}^n$ and $\bar{X}^n$ (in this section only) to denote the simulated process. These processes are constructed as follows. The process $\bar{X}^n$ will have jump times $\{t_i^n\}_{i\in\mathbb{N}}$ with $t_0^n = 0$. Then

$$\bar{X}^n(t_i^n) = x_n + \frac{1}{n}\sum_{j=0}^{i-1} w_j^n,\ i = 1, 2, \ldots,$$

where (t_k^n, w_k^n) are constructed recursively as follows. Conditional on $\bar{X}^n(t_k^n) = x$, $(t_{k+1}^n - t_k^n)$ and w_k^n are independent, with (conditional) distributions given as exponential with rate $n\bar{S}(x, t_k^n)$ and discrete measure on $\{e_j - e_i, i \neq j\}$ with weights $\bar{r}(x, e_j - e_i, t_k^n)$, respectively, where for $x \in \mathscr{S}$, we have

$$\bar{r}(x, e_j - e_i, t) = x_i(2 - x_j)e^{\left\langle e_j - e_i, -D\bar{V}(y, t_k^n)/\sqrt{\varkappa(n)n}\right\rangle},$$

$y = \sqrt{\varkappa(n)n}(x - X^0(t_k^n))$ and

$$\bar{S}(x, t) = \sum_{i\neq j} \bar{r}(x, e_j - e_i, t).$$

Letting $N^n \doteq \min\{i \in \mathbb{N} : t_i^n > 1\}$, the importance sampling estimate based on a single sample of $\{\bar{X}^n\}$ is

$$1_{(-c,c)^c}(\bar{X}_1^n(t_{N^n-1}^n) - X_1^0(t_{N^n-1}^n)) \prod_{i=0}^{N^n-2} e^{-n\left(t_{i+1}^n - t_i^n\right)\left(S(\bar{X}^n(t_i^n)) - \bar{S}(\bar{X}^n(t_i^n), t_i^n)\right)}$$
$$\times \frac{r(\bar{X}^n(t_i^n), w_i^n)}{\bar{r}(\bar{X}^n(t_i^n), w_i^n, t_i^n)} \frac{S(\bar{X}^n(t_{N^n-1}^n))}{\bar{S}(\bar{X}^n(t_{N^n-1}^n), t_{N^n-1}^n)} e^{-n\left(t_{N^n}^n - t_{N^n-1}^n\right)\left(S(\bar{X}^n(t_{N^n-1}^n)) - \bar{S}(\bar{X}^n(t_{N^n-1}^n), t_{N^n-1}^n)\right)},$$

where

$$r(x, e_j - e_i) = x_i(2 - x_j),\ S(x) = \sum_{i\neq j} r(x, e_j - e_i).$$

The theoretical values were estimated using 1,000,000 samples of the scheme. The following numerical results were computed with $x_n = x_0 = (1, 0, 0)$, $n = 200$, $\delta = 0.01$, and using 100,000 samples, where we recall that δ is the mollification parameter (Table 17.11). Thus the empirical measure starts with all processes in state 1, and we estimate a deviation from the LLN limit for this initial condition.

Table 17.11 Importance sampling for MD-based approximation

	$b = 0.12$	$b = 0.16$	$b = 0.2$
Theoretical value	9.10×10^{-5}	2.18×10^{-7}	1.25×10^{-10}
Estimate	9.13×10^{-5}	2.15×10^{-7}	1.23×10^{-10}
Standard error	0.08×10^{-5}	0.02×10^{-7}	0.02×10^{-10}
95% CI	$[8.98, 9.29] \times 10^{-5}$	$[2.10, 2.20] \times 10^{-7}$	$[1.20, 1.26] \times 10^{-10}$

Table 17.12 Splitting sampling for MD-based approximation

	$b = 0.12$	$b = 0.16$	$b = 0.2$
Theoretical value	9.10×10^{-5}	2.18×10^{-7}	1.25×10^{-10}
Estimate	9.10×10^{-5}	2.15×10^{-7}	1.26×10^{-10}
Standard error	1.80×10^{-6}	5.54×10^{-9}	4.03×10^{-12}
Average # steps	766.40	798.99	831.68
Max # steps	7306	13614	17738
95% CI	$[8.75, 9.46] \times 10^{-5}$	$[2.04, 2.26] \times 10^{-7}$	$[1.18, 1.34] \times 10^{-10}$
# success	5198	3291	2189

Table 17.12 corresponds to splitting with $n = 200$. The splitting rate is $R = 5$, and we take $\Delta = \log R/n$. The subsolution used to define splitting thresholds is

$$\bar{V}(y,t) = \bar{V}(y,t;w_1,g_1) \wedge \bar{V}(y,t;w_2,g_2).$$

We used 100,000 simulations for each estimate.

17.6 Escape from the Neighborhood of a Rest Point

The examples of this section illustrate numerically that for some rare event problems, the splitting schemes of the form introduced in Chap. 16 can be computationally more efficient than their importance sampling counterparts. The example considers a problem of escape of a small noise process from the neighborhood of a stable fixed point of the associated noiseless system. We present results for one- and two-dimensional models. In the one-dimensional setting, an importance sampling scheme was developed in [113], where it was observed that schemes based on a time-independent subsolution must degrade as the time horizon becomes larger. The paper also presented methods to improve the performance of the importance sampling estimator for one-dimensional models using a more complex time-dependent change of measure. In higher dimensions, such constructions become challenging, and thus it is of interest to examine the performance of splitting schemes, as possibly preferable to importance sampling methods for problems involving rest points.

17.6.1 Problem Formulation

Consider the following d-dimensional process model, which is a special case of the model of Chap. 4:

$$X_{i+1}^n = X_i^n - \frac{1}{n}\Lambda X_i^n + \frac{1}{n}v_i, \quad X_0^n = 0,$$

where the $\{v_i\}_{i=0}^{\infty}$ are iid $N(0, I)$ random variables and Λ is a symmetric positive definite $d \times d$ matrix. We also define as in Chap. 4 the usual piecewise linear interpolation with $X^n(i/n) = X_i^n$. The LLN dynamics, given by

$$\frac{d}{dt}X^0(t) = -\Lambda X^0(t), \quad X^0(0) = 0,$$

have a unique equilibrium point at the origin. Also, for every $T \in (0, \infty)$, $\{X^n\}_{n\in\mathbb{N}}$ satisfies the LDP on $\mathscr{C}([0, T] : \mathbb{R}^d)$ with a rate function that can be written in the form

$$I_T(\phi) = \inf\left[\int_0^T \frac{1}{2}\|u(s)\|^2\, ds : \dot{\phi} = -\Lambda\phi + u\right]$$

when ϕ is absolutely continuous and $\phi(0) = 0$, and $I_T(\phi) = \infty$ otherwise.

Let $W(x, y)$, as in Sect. 16.5, be the quasipotential for this rate function, and let $S(x) \doteq W(0, x)$. Thus for $x \in \mathbb{R}^d$,

$$S(x) = \inf\left[I_T(\phi) : \phi(T) = x, T < \infty\right].$$

It is easy to check by a verification argument that $S(x) = \langle x, \Lambda x\rangle$. We are interested in computing probabilities such as

$$p_T^{\varepsilon} \doteq P\left\{X^{\varepsilon}(t) \in B \text{ for some } t \le T\right\} \tag{17.17}$$

for various T and suitable B, and in particular for T large.

17.6.2 Associated PDE

Subsolutions to the Hamilton–Jacobi–Bellman equation associated with this escape probability must satisfy

$$\bar{V}_t(x, t) \ge \left\langle D\bar{V}(x, t), \Lambda x\right\rangle + \frac{1}{2}\left\|D\bar{V}(x, t)\right\|^2 \tag{17.18}$$

for $x \in B^c$ and $t \in [0, T]$, and the boundary condition

$$\bar{V}(x,t) \le 0 \text{ for } x \in B \tag{17.19}$$

(note that the terminal condition $\bar{V}(x,T) \le \infty$ for $x \in B^c$ is vacuous).

17.6.3 Subsolutions

Let $V(x,t)$ denote the *solution* to the corresponding HJB equation, and suppose we temporarily put the dependence on T into the notation as $V(x,t;T)$. Define $\bar{V}_\infty(x) \doteq C - S(x)$ for $x \in B^c$, where $C \doteq \inf_{x\in B} S(x)$. One can verify by direct computation that $\bar{V}_\infty$ is a classical-sense subsolution to (17.18) and (17.19). Also, as noted in Sect. 16.5, $\bar{V}_\infty$ is also a weak-sense subsolution in the sense of Definition 16.21. It follows that $\bar{V}_\infty$ can be used for the design of both importance sampling and splitting schemes that will estimate (17.17). Since $V(x,0;T)$ has the interpretation as the minimal cost for trajectories that enter B by time T after starting at x, one has $V(0,0;T) \downarrow \bar{V}_\infty(0)$ as $T \to \infty$. Therefore, for large T, $\bar{V}_\infty(x)$ is a subsolution with a nearly optimal value at the starting point $(x,t) = (0,0)$.

17.6.4 Examples

17.6.4.1 One-Dimensional Example

We first consider a one-dimensional example in which $\Lambda = 1$. The following estimates for importance sampling are based on 1,000,000 samples, while the estimates for splitting are based on only 20,000 samples. This is because for this problem, a single importance sampling estimate is less accurate (but also less costly), so that more samples are required for accurate estimates. We provide estimates of the relative errors (i.e., the standard deviation σ of the estimator divided by the probability p of interest) for various combinations of n and T for both algorithms, keeping only two significant digits. Table 17.13 gives results for importance sampling. Since $\bar{V}_\infty$ is a classical-sense subsolution, no mollification is needed. The splitting algorithm is the RESTART scheme, with $R = 2$ or $R = 5$ and hence $\Delta = (\log R)/n$. Results for splitting appear in Tables 17.14 and 17.15. A result of – indicates that no "successes" occurred in our sample.

The numerical results show that for fixed n, the splitting scheme is much more stable than its importance sampling counterpart as T gets large [both schemes give poor performance for small T, since $\bar{V}_\infty(0)$ is far from $V(0,0;T)$]. This does not contradict the claim that importance sampling is getting closer to asymptotic optimality as T grows, but rather points out a shortcoming of the use of asymptotic optimality as a criterion, which is that the exponential rate of decay of the second moment does not capture the impact of any prefactor terms. For a detailed discussion on these points

Table 17.13 Relative errors for importance sampling, one-dimensional problem

n/T	0.25	0.5	1	1.5	2.5	10	14	18
9	720	33	5.3	2.5	1.9	18	250	150
11	–	61	6.3	2.7	1.8	18	96	180
13	–	130	7.6	2.9	1.8	16	88	84
15	–	170	9.0	3.1	1.8	14	68	140
17	–	370	11	3.3	1.7	22	45	87

Table 17.14 Relative errors for RESTART, $R = 2$, one-dimensional problem

n/T	0.25	0.5	1	1.5	2.5	10	14	18
9	–	74	15	9.3	4.3	1.8	1.5	1.3
11	–	–	19	9.4	4.9	1.9	1.6	1.4
13	–	–	27	14	5.5	2.2	1.7	1.5
15	–	–	60	11	5.9	2.2	1.8	1.5
17	–	–	41	13	6.2	2.3	1.8	1.6

Table 17.15 Relative errors for RESTART, $R = 5$, one-dimensional problem

n/T	0.25	0.5	1	1.5	2.5	10	14	18
9	–	80	25	11	5.8	2.3	2.0	1.7
11	–	–	36	12	6.4	2.4	2.3	1.9
13	–	–	37	16	7.7	2.7	2.2	2.0
15	–	–	40	15	7.6	2.8	2.3	2.0
17	–	–	55	16	9.4	3.0	2.4	2.1

as well as why they apply specifically when the event of interest allows the process to wander in a neighborhood of a rest point, see [113].

For a more nuanced comparison, we bring in the fact that importance sampling and splitting schemes require different computational effort for a single sample. In order to measure the computational time required for comparable performance for the two schemes, we use the concept of *work*. We remark that the notion of work used here is designed to compare the numerical performance of the splitting scheme with the importance sampling scheme, and it is quite different from that introduced in Chap. 16 [see 16.11].

We define a unit of work as the time it takes to simulate one time step under the *original* dynamics. We emphasize that work is determined by the original dynamics because importance sampling requires a state-dependent change of measure that takes additional effort to calculate. We found that when we used Matlab, a single transition for the importance sampling scheme took approximately 35 times as long as a single transition under the original dynamics. Thus each transition step for the importance sampling scheme requires approximately 35 units of work. In the results

below, we account for the additional computation cost for the importance sampling change of measure. The splitting algorithm uses the RESTART scheme, which is much more efficient than ordinary splitting in this context, since in the latter case, every trajectory that does not escape must be simulated till time T. In order to compare the numerical results while accounting for work, we report in the next set of tables the estimated amount of work needed to get a relative error of 1. Recall that the relative error of a simulation based on n samples is $\sigma(\sqrt{n}p)^{-1}$, where σ is the standard deviation of a single sample and p is the probability of interest. Consequently, we need $n^* = (\sigma/p)^2$ samples to get a relative error of 1, and $wn^* = w(\sigma/p)^2$ units of work to get a relative error of 1, where w is the work required for a single sample. Note that the quantities p and σ are estimated from the Monte Carlo results. The work w can be approximated by tracking how many transition steps are needed on average for a single sample and multiplying it by a factor of 35 for the importance sampling scheme and a factor of 1 for the splitting scheme. Results are presented in Tables 17.16, 17.17 and 17.18.

Table 17.16 Work required to obtain a relative error of 1 for importance sampling, one-dimensional problem

n/T	0.25	0.5	1	1.5	2.5	10	14	18
9	5.5×10^7	1.9×10^5	8.8×10^3	3.0×10^3	2.2×10^3	2.2×10^5	4.6×10^7	1.7×10^7
11	–	7.9×10^5	1.5×10^4	4.1×10^3	2.6×10^3	2.8×10^5	8.4×10^6	2.9×10^7
13	–	4.1×10^6	2.6×10^4	5.6×10^3	2.9×10^3	2.8×10^5	8.6×10^6	7.8×10^6
15	–	8.4×10^6	4.3×10^4	7.3×10^3	3.3×10^3	2.7×10^5	6.0×10^6	2.7×10^7
17	–	4.3×10^7	7.1×10^4	9.5×10^3	3.8×10^3	7.1×10^5	3.0×10^6	1.1×10^7

Table 17.17 Work required to obtain a relative error of 1 for RESTART, $R = 2$, one-dimensional problem

n/T	0.25	0.5	1	1.5	2.5	10	14	18
9	–	3.1×10^4	2.9×10^3	2.1×10^3	9.2×10^2	7.6×10^2	7.6×10^2	7.8×10^2
11	–	–	5.7×10^3	2.7×10^3	1.5×10^3	1.2×10^3	1.2×10^3	1.2×10^3
13	–	–	1.3×10^4	7.5×10^3	2.4×10^3	2.0×10^3	1.8×10^3	1.7×10^3
15	–	–	7.7×10^4	5.3×10^3	3.2×10^3	2.5×10^3	2.5×10^3	2.2×10^3
17	–	–	4.1×10^4	8.8×10^3	4.3×10^3	3.1×10^3	3.0×10^3	3.0×10^3

Table 17.18 Work required to obtain a relative error of 1 for RESTART, $R = 5$, one-dimensional problem

n/T	0.25	0.5	1	1.5	2.5	10	14	18
9	–	3.1×10^4	3.6×10^3	2.0×10^3	1.4×10^3	1.2×10^3	9.9×10^2	1.0×10^3
11	–	–	7.2×10^3	4.1×10^3	2.3×10^3	1.7×10^3	1.5×10^3	1.5×10^3
13	–	–	4.8×10^4	3.8×10^3	3.2×10^3	2.3×10^3	2.2×10^3	2.1×10^3
15	–	–	4.9×10^4	8.7×10^3	4.1×10^3	2.8×10^3	3.0×10^3	3.0×10^3
17	–	–	5.5×10^4	1.4×10^4	7.1×10^3	3.8×10^3	3.6×10^3	3.8×10^3

17.6.4.2 Two-Dimensional Example

The following estimates are also based on 1,000,000 samples for importance sampling and 20,000 samples for splitting. We use

$$\Lambda = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

and the escape set

$$B \doteq \{(x, y) \in \mathbb{R}^2 : 2x^2 \geq 1 \text{ or } 2y^2 \geq 1\}.$$

We still use the quasipotential to define the subsolution for this example. The estimates are based on 1,000,000 samples for importance sampling, while 20,000 samples are used for splitting.

We provide estimates of the relative errors for various combinations of n and T for both algorithms, keeping only two significant digits, in Tables 17.19 and 17.20 A result of – indicates that we were unable to estimate the quantity because no "successes" occurred in our sample.

Tables 17.21 and 17.22 provide estimates of the required work for a relative error of 1, keeping only two significant digits.

Table 17.19 Relative errors for importance sampling, two-dimensional problem

n/T	0.25	0.5	1	1.5	2.5	10	14	18
9	26	7.7	5.3	9.9	15	66	580	83
11	110	9.6	5.1	6.5	18	56	440	68
13	94	14	5.0	5.8	15	130	120	120
15	360	17	5.5	5.8	15	500	97	130
17	520	22	5.7	5.1	12	63	68	120

Table 17.20 Relative errors for RESTART, $R = 5$, two-dimensional problem

n/T	0.25	0.5	1	1.5	2.5	10	14	18
9	50	15	5.9	3.7	2.3	1.0	0.84	0.75
11	–	23	6.7	5.3	2.6	1.1	0.92	0.82
13	–	27	8.8	5.5	2.8	1.2	1.0	0.87
15	–	38	10	4.8	3.1	1.3	1.2	0.92
17	–	77	11	5.6	3.3	1.3	1.1	0.96

Table 17.21 Work required to obtain a relative error of 1 for importance sampling, two-dimensional problem

n/T	0.25	0.5	1	1.5	2.5	10	14	18
9	7.0×10^4	1.0×10^4	7.9×10^3	3.2×10^4	7.7×10^4	1.5×10^6	1.1×10^8	2.4×10^6
11	1.3×10^6	1.9×10^4	9.1×10^3	1.7×10^4	1.4×10^5	1.4×10^6	8.3×10^7	2.0×10^6
13	1.2×10^6	4.8×10^4	1.0×10^4	1.7×10^4	1.1×10^5	9.4×10^6	7.6×10^6	7.0×10^6
15	1.8×10^7	8.3×10^4	1.5×10^4	2.0×10^4	1.3×10^5	1.5×10^8	5.8×10^6	1.1×10^7
17	4.7×10^7	1.6×10^5	1.8×10^4	1.7×10^4	9.5×10^4	2.8×10^6	3.3×10^6	1.0×10^7

Table 17.22 Work required to obtain a relative error of 1 for RESTART, $R = 5$, two-dimensional problem

n/T	0.25	0.5	1	1.5	2.5	10	14	18
9	8.7×10^3	2.1×10^3	9.4×10^2	7.1×10^2	5.5×10^2	4.9×10^2	4.6×10^2	4.7×10^2
11	–	4.7×10^3	1.5×10^3	2.1×10^3	1.0×10^3	9.2×10^2	9.1×10^2	9.1×10^2
13	–	8.1×10^3	3.4×10^3	2.9×10^3	1.6×10^3	1.5×10^3	1.5×10^3	1.5×10^3
15	–	1.9×10^4	5.5×10^3	2.8×10^3	2.7×10^3	2.3×10^3	2.8×10^3	2.3×10^3
17	–	8.6×10^4	7.4×10^3	4.6×10^3	3.7×10^3	3.1×10^3	3.1×10^3	3.1×10^3

17.7 Notes

In addition to the examples presented here, subsolutions have been constructed for many other types of problems. Queueing and related stochastic networks of various types, which are important examples of problems with discontinuous (in the spatial variable) statistical behavior, have been analyzed in [76, 77] (splitting) and [103, 105, 110, 117] (importance sampling). The paper [77] points out the role of the quasipotential in constructing subsolutions for large-time or time-independent problems, and a recent paper that rigorously analyzes the performance of splitting when the interval of interest is allowed to grow with the large deviation parameter is [49]. Importance sampling for processes with multiple scales including homogenization and multiple time scales are considered in [112, 115]. Finally, we note that there are important classes of problems either for which the needed subsolutions are easily found (e.g., using the quasipotential to construct subsolutions for exit problems over long times for reversible systems), or for which they exist already as part of the literature on deterministic optimal control (e.g., for occupancy and related combinatorial models in probability [119]).

Appendix A
Spaces of Measures

A.1 Weak Convergence of Probability Measures

Throughout this section, $\mathscr{X}$ is a Polish space with metric $d(x, y)$. For certain results we will need $\mathscr{X}$ also to be locally compact. In such cases, this assumption will be made explicit. Let $\mathscr{P}(\mathscr{X})$ denote the space of probability measures on $\mathscr{X}$ and let $\mathscr{C}_b(\mathscr{X})$ denote the space of bounded continuous functions mapping $\mathscr{X}$ into $\mathbb{R}$. Consider a sequence $\{\theta_n\}_{n\in\mathbb{N}}$ in $\mathscr{P}(\mathscr{X})$. We say that $\{\theta_n\}$ **converges weakly** to θ, and write $\theta_n \Rightarrow \theta$, if for each $g \in \mathscr{C}_b(\mathscr{X})$,

$$\lim_{n\to\infty} \int_{\mathscr{X}} g \, d\theta_n = \int_{\mathscr{X}} g \, d\theta.$$

Then $\mathscr{P}(\mathscr{X})$ is made into a topological space by taking as the basic open neighborhoods of $\gamma \in \mathscr{P}(\mathscr{X})$ the sets of the form

$$\left\{\theta \in \mathscr{P}(\mathscr{X}) : \left| \int_{\mathscr{X}} g_i \, d\theta - \int_{\mathscr{X}} g_i \, d\gamma \right| < \varepsilon, i = 1, 2, \ldots, k \right\},$$

where $\varepsilon > 0$, k is a positive integer, and $g_1, g_2, \ldots, g_k$ are in $\mathscr{C}_b(\mathscr{X})$. The resulting topology is called the **topology of weak convergence** or simply the **weak topology**.

To introduce a metric on $\mathscr{P}(\mathscr{X})$, for $A \subset \mathscr{X}$ and $\varepsilon > 0$ we define

$$A^{(\varepsilon)} \doteq \{x \in \mathscr{X} : d(x, A) < \varepsilon\}.$$

For γ and θ in $\mathscr{P}(\mathscr{X})$, we then define

$$\mathscr{L}(\gamma, \theta) \doteq \inf\{\varepsilon > 0 : \gamma(F) \le \theta(F^{(\varepsilon)}) + \varepsilon \text{ for all closed subsets } F \text{ of } \mathscr{X}\}.$$

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0

Then $\mathscr{L}(\gamma, \theta)$ defines a metric on $\mathscr{P}(\mathscr{X})$, known as the **Lévy–Prohorov metric** [126, p. 96].

As we state in the next theorem, the Lévy–Prohorov metric is compatible with the weak topology, and with respect to it, $\mathscr{P}(\mathscr{X})$ is Polish.

Theorem A.1 ([126, pp. 101 and 108]) *Let $\{\theta_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathscr{P}(\mathscr{X})$. Then $\theta_n \Rightarrow \theta \in \mathscr{P}(\mathscr{X})$ if and only if $\mathscr{L}(\theta_n, \theta) \to 0$. Furthermore, with respect to the Lévy–Prohorov metric, $\mathscr{P}(\mathscr{X})$ is complete and separable.*

The next result, known as the Portmanteau theorem, gives a number of useful conditions that are equivalent to weak convergence. For $\theta \in \mathscr{P}(\mathscr{X})$, a Borel set A whose boundary ∂A satisfies $\theta(\partial A) = 0$ is called a θ-**continuity set**.

Theorem A.2 ((PORTMANTEAU THEOREM). [24, p. 11]) *Let $\{\theta_n\}$ and θ be probability measures on $\mathscr{X}$. The following five conditions are equivalent:*

(a) $\theta_n \Rightarrow \theta$.
(b) $\lim_{n\to\infty} \int_{\mathscr{X}} g\, d\theta_n = \int_{\mathscr{X}} g\, d\theta$ *for all bounded uniformly continuous functions g mapping $\mathscr{X}$ into $\mathbb{R}$.*
(c) $\limsup_{n\to\infty} \theta_n(F) \le \theta(F)$ *for all closed subsets F of $\mathscr{X}$.*
(d) $\liminf_{n\to\infty} \theta_n(G) \ge \theta(G)$ *for all open subsets G of $\mathscr{X}$.*
(e) $\lim_{n\to\infty} \theta_n(A) = \theta(A)$ *for all θ-continuity sets A.*

Remark A.3 The standard proof that (b) implies (c) uses a collection of Lipschitz continuous functions. Using this observation, we can augment the Portmanteau theorem with the following additional equivalent condition: $\theta_n \Rightarrow \theta$ if and only if

$$\lim_{n\to\infty} \int_{\mathscr{X}} g\, d\theta_n = \int_{\mathscr{X}} g\, d\theta$$

for all bounded Lipschitz continuous functions g mapping $\mathscr{X}$ into $\mathbb{R}$.

We next state Prohorov's theorem, which characterizes relatively compact subsets of $\mathscr{P}(\mathscr{X})$. It is one of the main results in the theory. A family Φ of probability measures on $\mathscr{X}$ is said to be **tight** if for each $\varepsilon > 0$, there exists a compact set K such that

$$\inf_{\gamma\in\Phi} \gamma(K) \ge 1 - \varepsilon.$$

Theorem A.4 ((PROHOROV'S THEOREM). [126, p. 103]) *A family of probability measures on $\mathscr{X}$ is relatively compact with respect to weak convergence if and only if it is tight. In particular, if $\theta_n \Rightarrow \theta$, then $\{\theta_n\}$ is tight.*

Prohorov's theorem yields the following useful fact.

Corollary A.5 *If $\mathscr{X}$ is a compact Polish space, then $\mathscr{P}(\mathscr{X})$ is compact.*

The notion of uniform integrability is useful in proving L^1 convergence.

Definition A.6 A sequence $\{\mu_n\}_{n\in\mathbb{N}}$ of probability measures on $\mathbb{R}^d$ (resp. a sequence $\{\xi_n\}$ of $\mathbb{R}^d$-valued random variables) is said to be **uniformly integrable** if

$$\sup_{n\in\mathbb{N}} \int_{\{x:\|x\|\geq a\}} \|x\| \mu_n(dx) \to 0 \text{ as } a \to \infty,$$

resp.

$$\sup_{n\in\mathbb{N}} E[\|\xi_n\| 1_{\{\|\xi_n\|\geq a\}}] \to 0 \text{ as } a \to \infty.$$

The following definitions are useful for characterizing limits of sequences of probability measures.

Definition A.7 Let $\mathscr{V}$ be a subset of the class of continuous and bounded functions on $\mathscr{X}$. We say that $\mathscr{V}$ is **separating** if for all $\mu, \nu \in \mathscr{P}(\mathscr{X})$, whenever $\int_{\mathscr{X}} g\, d\mu = \int_{\mathscr{X}} g\, d\nu$ for all $g \in \mathscr{V}$, we must have $\mu = \nu$. The class $\mathscr{V}$ is said to be **convergence determining** if for every sequence $\{\theta_n\} \subset \mathscr{P}(\mathscr{X})$ and $\theta \in \mathscr{P}(\mathscr{X})$, $\theta_n \Rightarrow \theta$ if and only if $\int_{\mathscr{X}} g\, d\theta_n \to \int_{\mathscr{X}} g\, d\theta$ for all $g \in \mathscr{V}$.

Note that if $\mathscr{V}$ is convergence determining, then it is separating. One of the basic results in the theory says that there exists a countable convergence determining class $\mathscr{V}$ of bounded uniformly continuous (in fact Lipschitz continuous) functions (cf. [126, Proposition 3.4.4]). Such a class can be given explicitly as follows. Let $\{x_n\}$ be a countable dense set in $\mathscr{X}$. For $i, j \in \mathbb{N}$, let $f_{ij}(x) \doteq 2(1 - jd(x, x_i)) \vee 0, x \in \mathscr{X}$. For a finite subset Λ of $\mathbb{N} \times \mathbb{N}$, let

$$g_\Lambda(x) \doteq \left(\sum_{(i,j)\in\Lambda} f_{ij}(x) \right) \wedge 1, \ x \in \mathscr{X}.$$

Then for each Λ, g_Λ is a bounded Lipschitz continuous function and $\{g_\Lambda : \Lambda \subset \mathbb{N} \times \mathbb{N}\}$ defines a countable convergence determining class. Indeed, suppose $\mu_n, \mu \in \mathscr{P}(\mathscr{X})$ satisfy $\int_{\mathscr{X}} g_\Lambda\, d\mu_n \to \int_{\mathscr{X}} g_\Lambda\, d\mu$ for every Λ, and let $G \subset \mathscr{X}$ be an open set. For $m \in \mathbb{N}$, define

$$\Lambda_m \doteq \{(i, j) : i, j \leq m \text{ and } B(x_i, 1/j) \subset G\}.$$

Then $h_m \doteq g_{\Lambda_m}$ satisfies $h_m \leq 1_G$, and $h_m \uparrow 1_G$ as $m \to \infty$. Thus

$$\liminf_{n\to\infty} \mu_n(G) \geq \lim_{n\to\infty} \int h_m d\mu_n = \int h_m d\mu.$$

Sending $m \to \infty$, we have $\liminf_{n\to\infty} \mu_n(G) \geq \mu(G)$. Since G is an arbitrary open set, we have $\mu_n \Rightarrow \mu$.

A.2 Skorohod Representation Theorem

For the proof of the following theorem we refer the reader to [167, Theorem 3.30].

Theorem A.8 (SKOROHOD REPRESENTATION THEOREM) *Suppose $\xi, \xi_1, \xi_2, \ldots$ are random variables with values in some separable metric space $(\mathscr{S}, \rho)$ such that $\xi_n \Rightarrow \xi$ as $n \to \infty$. Then on some probability space $(\Omega, \mathscr{F}, P)$, there exist $\mathscr{S}$-valued random variables $\eta, \eta_1, \eta_2, \ldots$ such that the distribution of ξ is the same as the distribution of η, the distribution of ξ_i is same as the distribution of η_i for every i, and $\eta_i \to \eta$ P-a.s.*

A.3 Space of Finite Measures

For a Polish space $\mathscr{X}$, $\mathscr{M}(\mathscr{X})$ denotes the set of finite measures on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$, i.e., the set of measures γ for which $\gamma(\mathscr{X}) < \infty$. Let $\{\theta_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathscr{M}(\mathscr{X})$. We say that $\{\theta_n\}$ **converges weakly** to θ if for each $g \in \mathscr{C}_b(\mathscr{X})$,

$$\lim_{n\to\infty} \int_{\mathscr{X}} g \, d\theta_n = \int_{\mathscr{X}} g \, d\theta.$$

We next introduce a metric on $\mathscr{M}(\mathscr{X})$ having properties analogous to those of $\mathscr{L}(\cdot,\cdot)$. For γ and θ in $\mathscr{M}(\mathscr{X})$, define

$$m(\gamma,\theta) \doteq \big[\gamma(\mathscr{X}) \wedge \theta(\mathscr{X})\big] \cdot \mathscr{L}\left(\frac{\gamma}{\gamma(\mathscr{X})}, \frac{\theta}{\theta(\mathscr{X})}\right) + |\gamma(\mathscr{X}) - \theta(\mathscr{X})|.$$

The convention is that if γ or θ equals the zero measure on $\mathscr{X}$, then the first term in this definition is 0. Properties of m are given in the next theorem. The straightforward proof is omitted.

Theorem A.9 *The quantity $m(\gamma,\theta)$ defines a metric on $\mathscr{M}(\mathscr{X})$. The weak convergence on $\mathscr{M}(\mathscr{X})$ is equivalent to convergence under this metric. With respect to this metric, $\mathscr{M}(\mathscr{X})$ is complete and separable.*

A.4 Space of Locally Finite Measures

For a locally compact Polish space $\mathscr{X}$, we denote by $\Sigma(\mathscr{X})$ the space of all measures ν on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$ satisfying $\nu(K) < \infty$ for every compact $K \subset \mathscr{X}$. We endow $\Sigma(\mathscr{X})$ with the weakest topology such that for every $f \in \mathscr{C}_c(\mathscr{X})$ (the space of real continuous functions on $\mathscr{X}$ with compact support), the function $\nu \mapsto \langle f, \nu\rangle =$

$\int_{\mathscr{X}} f(u)\,\nu(du)$, $\nu \in \Sigma(\mathscr{X})$ is continuous. This topology can be metrized such that $\Sigma(\mathscr{X})$ is a Polish space. One metric that is convenient for this purpose is given in the following subsection.

A.4.1 Metric for Measures on a Locally Compact Polish Space

Let $\|f\|_\infty \doteq \sup_{x\in\mathscr{X}} |f(x)|$ and $\|f\|_L \doteq \sup_{x,y\in\mathscr{X}} |f(x)-f(y)|/d(x,y)$, where d denotes the metric on $\mathscr{X}$. According to [220, Theorem 9.5.21], for a locally compact set $\mathscr{X}$ there exists a sequence of open sets $\{O_j\}$ such that $\bar{O}_j \subset O_{j+1}$, each $\bar{O}_j$ is compact, and $\cup_{j=1}^\infty O_j = \mathscr{X}$. Let $\phi_j(x) \doteq \left[1 - d(x, O_j)\right] \vee 0$. Given any $\mu \in \Sigma(\mathscr{X})$, let $\mu^j \in \Sigma(\mathscr{X})$ be defined by

$$\left[d\mu^j/d\mu\right](x) = \phi_j(x). \tag{A.1}$$

Given $\mu, \nu \in \Sigma(\mathscr{X})$, let

$$\bar{d}(\mu,\nu) \doteq \sum_{j=1}^{\infty} 2^{-j} \left\|\mu^j - \nu^j\right\|_{BL},$$

where $\|\cdot\|_{BL}$ denotes the bounded Lipschitz norm [92, Chap. 11]

$$\left\|\mu^j - \nu^j\right\|_{BL} \doteq \sup_{f:\|f\|_\infty\le 1, \|f\|_L\le 1} \left[\int_{\mathscr{X}} f d\mu^j - \int_{\mathscr{X}} f d\nu^j\right].$$

It is straightforward to check that $\bar{d}(\mu,\nu)$ defines a metric under which $\Sigma(\mathscr{X})$ is a Polish space, and that convergence in this metric is essentially equivalent to weak convergence on each compact subset of $\mathscr{X}$. Specifically, $\bar{d}(\mu_n,\mu) \to 0$ if and only if for each $j \in \mathbb{N}$, $\mu_n^j \to \mu^j$ in the weak topology as finite nonnegative measures, i.e., for all $f \in \mathscr{C}_b(\mathscr{X})$,

$$\int_{\mathscr{X}} f d\mu_n^j \to \int_{\mathscr{X}} f d\mu^j.$$

A.4.2 Determining Convergence from a Countable Class

From the separability of $\mathscr{X}$ it follows that the space $\mathscr{C}_c(\mathscr{X})$ is separable in the uniform metric, from which it follows (see [167, Appendix A.2]) that there is a countable collection $\mathscr{J} \subset \mathscr{C}_c(\mathscr{X})$ such that for $\mu_n, \mu \in \Sigma(\mathscr{X})$, $\bar{d}(\mu_n,\mu) \to 0$ if and only if for every $f \in \mathscr{J}$,

$$\int_{\mathscr{X}} f d\mu_n \to \int_{\mathscr{X}} f d\mu.$$

As an immediate consequence of this fact we have the following result.

Lemma A.10 *Suppose that* $\{N_k\}_{k\in\mathbb{N}}$ *and* N *are* $\Sigma(\mathscr{X})$*-valued random variables defined on a probability space* $(\Omega, \mathscr{F}, P)$ *and that* $E\,|\langle g, N_k\rangle - \langle g, N\rangle| \to 0$ *for all* $g \in \mathscr{C}_c(\mathscr{X})$*. Then* $N_k \to N$ *in probability.*

A.4.3 Proof of Compactness of S_m^N Defined in Chap. 8

Fix $T < \infty$. Let $\mathscr{X}$ be a locally compact Polish space and let $\mathscr{X}_T = [0, T] \times \mathscr{X}$. Recall the function L_T defined in (8.17):

$$L_T(g) \doteq \int_{\mathscr{X}_T} \ell(g(t,x))\, \nu_T(dt \times dx).$$

Here $\ell : [0, \infty) \to [0, \infty)$ is defined by

$$\ell(r) \doteq r \log r - r + 1,\ r \in [0, \infty),$$

with the convention that $0 \log 0 = 0$.

Let $\mathbb{M} \doteq \Sigma(\mathscr{X}_T)$. For $m \in \mathbb{N}$, define

$$S_m^N \doteq \{g : \mathscr{X}_T \to [0, \infty) : L_T(g) \le m\}.$$

A function $g \in S_m^N$ can be identified with a measure $\nu_T^g \in \mathbb{M}$ according to $\nu_T^g(A) = \int_A g(s,x)\, \nu_T(ds \times dx)$, $A \in \mathscr{B}(\mathscr{X}_T)$. The following lemma shows that $\left\{\nu_T^g : g \in S_m^N\right\}$ is a compact subset of $\mathbb{M}$.

Lemma A.11 *For every* $m \in \mathbb{N}$, $\{\nu_T^g : g \in S_m^N\}$ *is a compact subset of* $\mathbb{M}$.

Proof We note that the metric $\bar{d}$ on $\mathbb{M}$ introduced in Sect. A.4.1, when $\mathscr{X}$ is replaced with $\mathscr{X}_T$, will be given as follows. There is a sequence of open sets $\left\{O_j,\ j \in \mathbb{N}\right\}$ such that $\bar{O}_j \subset O_{j+1}$, each $\bar{O}_j$ is compact, and $\cup_{j=1}^{\infty} O_j = \mathscr{X}_T$. Also, for $(t,x) \in \mathscr{X}_T$, let $\phi_j(t,x) = \left[1 - d((t,x), O_j)\right] \vee 0$, where d denotes the metric on $\mathscr{X}_T$. Given any $\mu \in \mathbb{M}$, let $\mu^j \in \mathbb{M}$ be defined by $\left[d\mu^j / d\mu\right](t,x) = \phi_j(t,x)$. Then given $\mu, \nu \in \mathbb{M}$, we have

$$\bar{d}(\mu, \nu) \doteq \sum_{j=1}^{\infty} 2^{-j} \left\|\mu^j - \nu^j\right\|_{BL},$$

where $\|\cdot\|_{BL}$ denotes the bounded Lipschitz norm on $\mathscr{M}_F(\mathscr{X}_T)$. Note that for $\mu_n, \mu \in \mathbb{M}$, we have $\bar{d}(\mu_n, \mu) \to 0$ if and only if for each $i \in \mathbb{N}$,

$$\int_{\mathscr{X}_T} f d\mu_n^i \to \int_{\mathscr{X}_T} f d\mu^i$$

for every continuous and bounded function on $\mathscr{X}_T$, where μ_n^i is defined according to $\left[d\mu_n^i/d\mu_n\right](t,x) = \phi_i(t,x)$.

Let $\mu_n \doteq \nu_T^{g_n}$. We first show that $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathbb{M}$ is relatively compact for every sequence $\{g_n\}_{n\in\mathbb{N}} \subset S^N$. For this, using a diagonalization method, it suffices to show that $\{\mu_n^i\} \subset \mathbb{M}$ is relatively compact for every i. Next, since $\{\mu_n^i\}$ are supported on the compact subset K^i of $\mathscr{X}_T$ given by the closure of $\{(t,x) : \phi_i(t,x) \neq 0\}$, to show that $\{\mu_n^i\} \subset \mathbb{M}$ is relatively compact, it suffices to show that $\sup_n \mu_n^i(\mathscr{X}_T) < \infty$. The last property will follow from the fact that $L_T(g_n) \leq m$ for all n and the superlinear growth of ℓ. Specifically, let $c \in (0,\infty)$ be such that $z \leq c(\ell(z)+1)$ for all $z \in [0,\infty)$. Then

$$\begin{aligned}\sup_{n\in\mathbb{N}} \mu_n^i(\mathscr{X}_T) &= \sup_{n\in\mathbb{N}} \int_{\mathscr{X}_T} \phi_i(t,x) g_n(t,x) \nu_T(dt\times dx) \\ &\leq \sup_{n\in\mathbb{N}} \int_{K^i} g_n(t,x)\nu_T(dt\times dx) \leq c(m+\nu_T(K^i)) < \infty.\end{aligned}$$

Next, suppose that along a subsequence (without loss of generality, also denoted by $\{n\}$), $\mu_n \to \mu$. We would like to show that μ is of the form ν_T^g, where $g \in S_m^N$. For this, we will use the lower semicontinuity property of relative entropy. The result holds trivially if $\mu = 0$. Suppose now $\mu \neq 0$. Then there exists $i_0 \in \mathbb{N}$ such that for all $i \geq i_0$, $\inf_{n\in\mathbb{N}} \nu_T^{g_n}(\bar{O}_i) > 0$. Introducing a slight notational inconsistency, let ν_T^i be defined by $\left[d\nu_T^i/d\nu_T\right](t,x) = \phi_i(t,x)$. For $i \geq i_0$, define

$$\begin{aligned} c^i &= \nu_T^i(\mathscr{X}_T), \quad \bar{\nu}_T^i = \nu_T^i/c^i, \\ b_n^i &= \mu_n^i(\mathscr{X}_T), \quad \bar{\mu}_n^i = \mu_n^i/b_n^i, \\ b^i &= \mu^i(\mathscr{X}_T), \quad \bar{\mu}^i = \mu^i/b^i. \end{aligned}$$

Then $\bar{\nu}_T^i$, $\bar{\mu}_n^i$, and $\bar{\mu}^i$ are probability measures, and

$$\begin{aligned} R(\bar{\mu}_n^i||\bar{\nu}_T^i) &= \frac{1}{b_n^i}\int_{\mathscr{X}_T}\left[\log(g_n(t,x)) + \log\left(\frac{c^i}{b_n^i}\right)\right] g_n(t,x)\phi_i(t,x)\nu_T(dt\times dx) \\ &= \frac{1}{b_n^i}\int_{\mathscr{X}_T}\left[\ell(g_n(t,x)) + g_n(t,x) - 1\right]\phi_i(t,x)\nu_T(dt\times dx) + \log\left(\frac{c^i}{b_n^i}\right) \\ &\leq \frac{1}{b_n^i}m + 1 - \frac{c^i}{b_n^i} + \log\left(\frac{c^i}{b_n^i}\right). \end{aligned}$$

Since $\mu_n^i \to \mu^i$, we have $b_n^i \to b^i$. Hence by the lower semicontinuity property of relative entropy,

$$\begin{aligned} R(\bar{\mu}^i \| \bar{\nu}_T^i) &\le \liminf_{n\to\infty} R(\bar{\mu}_n^i \| \bar{\nu}_T^i) \\ &\le \liminf_{n\to\infty} \left[\frac{1}{b_n^i} m + 1 - \frac{c^i}{b_n^i} + \log\left(\frac{c^i}{b_n^i}\right) \right] \\ &\le \frac{1}{b^i} m + 1 - \frac{c^i}{b^i} + \log\left(\frac{c^i}{b^i}\right) \\ &< \infty. \end{aligned} \tag{A.2}$$

Thus μ^i is absolutely continuous with respect to ν_T^i. Define $g^i = d\mu^i / d\nu_T^i$ and $g = g^i$ on $\bar{O}_i$. It is easily checked that g is defined consistently and that $\mu = \nu_T^g$. Also, by a direct calculation,

$$R(\bar{\mu}^i \| \bar{\nu}_T^i) = \frac{1}{b^i} \int_{\mathscr{X}_T} \ell(g(t,x)) \phi_i(t,x) \nu_T(dt \times dx) + 1 - \frac{c^i}{b^i} + \log\left(\frac{c^i}{b^i}\right).$$

Combining the last display with (A.2), we have $\int_{\mathscr{X}_T} \ell(g(t,x)) \phi_i(t,x) \nu_T(dt \times dx) \le m$ for all i. Sending $i \to \infty$, we see that $g \in S_m^N$. The result follows. □

Appendix B
Stochastic Kernels

B.1 Regular Conditional Probabilities

Throughout this section, $(\Omega, \mathscr{F}, P)$ is a probability space and $\mathscr{Y}$ is a Polish space. Let Y be a random variable mapping Ω into $\mathscr{Y}$ and $\mathscr{G}$ a sub-σ-field of $\mathscr{F}$. A **regular conditional distribution for Y given** $\mathscr{G}$ is a map $(\omega, A) \mapsto \hat{P}(A|\mathscr{G})(\omega)$ from $\Omega \times \mathscr{Y}$ to $[0, 1]$ with the following properties.

(a) For each Borel subset B of $\mathscr{Y}$, the function mapping $\omega \in \Omega \mapsto \hat{P}(B|\mathscr{G})(\omega)$ is measurable with respect to $\mathscr{G}$.

(b) For each $\Gamma \in \mathscr{G}$ and Borel subset B of $\mathscr{Y}$,

$$P\{\Gamma \cap \{Y \in B\}\} = \int_{\Gamma} \hat{P}(B|\mathscr{G})(\omega)\, P(d\omega).$$

(c) For each $\omega \in \Omega$, $\hat{P}(dy|\mathscr{G})(\omega)$ is a probability measure on $\mathscr{Y}$.

The first two properties state that $\hat{P}(B|\mathscr{G})(\omega)$ is a version of the conditional probability $P\{Y \in B|\mathscr{G}\}(\omega)$ for every Borel set B. The issue in proving the existence of a regular conditional distribution is to show that a version of the conditional probability can be chosen to be a probability measure on $\mathscr{Y}$ for each fixed value of ω. According to part (a) of the following standard result, this is possible when $\mathscr{Y}$ is a Polish space. Part (b) states the useful property that P-a.s. for $\omega \in \Omega$, conditional expectations can be obtained by integrating with respect to regular conditional distributions. The theorem is proved in Theorems 10.2.2 and 10.2.5 in [92]. If X is a random variable mapping $(\Omega, \mathscr{F})$ into a measurable space $(\mathscr{V}, \mathscr{A})$ and $\mathscr{G}$ denotes the sub-σ-algebra of $\mathscr{F}$ generated by X, then we will write $\hat{P}(dy|X)(\omega)$ for $\hat{P}(dy|\mathscr{G})(\omega)$ and call it a **regular conditional distribution for Y given X**.

Theorem B.1 *Let Y be a random variable mapping Ω into $\mathscr{Y}$, $\mathscr{G}$ a sub-σ-algebra of $\mathscr{F}$, and f a measurable function mapping $\mathscr{Y}$ into $\mathbb{R}$ such that $E\{|f(Y)|\} < \infty$. The following conclusions hold.*

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0

(a) *A regular conditional distribution* $\hat{P}(dy|\mathscr{G})(\omega)$ *for* Y *given* $\mathscr{G}$ *exists. It is unique in the sense that if* $\hat{Q}(dy|\mathscr{G})(\omega)$ *also satisfies the definition, then the two distributions* $\hat{P}(dy|\mathscr{G})(\omega)$ *and* $\hat{Q}(dy|\mathscr{G})(\omega)$ *agree* P*-a.s. for* $\omega \in \Omega$.

(b) P*-a.s. for* $\omega \in \Omega$, f *is integrable with respect to* $\hat{P}(dy|\mathscr{G})(\omega)$ *and*

$$E\left[f(Y)|\mathscr{G}\right](\omega) = \int_{\mathscr{Y}} f(y)\, \hat{P}(dy|\mathscr{G})(\omega).$$

Now let $(\mathscr{V}, \mathscr{A})$ be a measurable space and X a random variable mapping Ω into $\mathscr{V}$. A **regular conditional distribution for** Y **given** $X = x$ is defined to be a quantity $\hat{P}(dy|X = x)$ taking values in [0, 1] and having the following properties.

(a) For each Borel subset B of $\mathscr{Y}$, the function mapping $x \in \mathscr{X} \mapsto \hat{P}(B|X = x)$ is measurable.

(b) For each measurable subset A of $(\mathscr{V}, \mathscr{A})$ and Borel subset B of $\mathscr{Y}$,

$$P\{\{X \in A\} \cap \{Y \in B\}\} = \int_A \hat{P}(B|X = x)\, P\{X \in dx\}.$$

(c) For each $x \in \mathscr{X}$, $\hat{P}(dy|X = x)$ is a probability measure on $\mathscr{Y}$.

The first two properties state that $\hat{P}(B|X = x)$ is a version of the conditional probability, in that if $g(x) = \hat{P}(B|X = x)$, then $g(X) = P\{Y \in B|X\}$ a.s. The next result is an immediate consequence of Theorem B.1.

Theorem B.2 *Let* $(\mathscr{V}, \mathscr{A})$ *be a measurable space,* X *a random variable mapping* Ω *into* $\mathscr{V}$*, and* Y *a random variable mapping* Ω *into* $\mathscr{Y}$*. Then a regular conditional distribution* $\hat{P}(dy|X = x)$ *for* Y *given* $X = x$ *exists. It is unique in the sense that if* $\hat{Q}(dy|X = x)$ *also satisfies the definition, then for almost every* x*, the two measures* $\hat{P}(dy|X = x)$ *and* $\hat{Q}(dy|X = x)$ *agree with respect to the distribution of* X.

Proof A regular conditional distribution $\hat{P}(dy|X)$ is measurable with respect to the sub-σ-field generated by X, and so it is a measurable function of X, say $\varphi(X)$. The quantity $\varphi(x)$ is a regular conditional distribution for Y given $X = x$. The uniqueness follows from the uniqueness asserted in part (a) of Theorem B.1. □

B.2 Stochastic Kernels

Throughout this section, $\mathscr{X}$ and $\mathscr{Y}$ are Polish spaces and $(\mathscr{V}, \mathscr{A})$ is a measurable space. Let us recall the definition of a stochastic kernel, which was introduced in Sect. 1.4. Let $\tau(dy|x)$ be a family of probability measures on $\mathscr{Y}$ parametrized by $x \in \mathscr{V}$. We call $\tau(dy|x)$ a **stochastic kernel** on $\mathscr{Y}$ given $\mathscr{V}$ if for every Borel subset E of $\mathscr{Y}$, the function mapping $x \in \mathscr{V} \mapsto \tau(E|x) \in [0, 1]$ is measurable.

In order to establish a useful equivalent condition for a stochastic kernel, we need a preliminary fact given in the next lemma. It is a special case of Proposition 7.25 in [19].

Lemma B.3 *For $E \in \mathscr{B}(\mathscr{Y})$, define $f_E : \mathscr{P}(\mathscr{Y}) \to [0,1]$ by $f_E(\theta) \doteq \theta(E)$. Then*

$$\mathscr{B}(\mathscr{P}(\mathscr{Y})) = \sigma\left[\bigcup_{E \in \mathscr{B}(\mathscr{Y})} f_E^{-1}(\mathscr{B}(\mathbb{R}))\right].$$

In other words, $\mathscr{B}(\mathscr{P}(\mathscr{Y}))$ is the smallest σ-algebra with respect to which f_E is measurable for every $E \in \mathscr{B}(\mathscr{Y})$.

Proof We write $\mathscr{G} \doteq \sigma[\cup_{E \in \mathscr{B}(\mathscr{Y})} f_E^{-1}(\mathscr{B}(\mathbb{R}))]$. To prove that $\mathscr{G} \subset \mathscr{B}(\mathscr{P}(\mathscr{Y}))$, we show that f_E is $\mathscr{B}(\mathscr{P}(\mathscr{Y}))$-measurable for every $E \in \mathscr{B}(\mathscr{Y})$, so that for every $A \in \mathscr{B}(\mathbb{R})$, we have $f_E^{-1}(A) \in \mathscr{B}(\mathscr{P}(\mathscr{Y}))$. Let

$$\mathscr{D} \doteq \{E \in \mathscr{B}_{\mathscr{Y}} : f_E \text{ is } \mathscr{B}(\mathscr{P}(\mathscr{Y}))\text{-measurable}\}.$$

For every closed set $F \in \mathscr{B}(\mathscr{Y})$ and real number α, the Portmanteau theorem (Theorem A.2) implies that the set $\{\theta \in \mathscr{P}(\mathscr{Y}) : \theta(F) \geq \alpha\}$ is closed. Hence $F \in \mathscr{D}$. It is now straightforward to verify using the Dynkin class theorem [126] that $\mathscr{D}$ equals $\mathscr{B}(\mathscr{Y})$ and thus that $\mathscr{G} \subset \mathscr{B}(\mathscr{P}(\mathscr{Y}))$. The proof that $\mathscr{B}(\mathscr{P}(\mathscr{Y})) \subset \mathscr{G}$ is based on a standard approximation argument. By definition of $\mathscr{G}$, the function

$$\alpha_\varphi : \theta \in \mathscr{P}(\mathscr{Y}) \longmapsto \int_{\mathscr{Y}} \varphi \, d\theta \in \mathbb{R}$$

is $\mathscr{G}$-measurable when $\varphi \doteq 1_E$ for every $E \in \mathscr{B}(\mathscr{Y})$; indeed, in this case, $\alpha_\varphi(\theta) = f_E(\theta)$. Thus α_φ is $\mathscr{G}$-measurable when φ is a $\mathscr{B}(\mathscr{Y})$-simple function. Since when $\varphi \in \mathscr{C}_b(\mathscr{Y})$ there exists a sequence of $\mathscr{B}(\mathscr{Y})$-simple functions $\{\varphi_n\}$ that are uniformly bounded below and satisfy $\varphi_n \uparrow \varphi$, the monotone convergence theorem implies that $\alpha_{\varphi_n} \uparrow \alpha_\varphi$. Thus α_φ is $\mathscr{G}$-measurable for every $\varphi \in \mathscr{C}_b(\mathscr{Y})$. For $\gamma \in \mathscr{P}(\mathscr{Y})$, $\varphi \in \mathscr{C}_b(\mathscr{Y})$, and $\varepsilon > 0$, we define

$$N(\gamma, \varphi, \varepsilon) \doteq \left\{\theta \in \mathscr{P}(\mathscr{Y}) : \left| \int_{\mathscr{Y}} \varphi \, d\theta - \int_{\mathscr{Y}} \varphi \, d\gamma \right| < \varepsilon \right\}.$$

Since $N(\gamma, \varphi, \varepsilon) = \alpha_\varphi^{-1}(\int_{\mathscr{Y}} \varphi \, d\gamma - \varepsilon, \int_{\mathscr{Y}} \varphi \, d\gamma + \varepsilon)$, it follows that $N(\gamma, \varphi, \varepsilon)$ is an element of $\mathscr{G}$, and since the class of sets $\{N(\gamma, \varphi, \varepsilon)\}$ forms an open subbase for $\mathscr{B}(\mathscr{P}(\mathscr{Y}))$, we conclude that $\mathscr{B}(\mathscr{P}(\mathscr{Y})) \subset \mathscr{G}$. This completes the proof of the lemma. □

The following result, taken from Proposition 7.26 in [19], gives a useful equivalent condition for a stochastic kernel. In the latter reference it is assumed that $(\mathscr{V}, \mathscr{A})$ is a Borel space. However, the proof applies without change when $(\mathscr{V}, \mathscr{A})$ is a measurable space.

Theorem B.4 *Let $\tau(dy|x)$ be a family of probability measures on $\mathscr{Y}$ parametrized by $x \in \mathscr{V}$. Then $\tau(dy|x)$ is a stochastic kernel if and only if the function mapping*

$x \in \mathscr{V} \longmapsto \tau(\cdot|x) \in \mathscr{P}(\mathscr{Y})$ is measurable, i.e., if and only if $\tau(\cdot|x)$ is a random variable mapping $\mathscr{V}$ into $\mathscr{P}(\mathscr{Y})$.

Proof We define $g : \mathscr{V} \to \mathscr{P}(\mathscr{Y})$ by $g(x) \doteq \tau(\cdot|x)$, and for $E \in \mathscr{B}(\mathscr{Y})$, we define $h_E : \mathscr{V} \to [0, 1]$ by $h_E(x) \doteq \tau(E|x)$. For $E \in \mathscr{B}(\mathscr{Y})$ we also recall $f_E : \mathscr{P}(\mathscr{Y}) \to [0, 1]$ defined in the previous lemma by $f_E(\theta) \doteq \theta(E)$. These mappings are related by $h_E = f_E \circ g$. The assertion of the theorem is that g is $\mathscr{A}$-measurable if and only if h_E is $\mathscr{A}$-measurable for every $E \in \mathscr{B}(\mathscr{Y})$. Lemma B.3 implies that f_E is $\mathscr{B}(\mathscr{P}(\mathscr{Y}))$-measurable for every $E \in \mathscr{B}(\mathscr{Y})$. Since $h_E = f_E \circ g$, it follows that if g is $\mathscr{A}$-measurable, then h_E is $\mathscr{A}$-measurable for every $E \in \mathscr{B}(\mathscr{Y})$. Conversely, if h_E is $\mathscr{A}$-measurable for every $E \in \mathscr{B}(\mathscr{Y})$, then again by Lemma B.3,

$$
\begin{aligned}
g^{-1}\left(\mathscr{B}(\mathscr{P}(\mathscr{Y}))\right) &= g^{-1}\left(\sigma\left[\bigcup_{E\in\mathscr{B}(\mathscr{Y})} f_E^{-1}(\mathscr{B}(\mathbb{R}))\right]\right) \\
&= \sigma\left[\bigcup_{E\in\mathscr{B}(\mathscr{Y})} g^{-1}\left(f_E^{-1}(\mathscr{B}(\mathbb{R}))\right)\right] \\
&= \sigma\left[\bigcup_{E\in\mathscr{B}(\mathscr{Y})} h_E^{-1}(\mathscr{B}(\mathbb{R}))\right] \subset \mathscr{A}.
\end{aligned}
$$

We conclude that g is $\mathscr{A}$-measurable. This completes the proof. □

B.3 A Stochastic Kernel Needed in Sect. 4.8.4

In Sect. 4.8.4, it was required that we find stochastic kernels $\gamma^i, i = 1, 2$, on $\mathbb{R}^d$ given $\mathbb{R}^d \times \mathbb{R}^d$ and on $\mathbb{R}^d$ given $\mathbb{R}^d$, respectively, such that for all $(\xi, \beta^1) \in \mathbb{R}^d \times \mathbb{R}^d$ and $\beta^2 \in \mathbb{R}^d$,

$$
R\left(\gamma^1(\cdot|\xi, \beta^1) \,\|\theta(\cdot|\xi)\right) = L(\xi, \beta^1) \quad \text{and} \quad \int_{\mathbb{R}^d} y\gamma^1(dy|\xi, \beta^1) = \beta^1
$$

and

$$
R\left(\gamma^2(\cdot|\beta^2) \,\|\rho_\sigma(\cdot)\right) = \frac{1}{2\sigma^2}\|\beta^2\|^2 \quad \text{and} \quad \int_{\mathbb{R}^d} y\gamma^2(dy|\beta^2) = \beta^2.
$$

While part (g) of Lemma 4.16 can be directly applied to obtain γ^2, it does not directly give γ^1, since we may have $L(\xi, \beta^1) = \infty$ for some (ξ, β^1). Instead, we will mollify and obtain γ^1 as a limit. To simplify notation, we replace (ξ, β^1) by (x, β).

We first note that $L(x, \beta)$ is a lower semicontinuous function of $(x, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$, and thus is measurable on $\mathbb{R}^d \times \mathbb{R}^d$. In particular, $\{(x, \beta) : L(x, \beta) = \infty\}$ is measurable. Fix a sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ in $(0, 1)$ that converges to 0, and for $n \in \mathbb{N}$,

define

$$H^n(x,\alpha) \doteq \log \int_{\mathbb{R}^d} \exp\langle \alpha, y\rangle \ \theta^n(dy|x),$$
$$\tilde{H}^n(x,\alpha) \doteq \log \int_{\mathbb{R}^d} \exp\langle \alpha, y\rangle \ [\theta(dy|x) + \varepsilon_n \rho_1(dy)],$$

where

$$\theta^n(dy|x) \doteq \frac{1}{1+\varepsilon_n}[\theta(dy|x) + \varepsilon_n \rho_1(dy)]$$

and ρ_1 is the d-dimensional Gaussian distribution with covariance I. For $n \in \mathbb{N}$ and x, α, and β in $\mathbb{R}^d$, we also introduce

$$L^n(x,\beta) \doteq \sup_{\alpha\in\mathbb{R}^d} \left[\langle \alpha, \beta\rangle - H^n(x,\alpha)\right].$$

According to part (g) of Lemma 4.16, for each $n \in \mathbb{N}$, there is a measurable $\alpha^n(x,\beta)$ such that with

$$\bar{\gamma}^n(dy|x,\beta) \doteq e^{\langle \alpha^n(x,\beta), y\rangle - H^n(x,\alpha^n(x,\beta))}\theta^n(dy|x),$$

we have

$$R\left(\bar{\gamma}^n(\cdot|x,\beta) \,\|\, \theta^n(\cdot|x)\right) = L^n(x,\beta) \quad \text{and} \quad \int_{\mathbb{R}^d} y\bar{\gamma}^n(dy|x,\beta) = \beta. \tag{B.1}$$

Since $H^n(x,\alpha) = \tilde{H}^n(x,\alpha) - \log(1+\varepsilon_n)$ and $\tilde{H}^n(x,\alpha) \geq H(x,\alpha)$, it follows that if $L(x,\beta) < \infty$, then for all $n \in \mathbb{N}$,

$$L^n(x,\beta) = \sup_{\alpha\in\mathbb{R}^d} \left[\langle \alpha, \beta\rangle - \tilde{H}^n(x,\beta)\right] + \log(1+\varepsilon_n) \leq L(x,\beta) + \log 2 < \infty. \tag{B.2}$$

Also, by Condition 4.3, for each x and α in $\mathbb{R}^d$,

$$\sup_{n\in\mathbb{N}} H^n(x,\alpha) < \infty. \tag{B.3}$$

For x and β in $\mathbb{R}^d$, define

$$\gamma^n(dy|x,\beta) \doteq \begin{cases} \bar{\gamma}^n(dy|x,\beta) & \text{if } L(x,\beta) < \infty, \\ \delta_\beta(dy) & \text{if } L(x,\beta) = \infty. \end{cases}$$

Then for each $n \in \mathbb{N}$, γ^n is a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d \times \mathbb{R}^d$. We will show that for (x,β) such that $L(x,\beta) < \infty$, $\gamma^n(dy|x,\beta)$ converges to $\gamma(dy|x,\beta)$ that satisfies

$$R\left(\gamma(\cdot|x,\beta)\,\|\theta(\cdot|x)\right) = L(x,\beta) \quad \text{and} \quad \int_{\mathbb{R}^d} y\gamma(dy|x,\beta) = \beta, \tag{B.4}$$

which will complete the proof.

We therefore consider such (x,β). Suppose that n indexes a subsequence. It follows from (B.2), (B.1), and (B.3) that the relative entropies $R\left(\bar{\gamma}^n(\cdot|x,\beta)\,\|\theta^n(\cdot|x)\right)$ are uniformly bounded and that the moment-generating functions $H^n(x,\alpha)$ are uniformly bounded for each fixed α and x. The proof of Lemma 3.9 considered an analogous situation but without the n-dependence of the moment-generating function. However, with the uniform bound (B.3), the argument applies with only notational changes, and it shows that $\{\gamma^n(\cdot|x,\beta)\}_{n\in\mathbb{N}}$ is tight and uniformly integrable. Thus by letting n index a convergent subsubsequence with limit $\gamma(\cdot|x,\beta)$, we have

$$\int_{\mathbb{R}^d} y\gamma^n(dy|x,\beta) \to \int_{\mathbb{R}^d} y\gamma(dy|x,\beta).$$

It then follows from the lower semicontinuity of relative entropy (Lemma 2.4) that

$$\begin{aligned} L(x,\beta) &\le R\left(\gamma(\cdot|x,\beta)\,\|\theta(\cdot|x)\right) \\ &\le \liminf_{n\to\infty} R\left(\bar{\gamma}^n(\cdot|x,\beta)\,\middle\|\,\theta^n(\cdot|x)\right) \\ &= \liminf_{n\to\infty} L^n(x,\beta) \\ &\le L(x,\beta), \end{aligned}$$

where the last line is due to the fact that $L^n(x,\beta) \le L(x,\beta) - \log(1+\varepsilon_n)$ for all $n \in \mathbb{N}$. According to Lemma 2.4, $R(\cdot\,\|\cdot)$ is strictly convex in the first variable, which shows that $\gamma(\cdot|x,\beta)$ is the unique probability measure that satisfies (B.4). An argument by contradiction then shows that $\gamma^n(\cdot|x,\beta)$ converges to $\gamma(\cdot|x,\beta)$ along the entire sequence $n \in \mathbb{N}$. Since this is true for every (x,β) such that $L(x,\beta) < \infty$, this completes the proof. □

Appendix C
Further Properties of Relative Entropy

C.1 Proof of Part (e) of Lemma 2.4

We denote by Π the class of all finite measurable partitions of the Polish space $\mathscr{X}$. Part (e) of Lemma 2.4 states that for each γ and θ in $\mathscr{P}(\mathscr{X})$,

$$R(\gamma\|\theta) = \sup_{\pi\in\Pi} \sum_{A\in\pi} \gamma(A) \log \frac{\gamma(A)}{\theta(A)}, \tag{C.1}$$

where the summand equals 0 if $\gamma(A) = 0$ and equals ∞ if $\gamma(A) > 0$ and $\theta(A) = 0$. In addition, if A is any Borel subset of $\mathscr{X}$, then

$$R(\gamma\|\theta) \geq \gamma(A) \log \frac{\gamma(A)}{\theta(A)} - 1. \tag{C.2}$$

We first prove that for every finite measurable partition π of $\mathscr{X}$,

$$R(\gamma\|\theta) \geq \sum_{A\in\pi} \gamma(A) \log \frac{\gamma(A)}{\theta(A)}.$$

If $R(\gamma\|\theta) = \infty$, there is nothing to prove, so we assume that $R(\gamma\|\theta) < \infty$. In this case, $\gamma \ll \theta$, and setting $B \doteq \cup_{\{A\in\pi:\gamma(A)=0\}} A$, we define for $m \in \mathbb{N}$ the bounded measurable function

$$\psi_m(x) \doteq \sum_{\{A\in\pi:\gamma(A)>0\}} \left(\log \frac{\gamma(A)}{\theta(A)}\right) 1_A(x) - m\, 1_B(x).$$

The Donsker–Varadhan variational formula stated in part (a) of Lemma 2.1 implies that

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0

$$R(\gamma\|\theta) \geq \int_{\mathscr{X}} \psi_m \, d\gamma - \log \int_{\mathscr{X}} e^{\psi_m} \, d\theta$$
$$= \sum_{\{A\in\pi : \gamma(A)>0\}} \gamma(A) \, \log \frac{\gamma(A)}{\theta(A)} - \log(1 + e^{-m}\theta(B)).$$

This yields the desired formula, since $\lim_{m\to\infty} \log(1 + e^{-m}\theta(B)) = 0$.

In order to complete the proof of equation (C.1), we determine a sequence $\{\pi_n\}_{n\in\mathbb{N}}$ of finite measurable partitions of $\mathscr{X}$ having the property that

$$R(\gamma\|\theta) = \lim_{n\to\infty} \sum_{A\in\pi_n} \gamma(A) \, \log \frac{\gamma(A)}{\theta(A)}. \tag{C.3}$$

We carry this out via a standard technique, using unpublished notes of Barron [15]. If γ is not absolutely continuous with respect to θ, then the proof is straightforward. Indeed, in this case there exists a Borel subset A of $\mathscr{X}$ having the property that $\theta(A) = 0$ and $\gamma(A) > 0$. We obtain formula (C.3) by setting $\pi_n \doteq \{A, A^c\}$ for each $n \in \mathbb{N}$.

We now suppose that γ is absolutely continuous with respect to θ and let $f \doteq d\gamma/d\theta$. For each $n \in \mathbb{N}$, we then define π_n to be the finite measurable partition of $\mathscr{X}$ consisting of the disjoint Borel sets

$$A_{n,k} \doteq \begin{cases} \{x \in \mathscr{X} : \log f(x) \leq -\sqrt{n}\} & \text{if } k = -n, \\ \left\{x \in \mathscr{X} : \frac{k-1}{\sqrt{n}} < \log f(x) \leq \frac{k}{\sqrt{n}}\right\} & \text{if } k \in \{-n+1, -n+2, \ldots, n-1, n\}, \\ \{x \in \mathscr{X} : \log f(x) > \sqrt{n}\} & \text{if } k = n+1. \end{cases}$$

For $-n + 1 \leq k \leq n + 1$,

$$\gamma(A_{n,k}) = \int_{A_{n,k}} \exp(\log f) \, d\theta \geq \exp\left[\frac{k-1}{\sqrt{n}}\right] \theta(A_{n,k}). \tag{C.4}$$

The error in the approximation of the relative entropy by the sum over the partition π_n equals

$$R(\gamma\|\theta) - \sum_{A\in\pi_n} \gamma(A) \, \log \frac{\gamma(A)}{\theta(A)} = \sum_{k=-n}^{n+1} 1_{\{j:\gamma(A_{n,j})>0\}}(k) \int_{A_{n,k}} \log\left(f \, \frac{\theta(A_{n,k})}{\gamma(A_{n,k})}\right) d\gamma.$$

We now bound each term in this sum. For $-n + 1 \leq k \leq n$ and $x \in A_{n,k}$, if $\gamma(A_{n,k}) > 0$, then from (C.4), the integrand satisfies

$$\log\left(f(x) \, \frac{\theta(A_{n,k})}{\gamma(A_{n,k})}\right) \leq \frac{k}{\sqrt{n}} - \frac{k-1}{\sqrt{n}} = \frac{1}{\sqrt{n}},$$

which implies that

$$\sum_{k=-n+1}^{n} 1_{\{j:\gamma(A_{n,j})>0\}}(k) \int_{A_{n,k}} \log\left(f \, \frac{\theta(A_{n,k})}{\gamma(A_{n,k})} \right) d\gamma \le \frac{1}{\sqrt{n}} \sum_{k=-n+1}^{n} \gamma(A_{n,k}) \le \frac{1}{\sqrt{n}}.$$

For $k = -n$ and $x \in A_{n,-n}$, if $\gamma(A_{n,-n}) > 0$, then the integrand satisfies

$$\log\left(f(x) \, \frac{\theta(A_{n,-n})}{\gamma(A_{n,-n})} \right) \le \log\left(e^{-\sqrt{n}} \frac{1}{\gamma(A_{n,-n})} \right),$$

and so, since $s \log s \ge -e^{-1}$ for $s \in [0, \infty)$,

$$\int_{A_{n,-n}} \log\left(f \, \frac{\theta(A_{n,-n})}{\gamma(A_{n,-n})} \right) d\gamma \le -\gamma(A_{n,-n}) \log\left(e^{\sqrt{n}} \gamma(A_{n,-n}) \right) \le e^{-\sqrt{n}-1}.$$

Finally, for $k = n+1$, if $\gamma(A_{n,n+1}) > 0$, then (C.4) implies that $\theta(A_{n,n+1})/\gamma(A_{n,n+1}) \le 1$. Thus

$$\int_{A_{n,n+1}} \log\left(f \, \frac{\theta(A_{n,n+1})}{\gamma(A_{n,n+1})} \right) d\gamma \le \int_{A_{n,n+1}} (\log f) \, d\gamma = \int_{\{\log f > \sqrt{n}\}} (\log f) \, d\gamma.$$

Combining these inequalities yields

$$\begin{aligned} 0 \le R(\gamma \| \theta) &- \sum_{A \in \pi_n} \gamma(A) \, \log \frac{\gamma(A)}{\theta(A)} \\ &\le \frac{1}{\sqrt{n}} + e^{-\sqrt{n}-1} + \int_{\{\log f > \sqrt{n}\}} (\log f) \, d\gamma. \end{aligned}$$

If $R(\gamma \| \theta) < \infty$, then the integral in this inequality converges to 0 as $n \to \infty$, and thus

$$\lim_{n \to \infty} \sum_{A \in \pi_n} \gamma(A) \, \log \frac{\gamma(A)}{\theta(A)} = R(\gamma \| \theta).$$

Now assume that γ is absolutely continuous with respect to θ but that $R(\gamma \| \theta) = \infty$. For any Borel set B, if $\theta(B) = 0$, then $\gamma(B) = 0$ and $\gamma(B) \, \log[\gamma(B)/\theta(B)] = 0$, while if $\theta(B) > 0$, then since $s \, \log s \ge s - 1$ for $s \in [0, \infty)$, it follows that

$$\gamma(B) \log \frac{\gamma(B)}{\theta(B)} = \theta(B) \left[\frac{\gamma(B)}{\theta(B)} \, \log \frac{\gamma(B)}{\theta(B)} \right] \ge \theta(B) \left[\frac{\gamma(B)}{\theta(B)} - 1 \right] \ge -1. \quad \text{(C.5)}$$

Since $\{A_{n,k}, -n \le k \le n\}$ is a finite measurable partition of $\{\log f \le \sqrt{n}\}$, similar estimates as in the case $R(\gamma \| \theta) < \infty$ yield

$$\int_{\{\log f \le \sqrt{n}\}} (\log f) \, d\gamma - \sum_{k=-n}^{n} \gamma(A_{n,k}) \, \log \frac{\gamma(A_{n,k})}{\theta(A_{n,k})} \le \frac{1}{\sqrt{n}} + e^{-\sqrt{n}-1}.$$

Thus

$$\sum_{A\in\pi_n} \gamma(A)\,\log\frac{\gamma(A)}{\theta(A)} \geq \int_{\{\log f\leq\sqrt{n}\}} (\log f)\,d\gamma - 1 - \frac{1}{\sqrt{n}} - e^{-\sqrt{n}-1}.$$

Since the right-hand side converges to $\infty = R(\gamma\|\theta)$ as $n\to\infty$, we have completed the proof of (C.3) and thus the proof of (C.1).

We now prove formula (C.2). Given A a Borel subset of $\mathscr{X}$, (C.1) yields for the finite measurable partition $\pi \doteq \{A, A^c\}$,

$$R(\gamma\|\theta) \geq \gamma(A)\,\log\frac{\gamma(A)}{\theta(A)} + \gamma(A^c)\,\log\frac{\gamma(A^c)}{\theta(A^c)}.$$

If $\theta(A^c)=0$, then the last term in this display equals either 0 or ∞ depending on whether $\gamma(A^c)$ equals 0 or is positive. In either case, formula (C.2) follows. On the other hand, if $\theta(A^c)>0$, then by (C.5),

$$R(\gamma\|\theta) \geq \gamma(A)\,\log\frac{\gamma(A)}{\theta(A)} - 1.$$

This is what we wanted to prove. The proof of part (e) of Lemma 2.4 is complete. □

C.2 Proof of Part (f) of Lemma 2.4

According to part (e) of Lemma 2.4,

$$\begin{aligned} R(\Delta_\psi\nu\|\Delta_\psi\mu) &= \sup_{\pi\in\Pi_\mathscr{Y}} \sum_{A\in\pi} \Delta_\psi\nu(A)\,\log\frac{\Delta_\psi\nu(A)}{\Delta_\psi\mu(A)} \\ &= \sup_{\pi\in\Pi_\mathscr{Y}} \sum_{A\in\pi} \nu(\psi^{-1}(A))\,\log\frac{\nu(\psi^{-1}(A))}{\mu(\psi^{-1}(A))}, \end{aligned}$$

where $\Pi_\mathscr{Y}$ denotes the class of all finite measurable partitions of $\mathscr{Y}$. For each $\pi\in\Pi_\mathscr{Y}$, we define $\psi^{-1}(\pi)\doteq\{\psi^{-1}(A): A\in\pi\}$, which is a finite measurable partition of $\mathscr{X}$. Thus, denoting by $\Pi_\mathscr{X}$ the class of all finite measurable partitions of $\mathscr{X}$, we have

$$\begin{aligned} R(\Delta_\psi\nu\|\Delta_\psi\mu) &= \sup_{\pi\in\Pi_\mathscr{Y}} \sum_{A\in\pi} \nu(\psi^{-1}(A))\,\log\frac{\nu(\psi^{-1}(A))}{\mu(\psi^{-1}(A))} \\ &\leq \sup_{\pi\in\Pi_\mathscr{X}} \sum_{A\in\pi} \nu(A)\,\log\frac{\nu(A)}{\mu(A)} \\ &= R(\nu\|\mu). \end{aligned}$$

This proves the first part of the lemma. Finally, when ψ is one-to-one and ψ^{-1} is measurable, each $\pi_* \in \Pi_{\mathscr{X}}$ has the form $\psi^{-1}(\pi)$ for some $\pi \in \Pi_{\mathscr{Y}}$. In such a case, the inequality in the above display can be replaced by an equality. This completes the proof. □

C.3 Proof of Proposition 2.3

To prove part (a), we note that since k is bounded from below, the right-hand side of equation (2.2) is well defined. Since for $N \in \mathbb{N}$, $k \wedge N$ is bounded and measurable, part (a) of Proposition 2.2 implies that

$$\begin{aligned} -\log \int_{\mathscr{X}} e^{-(k\wedge N)}\, d\theta &= \inf_{\gamma\in\mathscr{P}(\mathscr{X})} \left[R(\gamma\|\theta) + \int_{\mathscr{X}} (k\wedge N)\, d\gamma \right] \\ &\le \inf_{\gamma\in\mathscr{P}(\mathscr{X})} \left[R(\gamma\|\theta) + \int_{\mathscr{X}} k\, d\gamma \right]. \end{aligned}$$

Thus by the dominated convergence theorem,

$$-\log \int_{\mathscr{X}} e^{-k}\, d\theta = \lim_{N\to\infty} \left(-\log \int_{\mathscr{X}} e^{-(k\wedge N)}\, d\theta \right) \le \inf_{\gamma\in\mathscr{P}(\mathscr{X})} \left[R(\gamma\|\theta) + \int_{\mathscr{X}} k\, d\gamma \right].$$

In order to prove that

$$\inf_{\gamma\in\mathscr{P}(\mathscr{X})} \left[R(\gamma\|\theta) + \int_{\mathscr{X}} k\, d\gamma \right] \le -\log \int_{\mathscr{X}} e^{-k}\, d\theta,$$

we assume that $-\log \int_{\mathscr{X}} e^{-k}\, d\theta < \infty$, since otherwise, there is nothing to prove. Given $N \in \mathbb{N}$ and $\varepsilon > 0$, there exists a probability measure γ_N on $\mathscr{X}$ such that

$$\begin{aligned} R(\gamma_N\|\theta) + \int_{\mathscr{X}} (k\wedge N)\, d\gamma_N &\le \inf_{\gamma\in\mathscr{P}(\mathscr{X})} \left[R(\gamma\|\theta) + \int_{\mathscr{X}} (k\wedge N)\, d\gamma \right] + \varepsilon \\ &= -\log \int_{\mathscr{X}} e^{-(k\wedge N)}\, d\theta + \varepsilon \\ &\le -\log \int_{\mathscr{X}} e^{-k}\, d\theta + \varepsilon < \infty. \end{aligned}$$

Since k is bounded from below, it follows that $\sup_{N\in\mathbb{N}} R(\gamma_N\|\theta) < \infty$. This implies that the sequence $\{\gamma_N\}_{N\in\mathbb{N}}$ is relatively compact with respect to the weak topology [part (c) of Lemma 2.4]. Moreover, if γ_N converges along a subsequence to $\bar{\gamma}$, then by Lemma 2.5, for every bounded and measurable function ψ,

$$\lim_{N\to\infty} \int_{\mathscr{X}} \psi \, d\gamma_N = \int_{\mathscr{X}} \psi \, d\bar{\gamma}.$$

Thus along the convergent subsequence, we have

$$\begin{aligned}
-\log \int_{\mathscr{X}} e^{-k}\, d\theta + \varepsilon &\geq \liminf_{N\to\infty} \left[R(\gamma_N \| \theta) + \int_{\mathscr{X}} (k \wedge N)\, d\gamma_N \right] \\
&\geq \liminf_{M\to\infty} \liminf_{N\to\infty} \left[R(\gamma_N \| \theta) + \int_{\mathscr{X}} (k \wedge N \wedge M)\, d\gamma_N \right] \\
&\geq \liminf_{M\to\infty} \left[R(\bar{\gamma} \| \theta) + \int_{\mathscr{X}} (k \wedge M)\, d\bar{\gamma} \right] \\
&\geq \left[R(\bar{\gamma} \| \theta) + \int_{\mathscr{X}} k \, d\bar{\gamma} \right] \\
&\geq \inf_{\gamma \in \mathscr{P}(\mathscr{X})} \left[R(\gamma \| \theta) + \int_{\mathscr{X}} k \, d\gamma \right],
\end{aligned}$$

where the third inequality uses the lower semicontinuity of $\gamma \mapsto R(\gamma \| \theta)$ and Lemma 2.5, and the fourth inequality follows from the monotone convergence theorem. Sending $\varepsilon \to 0$ completes the proof of the variational formula under the assumption that k is bounded from below.

Next consider (b). Since the infimum is restricted to probability measures γ satisfying $R(\gamma \| \theta) < \infty$, the right-hand side of equation (2.3) is well defined. For $N \in \mathbb{N}$, $k \vee (-N)$ is bounded and measurable, and so by part (a) of Proposition 2.2,

$$\begin{aligned}
-\log \int_{\mathscr{X}} e^{-[k\vee(-N)]}\, d\theta &= \inf_{\gamma \in \Delta(\mathscr{X})} \left[R(\gamma \| \theta) + \int_{\mathscr{X}} [k \vee (-N)]\, d\gamma \right] \\
&\geq \inf_{\gamma \in \Delta(\mathscr{X})} \left[R(\gamma \| \theta) + \int_{\mathscr{X}} k \, d\gamma \right].
\end{aligned}$$

The monotone convergence theorem yields

$$\begin{aligned}
-\log \int_{\mathscr{X}} e^{-k}\, d\theta &= \lim_{N\to\infty} \left(-\log \int_{\mathscr{X}} e^{-[k\vee(-N)]}\, d\theta \right) \\
&\geq \inf_{\gamma \in \Delta(\mathscr{X})} \left[R(\gamma \| \theta) + \int_{\mathscr{X}} k \, d\gamma \right].
\end{aligned} \tag{C.6}$$

Let $\varepsilon > 0$ be given. In order to prove that

$$-\log \int_{\mathscr{X}} e^{-k}\, d\theta \leq \inf_{\gamma \in \Delta(\mathscr{X})} \left[R(\gamma \| \theta) + \int_{\mathscr{X}} k \, d\gamma \right],$$

we assume that the right-hand side is less than ∞, for otherwise, there is nothing to prove. We choose a probability measure $\tilde{\gamma} \in \Delta(\mathscr{X})$ such that

$$R(\tilde{\gamma}\,\|\theta) + \int_{\mathscr{X}} k\,d\tilde{\gamma} \le \inf_{\gamma\in\Delta(\mathscr{X})}\left[R(\gamma\,\|\theta) + \int_{\mathscr{X}} k\,d\gamma\right] + \varepsilon < \infty.$$

Then

$$\begin{aligned} -\log\int_{\mathscr{X}} e^{-[k\vee(-N)]}\,d\theta &= \inf_{\gamma\in\Delta(\mathscr{X})}\left[R(\gamma\,\|\theta) + \int_{\mathscr{X}} [k\vee(-N)]\,d\gamma\right] \\ &\le R(\tilde{\gamma}\,\|\theta) + \int_{\mathscr{X}} [k\vee(-N)]\,d\tilde{\gamma}. \end{aligned}$$

Since $R(\tilde{\gamma}\,\|\theta) < \infty$, the monotone convergence theorem yields

$$\begin{aligned} -\log\int_{\mathscr{X}} e^{-k}\,d\theta &= \lim_{N\to\infty}\left(-\log\int_{\mathscr{X}} e^{-[k\vee(-N)]}\,d\theta\right) \\ &\le \lim_{N\to\infty}\left(R(\tilde{\gamma}\,\|\theta) + \int_{\mathscr{X}} [k\vee(-N)]\,d\tilde{\gamma}\right) \\ &= R(\tilde{\gamma}\,\|\theta) + \int_{\mathscr{X}} k\,d\tilde{\gamma} \\ &\le \inf_{\gamma\in\Delta(\mathscr{X})}\left[R(\gamma\,\|\theta) + \int_{\mathscr{X}} k\,d\gamma\right] + \varepsilon. \end{aligned} \tag{C.7}$$

Sending $\varepsilon \to 0$ completes the proof of (2.3) under the assumption that k is bounded from above.

Next we consider (c), where k is not assumed bounded below or above. To simplify notation, we prove the equivalent claim

$$\log\int_{\mathbb{R}^d} e^{k}\,d\theta = \sup_{\gamma\in\Delta(\mathbb{R}^d)}\left[\int_{\mathbb{R}^d} k\,d\gamma - R(\gamma\,\|\theta)\right]. \tag{C.8}$$

By assumption, there exists $\zeta > 0$ such that $\int_{\mathbb{R}^d} e^{\zeta\|x\|}\theta(dx) < \infty$. Recalling the inequality

$$ab \le e^{a} + \ell(b) \text{ for all } a, b \ge 0, \tag{C.9}$$

we have, for every $\gamma \in \Delta(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} \|x\|\,d\gamma \le \frac{1}{\zeta}\int_{\mathbb{R}^d} e^{\zeta\|x\|}\theta(dx) + \frac{1}{\zeta}R(\gamma\,\|\theta) < \infty. \tag{C.10}$$

Thus the right side in (C.8) is well defined.

The first issue is to show that if the left side of (C.8) is ∞, then the right side is also. For $N \in \mathbb{N}$, let $f_N(x) = k(x) \wedge N$. Then f_N is bounded from above, and so the

probability measure

$$\gamma_N(dx) = \frac{1}{Z_N} e^{f_N(x)} \theta(dx), \quad Z_N \doteq \int_{\mathbb{R}^d} e^{f_N(x)} \theta(dx)$$

is well defined, and by Fatou's lemma, $Z_N \to \infty$ as $N \to \infty$. With this choice of γ_N, since $f_N(x) \le k(x)$, the right side of (C.8) is bounded below by

$$\begin{aligned} \left[\int_{\mathbb{R}^d} k \, d\gamma_N - R(\gamma_N \| \theta) \right] &= \left[\int_{\mathbb{R}^d} k \, d\gamma_N - \int_{\mathbb{R}^d} \log \left(\frac{e^{f_N}}{Z_N} \right) d\gamma_N \right] \\ &= \log Z_N + \int_{\mathbb{R}^d} [k - f_N] \, d\gamma_N \\ &\ge \log Z_N. \end{aligned}$$

Letting $N \to \infty$ shows that the right side in (C.8) also equals ∞. If the left-hand side of (C.8) is finite, then $\log Z_N$ converges to that value, and in this case, sending $N \to \infty$ shows that

$$\log \int_{\mathbb{R}^d} e^k \, d\theta \le \sup_{\gamma \in \Delta(\mathbb{R}^d)} \left[\int_{\mathbb{R}^d} k \, d\gamma - R(\gamma \| \theta) \right].$$

We now argue the reverse inequality. It suffices to show that for all $\gamma \in \Delta(\mathbb{R}^d)$,

$$R(\gamma \| \theta) \ge \int_{\mathbb{R}^d} k \, d\gamma - \log \int_{\mathbb{R}^d} e^k \, d\theta. \tag{C.11}$$

For $M \in \mathbb{N}$, let

$$F_M(x) \doteq k(x) 1_{\{|k(x)| \le M\}} + \frac{Mk(x)}{|k(x)|} 1_{\{|k(x)| > M\}}.$$

From part (a) of Lemma 2.4, we have

$$R(\gamma \| \theta) \ge \int_{\mathbb{R}^d} F_M d\gamma - \log \int_{\mathbb{R}^d} e^{F_M} d\theta. \tag{C.12}$$

Next note that by the dominated convergence theorem,

$$\lim_{M \to \infty} \int_{\mathbb{R}^d} F_M d\gamma = \int_{\mathbb{R}^d} k d\gamma$$

and

$$\lim_{M \to \infty} \int_{\{k<0\}} e^{F_M} d\theta = \int_{\{k<0\}} e^k d\theta.$$

Also by the monotone convergence theorem,

$$\lim_{M\to\infty}\int_{\{k\geq 0\}} e^{F_M}d\theta = \int_{\{k\geq 0\}} e^{k}d\theta.$$

Using the above three convergence properties and sending $M \to \infty$ in (C.12), we have (C.11), completing the proof of the reverse inequality. □

Appendix D
Martingales and Stochastic Integration

We begin with some basic definitions. Fix a finite-time horizon $T \in (0, \infty)$. Let $(\Omega, \mathscr{F}, P)$ be a probability space that is equipped with a **filtration** $\{\mathscr{F}_t\}_{0 \le t \le T}$, which means that $\mathscr{F}_s \subset \mathscr{F}_t \subset \mathscr{F}$ for all $0 \le s \le t \le T$. We will assume throughout that $\mathscr{F}$ is P-complete and the filtration satisfies the **usual conditions**, namely that the filtration is right continuous and for every $t \in [0, T]$, $\mathscr{F}_t$ contains all P-null sets in $\mathscr{F}$.

We say that a stochastic process $X = \{X(t)\}_{0 \le t \le T}$ on $(\Omega, \mathscr{F}, P)$ with values in some Polish space $\mathscr{E}$ is **RCLL** (resp. **LCRL**) if for every $\omega \in \Omega$, the map $t \mapsto X(t, \omega)$ from $[0, T]$ to $\mathscr{E}$ is right continuous on $[0, T)$ (resp. left continuous on $(0, T]$) and has left limits on $(0, T]$ (resp. has right limits on $[0, T)$). An $\mathscr{E}$-valued stochastic process X is said to be $\mathscr{F}_t$**-adapted** if for every $t \in [0, T]$, $X(t)$ is $\mathscr{F}_t$-measurable. It is said to be $\mathscr{F}_t$**-progressively measurable** if for every $t \in [0, T]$, the mapping $(s, \omega) \mapsto X(s, \omega)$ from $([0, t] \times \Omega, \mathscr{B}([0, t]) \otimes \mathscr{F}))$ to $(\mathscr{E}, \mathscr{B}(\mathscr{E}))$ is measurable. Denote by $\mathscr{P}\mathscr{F}$ the σ-field on $[0, T] \times \Omega$ generated by the collection of all real $\mathscr{F}_t$-adapted LCRL processes (note that this is the same σ-field as that generated by the elementary functions or simple functions, as used, for example, in Definition 8.2). This σ-field is called the $\mathscr{F}_t$**-predictable** σ**-field**. For a Polish space $\mathscr{E}$, a $\mathscr{P}\mathscr{F}/\mathscr{B}(\mathscr{E})$-measurable map $X : [0, T] \times \Omega \to \mathscr{E}$ is referred to as an $\mathscr{E}$-valued $\mathscr{F}_t$**-predictable process**.

A $[0, T]$-valued random variable τ on $(\Omega, \mathscr{F})$ is said to be an $\mathscr{F}_t$**-stopping time** if $\{\tau \le t\} \in \mathscr{F}_t$ for every $t \in [0, T]$.

D.1 Martingales

Let $\{X(t)\}_{0 \le t \le T}$ be a real-valued $\mathscr{F}_t$-adapted process such that $E|X(t)| < \infty$ for every $t \in [0, T]$. Such a process is called an $\mathscr{F}_t$**-submartingale** (resp. an $\mathscr{F}_t$**-supermartingale**) if for all $0 \le s \le t \le T$, $E[X(t) | \mathscr{F}_s] \ge X(s)$ [resp. $E[X(t) | \mathscr{F}_s] \le X(s)$]. A process that is both an $\mathscr{F}_t$-submartingale and an $\mathscr{F}_t$-supermartingale is an $\mathscr{F}_t$**-martingale**. A martingale admits an RCLL modification, which is a

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0

martingale with respect to the same filtration, and thus without loss of generality, we use RCLL modifications of martingales. A real-valued stochastic process X is called an $\mathscr{F}_t$**-local martingale** if there is a sequence of $\mathscr{F}_t$-stopping times τ^n increasing to T such that $\{X^{(n)}(t) \doteq X(t \wedge \tau_n)\}_{0 \le t \le T}$ is an $\mathscr{F}_t$-martingale for every n. A **locally square-integrable martingale** is defined in the analogous way. For two square-integrable $\mathscr{F}_t$-martingales X and Y with $X(0) = Y(0) = 0$, their **quadratic covariation**, denoted by $[X, Y]$, is the unique adapted RCLL process A with paths of bounded variation such that $A(0) = 0$, $XY - A$ is a martingale, and $\Delta A = \Delta X \Delta Y$, where for a real RCLL stochastic process Z, $\Delta Z(t) = Z(t) - Z(t-)$, $t \in [0, T]$. For such a process A, there is a unique decomposition $A = M + \tilde{A}$, where $M(0) = \tilde{A}(0) = 0$, M is a martingale, and $\tilde{A}$ is an $\mathscr{F}_t$-predictable process with paths of bounded variation. The process $\tilde{A}$ is called the **predictable quadratic covariation** of X and Y and is denoted by $\langle X, Y \rangle$. These definitions can be extended to local martingales (see [210]). When $X = Y$, these processes are sometimes denoted by $[X]$ and $\langle X \rangle$, and referred to as the **quadratic variation** (resp. **predictable quadratic variation**) of X. When X and Y are continuous, $[X, Y]$ is a continuous adapted process and hence predictable, in which case $[X, Y]$ coincides with $\langle X, Y \rangle$.

The following are some of the martingale inequalities used in this book. Discrete-time analogues of the first three are well known (see, for example, [173, Theorem 11.2] for the first two and [199] for the third). For the continuous time setting, see [172, Theorem 1.3.8] for the first two, [210, Theorem IV.4.48] for the third, and [180, Lemma 2.4] for the last.

Doob's submartingale inequality. For every nonnegative submartingale X, $c \in (0, \infty)$, and $t \in [0, T]$,

$$P\left[\sup_{0 \le s \le t} X(s) \ge c\right] \le \frac{1}{c} E[X(t)]. \tag{D.1}$$

Doob's maximal inequality. For every martingale M and $t \in [0, T]$,

$$E\left[\sup_{0 \le s \le t} |M(s)|^2\right] \le 4E[|M(t)|^2]. \tag{D.2}$$

Burkholder–Davis–Gundy inequality. For every $p \ge 1$, there exist $C_p \in (0, \infty)$ such that for every locally square-integrable martingale M with $M(0) = 0$ and $t \in [0, T]$,

$$E\left[\sup_{0 \le s \le t} |M(s)|^p\right] \le C_p E[[M, M](t)]^{p/2}. \tag{D.3}$$

Lenglart–Lepingle–Pratelli inequality. For $0 < p \le 2$, there exist $C_p \in (0, \infty)$ such that for every locally square-integrable martingale M with $M(0) = 0$ and $t \in [0, T]$,

$$E\left[\sup_{0 \le s \le t} |M(s)|^p\right] \le C_p E[\langle M, M \rangle(t)]^{p/2}. \tag{D.4}$$

The notion of quadratic variation can be extended to vector-valued martingales. Let $M = (M_1, \ldots, M_k)^T$ be an $\mathbb{R}^k$-valued stochastic process such that $\{M_i(t)\}_{0\le t\le T}$ is an $\{\mathscr{F}_t\}$-martingale for each $i = 1, \ldots, k$. We refer to M as a k-dimensional $\{\mathscr{F}_t\}$-martingale. Let M and N be k-dimensional and r-dimensional $\{\mathscr{F}_t\}$-martingales, respectively. Then $\langle\langle M, N\rangle\rangle$ is the $(k \times r)$-dimensional stochastic process given by

$$(\langle\langle M, N\rangle\rangle(t))_{ij} \doteq \langle M_i, N_j\rangle_t, \ 1 \le i \le k, 1 \le j \le r, \ t \in [0, T].$$

The martingale inequalities above can be extended to k-dimensional martingales. In particular, the Burkholder–Davis–Gundy inequality for $p = 2$ and a k-dimensional $\{\mathscr{F}_t\}$-martingale M says that

$$E\left[\sup_{0\le s\le t} \|M(s)\|^2\right] \le C_2 E[\mathrm{tr}(\langle\langle M, M\rangle\rangle(t))]. \tag{D.5}$$

D.2 Stochastic Integration

In this section we summarize the various types of stochastic integrals used in this book. We begin with the setting of d-dimensional Brownian motion.

D.2.1 Brownian Motion in $\mathbb{R}^d$

Let W be a d-dimensional $\mathscr{F}_t$-Brownian motion as introduced in Sect. 3.2. Let $\bar{\mathscr{A}}$ as in Definition 3.12 denote the collection of all $\mathbb{R}^d$-valued $\mathscr{F}_t$-progressively measurable processes $\{v(t)\}_{0\le t\le T}$ that satisfy $E[\int_0^T \|v(t)\|^2 dt] < \infty$. Then the stochastic integral $M_v(t) = \int_0^t v(s)dW(s)$ (see [172, Chap. 3]) is a square-integrable continuous $\mathscr{F}_t$-martingale, and for $v_1, v_2 \in \bar{\mathscr{A}}$,

$$\langle M_{v_1}, M_{v_2}\rangle(t) = [M_{v_1}, M_{v_2}](t) = \int_0^t \langle v_1(s)v_2(s)\rangle ds.$$

A similar result holds when $d = \infty$. More precisely, let $\{\beta_i\}_{i=1}^\infty$ be a sequence of independent one-dimensional $\{\mathscr{F}_t\}$-Brownian motions. Let $f_i \in \bar{\mathscr{A}}$ (with $d = 1$) for each $i \in \mathbb{N}$, and suppose that $\sum_{i=1}^\infty E[\int_0^T |f_i(t)|^2 dt] < \infty$. Then

$$M(t) \doteq \sum_{i=1}^\infty \int_0^t f_i(s)d\beta_i(s), \ t \in [0, T]$$

is a continuous $\{\mathscr{F}_t\}$-martingale and $\langle M\rangle_t = \sum_{i=1}^\infty \int_0^t |f_i(s)|^2 ds$ for $t \in [0, T]$.

In considering discontinuous martingales, we will consider integrands v that instead of being progressively measurable, lie in the smaller class of predictable processes. In the setting of Brownian motions, there is not much difference between the two classes, since for every $v \in \bar{\mathscr{A}}$, there is a predictable $\tilde{v}$ such that $v = \tilde{v}$ a.s. $dt \otimes P$, and the stochastic integrals M_v and $M_{\tilde{v}}$ agree a.s.

D.2.2 Point Processes

A reference for the topic of this section is [159]. Let $(\Omega, \mathscr{F}, P)$ and $\{\mathscr{F}_t\}_{0\le t\le T}$ be as at the beginning of this appendix. Let $\mathscr{X}$, $\mathscr{Y}$, $\mathscr{X}_T$, $\mathscr{Y}_T$, ν, $\bar{\nu}$, ν_T, $\bar{\nu}_T$ be as in Sect. 8.2.1. Also let $\bar{N}$ be a Poisson random measure (PRM) with respect to $\{\mathscr{F}_t\}$ on $\mathscr{Y}_T$ with intensity measure $\bar{\nu}_T$ (see Sect. 8.2.1). Let $\bar{\mathscr{A}}$ be as introduced below (8.19), and for $\varphi \in \bar{\mathscr{A}}$, N^φ is defined as in Sect. 8.2.1 through (8.16). We will also consider the compensated point processes $\bar{N}_c(ds \times dr) = \bar{N}(ds \times dr) - \bar{\nu}_T(ds \times dr)$ and $N_c^\varphi(ds \times dx) = N^\varphi(ds \times dx) - \varphi(s,x)\nu_T(ds \times dx)$. Let $\mathscr{P}\mathscr{F}$ be the predictable σ-field associated with $\{\mathscr{F}_t\}$. For $t \in [0,T]$ and $\psi : [0,T] \times \Omega \times \mathscr{Y} \to \mathbb{R}$, that is, $(\mathscr{P}\mathscr{F} \otimes \mathscr{B}(\mathscr{Y}))/\mathscr{B}(\mathbb{R})$ measurable and satisfying

$$E \int_{\mathscr{Y}_T} |\psi(s,y)| \bar{\nu}_T(ds \times dy) < \infty,$$

the stochastic integral

$$M_\psi(t) \doteq \int_{[0,t]\times\mathscr{Y}} \psi(s,y) \bar{N}_c(ds \times dy)$$

is well defined, and the stochastic process M_ψ is a martingale. Thus

$$E \int_{\mathscr{Y}_T} \psi(s,y) \bar{N}(ds \times dy) = E \int_{\mathscr{Y}_T} \psi(s,y) \bar{\nu}_T(ds \times dy). \tag{D.6}$$

If in addition

$$E \int_{\mathscr{Y}_T} \psi(s,y)^2 \bar{\nu}_T(ds \times dy) < \infty,$$

then M_ψ is a square-integrable martingale with quadratic variation

$$[M_\psi](t) = \int_{[0,t]\times\mathscr{Y}} \psi(s,y)^2 \bar{N}(ds \times dy).$$

For ψ_i, $i = 1, 2$, as above,

$$\langle M_{\psi_1}, M_{\psi_2} \rangle(t) = \int_{[0,t]\times\mathscr{Y}} \psi_1(s,y)\psi_2(s,y) \bar{\nu}_T(ds \times dy).$$

Similarly, if $\psi : [0, T] \times \Omega \times \mathscr{X} \to \mathbb{R}$ is $(\mathscr{P}\mathscr{F} \otimes \mathscr{B}(\mathscr{X}))/\mathscr{B}(\mathbb{R})$-measurable and

$$E \int_{\mathscr{X}_T} (|\psi(s,x)| \vee |\psi(s,x)|^2)\varphi(s,x)\nu_T(ds \times dx) < \infty,$$

then the stochastic integral

$$M_\psi(t) \doteq \int_{[0,t]\times\mathscr{X}} \psi(s,x) N_c^\varphi(ds \times dx)$$

is well defined, and the stochastic process M_ψ is a square-integrable martingale with quadratic variation

$$[M_\psi](t) = \int_{[0,t]\times\mathscr{X}} \psi(s,x)^2 N^\varphi(ds \times dx).$$

For two such integrands ψ_1, ψ_2, we have

$$\langle M_{\psi_1}, M_{\psi_2}\rangle(t) = \int_{[0,t]\times\mathscr{X}} \psi_1(s,x)\psi_2(s,x)\varphi(s,x)\nu_T(ds \times dx).$$

D.2.3 Hilbert Space Valued Brownian Motion

Let $(\mathscr{H}, \langle\cdot,\cdot\rangle)$ be a real separable Hilbert space. Let Λ be a strictly positive symmetric trace class operator on $\mathscr{H}$. Let $\{W(t)\}_{0\le t\le T}$ be a Λ-Wiener process with respect to $\{\mathscr{F}_t\}_{0\le t\le T}$ as introduced in Definition 8.1. Also let $(\mathscr{H}_0, \langle\cdot,\cdot\rangle_0)$ be the Hilbert space introduced in Sect. 8.1, i.e., $\mathscr{H}_0 \doteq \Lambda\mathscr{H}$ and $\langle h, k\rangle_0 \doteq \langle \Lambda^{-1/2}h, \Lambda^{-1/2}k\rangle$. Let $\bar{\mathscr{A}}$ be the class of $\mathscr{H}_0$-valued $\mathscr{F}_t$-predictable processes v that satisfy

$$P\left\{\int_0^T \|v(s)\|_0^2 ds < \infty\right\} = 1,$$

as introduced below (8.2). Then for every $\psi \in \bar{\mathscr{A}}$ and $t \in [0, T]$, the stochastic integral $M_t \doteq \int_0^t \langle\psi(s), dW(s)\rangle_0$ is defined as in [69, Sect. 4.2]. Furthermore, M is a continuous $\{\mathscr{F}_t\}$-local martingale, which is a martingale if $E\int_0^T \|\psi(s)\|_0^2 ds < \infty$, in which case $\langle M\rangle_t = \int_0^t \|\psi(s)\|_0^2 ds$ for $t \in [0, T]$.

D.2.4 Brownian Sheet

Let $\mathscr{O}$ be a bounded open subset of $\mathbb{R}^d$. Let $\{B(t,x),\ (t,x) \in [0,T]\times\mathscr{O}\}$ be a Brownian sheet on $(\Omega, \mathscr{F}, P)$ with respect to the filtration $\{\mathscr{F}_t\}$ as introduced in

Definition 11.5. Let $\bar{\mathscr{A}}$ be, as introduced below Definition 11.7, the class of all $\{\mathscr{F}_t\}$-predictable processes f such that $\int_{[0,T]\times\mathscr{O}} f^2(s,x)dsdx < \infty$ a.s. Then the stochastic integral $M_t(f) \doteq \int_{[0,t]\times\mathscr{O}} f(s,u)B(ds \times du)$, $t \in [0,T]$, is defined as in Chap. 2 of [243]. Furthermore, $\{M_t(f)\}$ is a continuous $\{\mathscr{F}_t\}$-local martingale, which is a martingale if $E\int_{[0,T]\times\mathscr{O}} f^2(s,x)dsdx < \infty$, in which case the quadratic variation is given by $\langle M(f)\rangle_t = \int_{[0,t]\times\mathscr{O}} f^2(s,x)dsdx$.

D.3 Girsanov's Theorem

In this section we summarize some variations of Girsanov's theorem, in addition to those already presented in Chap. 8, that are appealed to in this book. We begin with the classical setting of a finite dimensional Brownian motion. A proof can be found in [172, Sect. 3.5].

Theorem D.1 *Let W be a d-dimensional $\mathscr{F}_t$-Brownian motion and $\{v(t)\}_{0\le t\le T}$ a $\mathbb{R}^d$-valued $\mathscr{F}_t$-progressively measurable process that satisfies $E[\int_0^T \|v(t)\|^2 dt] < \infty$. Suppose*

$$E\left[\exp\left\{\int_0^T v(s)\,dW(s) - \frac{1}{2}\int_0^T \|v(s)\|^2 ds\right\}\right] = 1.$$

Then the process

$$\tilde{W}(t) \doteq W(t) - \int_0^t v(s)ds,$$

$t \in [0,T]$, is an $\{\mathscr{F}_t\}$-Brownian motion on $(\Omega, \mathscr{F}, Q)$, where Q is the probability measure defined by

$$\frac{dQ}{dP} = \exp\left\{\int_0^T v(s)\,dW(s) - \frac{1}{2}\int_0^T \|v(s)\|^2 ds\right\}.$$

A similar result holds for $d = \infty$ (see [69, Theorem 10.14]). For that case, W is replaced by a sequence $\{\beta_i\}_{i=1}^\infty$ of independent one-dimensional $\{\mathscr{F}_t\}$-Brownian motions, v with a sequence $f_i \in \bar{\mathscr{A}}$ (with $\bar{\mathscr{A}}$ as in Definition 3.12 and $d = 1$) such that $E[\int_0^T \sum_{i=1}^\infty |f_i(t)|^2 dt] < \infty$, and the integrals $\int_0^T v(s)\,dW(s)$ and $\int_0^T \|v(s)\|^2 ds$ replaced by $\sum_{i=1}^\infty \int_0^T f_i(t)d\beta_i(t)$ and $\sum_{i=1}^\infty \int_0^T |f_i(t)|^2 dt$, respectively. We omit the precise statement.

Girsanov's theorem for a Brownian sheet takes the following form (see [206, Proposition 1.6].

Theorem D.2 *Let $\mathscr{O}$ be a bounded open subset of $\mathbb{R}^d$ and suppose that $\{B(t,x), (t,x) \in [0,T]\times\mathscr{O}\}$ is a Brownian sheet on $(\Omega, \mathscr{F}, P)$ with respect to the filtration $\{\mathscr{F}_t\}$. Let f be $\{\mathscr{F}_t\}$-predictable in the sense of Definition 11.7 and satisfy $E\int_{[0,T]\times\mathscr{O}} f^2(s,x)ds\,dx < \infty$. Suppose that*

$$E\left[\exp\left\{\int_{[0,T]\times\mathscr{O}} f(s,u)B(ds\times du) - \frac{1}{2}\int_{[0,T]\times\mathscr{O}} f^2(s,x)dsdx\right\}\right] = 1.$$

Then the random field $\left\{\tilde{B}(t,x),\ (t,x)\in[0,T]\times\mathscr{O}\right\}$ *defined by*

$$\tilde{B}(t,x) \doteq B(t,x) - \int_0^t \int_{(-\infty,x]\cap\mathscr{O}} f(s,y)dyds$$

is a Brownian sheet with respect to $\{\mathscr{F}_t\}$ *on* $(\Omega,\mathscr{F},Q)$*, where* Q *is the probability measure defined by*

$$\frac{dQ}{dP} = \exp\left\{\int_{[0,T]\times\mathscr{O}} f(s,u)B(ds\times du) - \frac{1}{2}\int_{[0,T]\times\mathscr{O}} f^2(s,x)dsdx\right\}.$$

Finally, we present a version of Girsanov's theorem for systems with both Brownian and Poisson noise. We will not aim for maximum generality but rather state the result in the form in which it is used in the book. The result follows from the independence of the Brownian motion and PRM, and their corresponding versions of Girsanov's theorem (cf. [161, Theorem III.3.24]). We consider only a finite dimensional Brownian motion here; extensions to settings with an infinite dimensional Brownian motion can be written similarly.

Theorem D.3 *With notation and processes* W *and* N *that satisfy the conditions of Sect. 8.3, let* $u=(\psi,\varphi)\in\bar{\mathscr{A}}_b$*. Let*

$$\mathscr{E}_1^\varepsilon(t) \doteq \exp\left[\int_{\mathscr{X}_t\times[0,\infty)} 1_{[0,\varepsilon^{-1}\varphi(s,y)]}(r)\log(\tilde{\varphi}(s,y))\bar{N}(ds\times dy\times dr)\right.$$
$$\left.+\int_{\mathscr{X}_t\times[0,\infty)} 1_{[0,\varepsilon^{-1}\varphi(s,y)]}(r)(-\tilde{\varphi}(s,y)+1)\bar{\nu}_T(ds\times dy\times dr)\right],$$

$$\mathscr{E}_2^\varepsilon(t) \doteq \exp\left[-\frac{1}{\sqrt{\varepsilon}}\int_0^t \psi(s)dW(s) - \frac{1}{2\varepsilon}\int_0^t \|\psi(s)\|^2 ds\right],$$

and $\mathscr{E}^\varepsilon(t) \doteq \mathscr{E}_1^\varepsilon(t)\mathscr{E}_2^\varepsilon(t)$*. Then* $\{\mathscr{E}^\varepsilon(t)\}_{0\le t\le T}$ *is an* $\mathscr{F}_t$*-martingale, and consequently,*

$$\bar{Q}^\varepsilon(A) = \int_A \mathscr{E}^\varepsilon(T)dP,\ A\in\mathscr{F}$$

defines a probability measure on $(\Omega,\mathscr{F})$*. Furthermore,*

$$\left(W + \frac{1}{\sqrt{\varepsilon}}\int_0^\cdot \psi(s)ds,\ \varepsilon N^{\varphi/\varepsilon}\right)$$

under $\bar{Q}^\varepsilon$ *has the same probability law as* $(W,\varepsilon N^{1/\varepsilon})$ *under* P*.*

D.4 Criteria for Tightness

The next result, due to Aldous [3], considers tightness of a sequence of random variables $\{X^n\}_{n\in\mathbb{N}}$ with values in $\mathscr{D}([0,T]:\mathscr{E})$. For a proof, see [179, Theorem 2.7]. To simplify notation, it is assumed that all processes are defined on a common probability space $(\Omega, \mathscr{F}, P)$. Recall that τ is an $\mathscr{F}_t$-stopping time if $\{\tau \le t\} \in \mathscr{F}_t$ for all $t \in [0,T]$.

Theorem D.4 *Let $\{X^n\}_{n\in\mathbb{N}}$ be a sequence of processes with paths in $\mathscr{D}([0,T]:\mathscr{E})$ and let $\mathscr{F}_t^n$ be the σ-algebra generated by $\{X^n(s), 0 \le s \le t\}$. Suppose that $\{X^n(t)\}_{n\in\mathbb{N}}$ is tight for each rational $t \in [0,T]$, and that for every sequence of $\mathscr{F}_t^n$-stopping times $\{\tau_n\}$ such that $\tau_n \le T$ and every sequence of nonnegative numbers $\{\delta_n\}$ converging to zero as $n \to \infty$,*

$$d(X^n(\tau_n + \delta_n), X^n(\tau_n)) \to 0$$

in probability as $n \to \infty$. Then $\{X^n\}_{n\in\mathbb{N}}$ is tight.

The theorem is also true if $\mathscr{D}([0,T]:\mathscr{E})$ is replaced by $\mathscr{C}([0,T]:\mathscr{E})$.

D.5 Diffeomorphic Properties of Solutions of Itô SDEs

The following is [178, Theorem 4.6.5].

Theorem D.5 *Suppose that the local characteristic (a,b) of a $\mathscr{C}^{k,\nu}$-Brownian motion $\{\Phi(t)\}_{t\ge 0}$ satisfies Condition 12.2 with some $\delta > \nu$. Then the solution of Itô's stochastic differential equation based on the Brownian motion Φ has a modification $\{\phi_{s,t}\}_{0\le s\le t\le T}$ that is a forward stochastic flow of $\mathscr{C}^k$-diffeomorphisms.*

Appendix E
Analysis and Measure Theory

E.1 Measure Theory

The following result is well known. A proof can be found in [167].

Lemma E.1 *Let $\mathscr{X}_1$, $\mathscr{X}_2$ be Polish spaces and let X be an $\mathscr{X}_1$-valued Borel measurable map defined on some measurable space $(\Omega, \mathscr{F})$. Let $\mathscr{G} = \sigma\{X\}$. Suppose Y is an $\mathscr{X}_2$-valued Borel measurable map given on $(\Omega, \mathscr{F})$ that is $\mathscr{G}$-measurable. Then there is a Borel measurable map $g : \mathscr{X}_1 \to \mathscr{X}_2$ such that $Y = g(X)$.*

E.2 Gronwall's Inequality

Lemma E.2 (GRONWALL'S LEMMA) *Let f, g be measurable maps from $[0, \infty)$ to $[0, \infty)$. Suppose that for some $a \in [0, \infty)$,*

$$f(t) \le a + \int_0^t f(s)g(s)ds \quad \text{for all } t \in [0, \infty). \tag{E.1}$$

Also suppose that $\sup_{0\le s\le t} f(s) < \infty$ for each fixed $t \in [0, \infty)$. Then

$$f(t) \le a e^{\int_0^t g(s)ds} \quad \text{for all } t \in [0, \infty).$$

Proof Fix $t \in [0, \infty)$. We assume without loss of generality that $\int_0^t g(s)ds < \infty$. Iterating (E.1) n times, we get

$$f(t) \le a + a \sum_{k=1}^{n} \int_0^t g(s_1) \int_0^{s_1} g(s_2) \cdots \int_0^{s_{k-1}} g(s_k)\, ds_k \cdots ds_1 + R_n(t), \tag{E.2}$$

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0

where

$$R_n(t) = \int_0^t g(s_1) \int_0^{s_1} g(s_2) \cdots \int_0^{s_{n-1}} g(s_n) \int_0^{s_n} f(s_{n+1}) g(s_{n+1})\, ds_{n+1} ds_n \cdots ds_1.$$

Note that

$$R_n(t) \le \left(\int_0^t g(s) f(s) ds \right) \frac{\left(\int_0^t g(s) ds \right)^n}{n!} \le \left(\sup_{0 \le s \le t} f(s) \right) \left(\int_0^t g(s) ds \right) \frac{\left(\int_0^t g(s) ds \right)^n}{n!}.$$

Using the fact that $\int_0^t g(s) ds < \infty$, we see that $R_n(t) \to 0$ as $n \to \infty$. Sending $n \to \infty$ in (E.2), we have

$$\begin{aligned} f(t) &\le a + a \sum_{k=1}^{\infty} \int_0^t g(s_1) \int_0^{s_1} g(s_2) \cdots \int_0^{s_{k-1}} g(s_k) ds_k \cdots ds_1 \\ &= a \sum_{k=0}^{\infty} \frac{1}{k!} \left(\int_0^t g(s) ds \right)^k \\ &= a e^{\int_0^t g(s) ds}. \end{aligned}$$

This completes the proof of the lemma. □

E.3 Measurable Selection and Approximation of Measurable Functions

Let $(\mathscr{X}_2, \rho_2)$ be a complete and separable metric space and let $(\mathscr{X}_1, \rho_1)$ be a metric space. Suppose that for each $x \in \mathscr{X}_1$, $\Gamma_x \subset \mathscr{X}_2$. A **measurable selection** of $\{\Gamma_x\}_{x \in \mathscr{X}_1}$ is a $\mathscr{B}(\mathscr{X}_1) \backslash \mathscr{B}(\mathscr{X}_2)$-measurable function $f : \mathscr{X}_1 \to \mathscr{X}_2$ such that $f(x) \in \Gamma_x$ for every $x \in \mathscr{X}_1$. The following result is proved in Corollary 10.3 in Appendix 10 of [126].

Corollary E.3 *Suppose that if $y_n \in \Gamma_{x_n}$ for $n \in \mathbb{N}$ and $x_n \to x$ as $n \to \infty$, then $\{y_n\}_{n \in \mathbb{N}}$ has a limit point in Γ_x. Then a measurable selection of $\{\Gamma_x\}_{x \in \mathscr{X}_1}$ exists.*

We next state an approximation result (see [90, Theorem V.16a]).

Theorem E.4 *Let $\mathscr{X}$ be a Polish space and suppose $\lambda \in \mathscr{P}(\mathscr{X})$. Let $f : \mathscr{X} \to \mathbb{R}$ be a Borel measurable function. Then there is a sequence of continuous functions $\{f_j\}_{j \in \mathbb{N}}$, $f_j : \mathscr{X} \to \mathbb{R}$, such that*

$$f_j \to f \qquad \lambda\text{-a.e.}$$

as $j \to \infty$. If the function f is bounded in absolute value by B, then all the approximating functions can be taken to be bounded in absolute value by B as well.

E.4 Hilbert Spaces

The definitions in this section are taken from [66, 226].

A real vector space H is called an **inner product space** if for each pair $x, y \in H$ there is a real number $\langle x, y\rangle$ such that the following properties hold for every $x, y, z \in H$ and $\alpha \in \mathbb{R}$: (a) $\langle x, y\rangle = \langle y, x\rangle$, (b) $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$, (c) $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$, (d) $\langle x, x\rangle \geq 0$, (e) $\langle x, x\rangle = 0$ if and only if $x = 0$. Such a space can be normed by defining $\|x\|^2 \doteq \langle x, x\rangle$. If the resulting metric space is complete, we call it a **Hilbert space**.

A subset H_0 of a Hilbert space H is a said to be an **orthonormal set** if (a) for every $h \in H_0$, $\|h\| = 1$; (b) if $h_1, h_2 \in H_0$ are such that $h_1 \neq h_2$, then $\langle h_1, h_2\rangle = 0$. A maximal orthonormal set is said to be a **complete orthonormal system** (CONS). Every separable Hilbert space has a countable CONS.

For the rest of this section, H will be a separable Hilbert space with inner product $\langle \cdot, \cdot\rangle$ and norm $\|\cdot\|$. A linear mapping $A : H \to H$ is called a **bounded linear operator** on H if

$$\|A\| \doteq \sup_{x \in H : \|x\| \leq 1} \|Ax\| < \infty.$$

In this case, the mapping is continuous, and $\|A\|$ is called the norm of A.

For every bounded linear operator A on H, A^* is the unique bounded linear operator on H, referred to as the **adjoint** of A, with the property that

$$\langle Ax, y\rangle = \langle x, A^* y\rangle \text{ for all } x, y \in H.$$

A bounded linear operator A on H is called **self-adjoint** or **symmetric** if $A = A^*$. It is called **positive** if $\langle Ax, x\rangle \geq 0$ for all $x \in H$, and it is called **strictly positive** if $\langle Ax, x\rangle > 0$ for all nonzero $x \in H$. A positive and self-adjoint operator has a unique positive **square root** S, which is a positive operator satisfying $S^2 = A$.

Let A be a bounded linear operator on H. Let $\{e_i\}_{i \in \mathbb{N}}$ be a CONS in H. Define $\|A\|_2 \doteq [\sum_{i \in \mathbb{N}} \|Ae_i\|^2]^{1/2}$. It can be checked that $\|A\|_2$ thus defined does not depend on the choice of the CONS. We say that A is a **Hilbert–Schmidt operator** if $\|A\|_2 < \infty$, and we refer to $\|A\|_2$ as the **Hilbert–Schmidt norm** of A.

A bounded linear operator A on H is called a **trace class operator** if $A = BC$, where B, C are Hilbert–Schmidt operators. For such an operator, $\sum_{i \in \mathbb{N}} |\langle Ae_i, e_i\rangle| < \infty$ for every CONS $\{e_i\}$, and the sum $\sum_{i \in \mathbb{N}} \langle Ae_i, e_i\rangle$ is independent of the choice of the CONS. This quantity is referred to as the **trace** of the operator A.

Conventions and Standard Notation

Conventions. The following conventions are used throughout the book.

1. The infimum of the empty set is ∞.
2. $0 \log(0/x) = 0$ and $y \log(y/0) = \infty$ for $x \in [0, \infty)$ and $y \in (0, \infty)$.
3. Sigma fields on topological spaces will always be taken to be Borel σ-fields. A set in a Borel σ-field will be referred to as a Borel set. Mappings on a topological space are Borel measurable.
4. Two types of constants are used. Meaningful constants, such as Lipschitz constants, are denoted by uppercase letters, and constants that are used only in the course of a proof are set lowercase; they take values in $(0, \infty)$.

Standard notation, terminology, and abbreviations. The following standard notation is used throughout the book. A list of more specialized notation is given in the list of Specialized Symbols that follows this section.

General

$\mathscr{B}(S)$	the Borel σ-algebra on a Polish space S.
$(\mathscr{H}, \langle \cdot, \cdot \rangle)$	a real separable Hilbert space.
$\mathscr{P}(S)$	the probability measures on the measurable space $(S, \mathscr{F})$.
1_A	the indicator function of the set A.
δ_x	the probability measure with mass 1 at the point x.
$\gamma \ll \theta$	the measure γ is absolutely continuous with respect to θ.
$\frac{d\gamma}{d\theta}$	the Radon–Nikodym derivative of γ with respect to θ when the measure γ is absolutely continuous with respect to θ.
$\theta_n \Rightarrow \theta$	for $\{\theta_n\}_{n \in N} \cup \{\theta\} \subset \mathscr{P}(S)$, with S a metric space, $\int_S f d\theta_n \to \int_S f d\theta$ for all $f \in \mathscr{C}_b(S)$ and called weak convergence; for random variables $\{X_n\}_{n \in \mathbb{N}}$, X, $X_n \Rightarrow X$ means that the induced measures converge weakly, also called convergence in distribution.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0

$d(x, F)$	$\inf\{d(x, y) : y \in F\}$, the distance from the point x to the set F in a metric space with distance $d(\cdot, \cdot)$
$x \vee y, x \wedge y$	maximum (resp. minimum) of two real numbers x, y.
x^+, x^-	the positive part (resp. the negative part) of a real number x, equivalently $x \vee 0$ (resp. $(-x) \vee 0$).
$\lfloor x \rfloor$	integer part of x.
$\binom{a}{i}$	$\frac{\prod_{j=0}^{i-1}(a-j)}{i!}$, for $a \in \mathbb{R}$, $a \neq 0$ and $i \in \mathbb{N}$.
$B(x, \delta)$	$\{y : d(y, x) < \delta\}$, the open ball of radius δ centered at x in a metric space with distance $d(\cdot, \cdot)$
$\bar{B}, B^\circ, \partial B$	closure, interior, and boundary of a set B, respectively.
$f_n \uparrow f$	for functions $f_n, f : S \to \mathbb{R}$, $f_n(x)$ increases monotonically $f(x)$ for all $x \in S$.
$f \circ g$	composition of two functions f and g.
$\theta \times \sigma$	for $\theta \in \mathscr{P}(\mathscr{X})$ and $\sigma \in \mathscr{P}(\mathscr{Y})$, $(\mathscr{X}, \mathscr{F})$ and $(\mathscr{Y}, \mathscr{G})$ measurable spaces, the unique probability measure on the product space $\mathscr{X} \times \mathscr{Y}$ that satisfies $[\theta \times \sigma](A \times B) = \theta(A)\sigma(B)$ for all $A \in \mathscr{F}$, $B \in \mathscr{G}$.
$\sigma(dy\|x)$	with $x \in \mathscr{V}$, $y \in \mathscr{Y}$, $\mathscr{V}$ a measurable space and $\mathscr{Y}$ a Polish space, a stochastic kernel on $\mathscr{Y}$ given $\mathscr{V}$: $\sigma(\cdot\|x) \in \mathscr{P}(\mathscr{Y})$ for all $x \in \mathscr{V}$ and $x \mapsto \sigma(A\|x)$ is measurable for every $A \in \mathscr{B}(\mathscr{Y})$.
$\theta \otimes \sigma$	for $\theta \in \mathscr{P}(\mathscr{X})$, $(\mathscr{X}, \mathscr{F})$ a measurable space, and $\sigma(dy\|x)$ a stochastic kernel on $\mathscr{Y}$ given $\mathscr{X}$, $\mathscr{Y}$ a Polish space, the unique probability measure on the product space $\mathscr{X} \times \mathscr{Y}$ obtained from a probability measure θ on $\mathscr{X}$ and a stochastic kernel $\sigma(dy\|x)$ on $\mathscr{Y}$ given $\mathscr{X}$ that satisfies $[\theta \times \sigma](A \times B) = \int_A \theta(dx)\sigma(B\|x)$ for all $A \in \mathscr{F}$, $B \in \mathscr{B}(\mathscr{Y})$.
$[\alpha]_i, [\alpha]_{j\|i}$	for $\alpha \in \mathscr{P}(\mathscr{X}_1 \times \cdots \times \mathscr{X}_k)$ with each $\mathscr{X}_i$ a Polish space and the product σ-algebra used, $[\alpha]_i$ is the marginal distribution on $\mathscr{X}_i$, and $[\alpha]_{j\|i}$ is the conditional distribution on $\mathscr{X}_j$ given a point in $\mathscr{X}_i$; $[\alpha]_{i_1,\ldots,i_m}$ and $[\alpha]_{j_1,\ldots,j_l\|i_1,\ldots,i_m}$ are defined in an analogous way.
Distribution of X	The probability measure induced by a random variable X on the space S in which X takes values, also called distribution induced by X.
$\mathscr{F}/\mathscr{G}$-measurable map	for $f : \mathscr{X} \to \mathscr{Y}$, $(\mathscr{X}, \mathscr{F})$ and $(\mathscr{Y}, \mathscr{G})$ measurable spaces, $\{x : f(x) \in B\} \in \mathscr{F}$ for all $B \in \mathscr{G}$.
Level set	for $F : S \to [0, \infty]$, a set of the form $\{x \in S : F(x) \leq M\}$
$\alpha \mapsto f(\alpha)$	the function on space S that maps points $\alpha \in S$ to $f(\alpha)$.
σ^T	the transpose of a vector or a matrix.
$\mathrm{tr}(A)$	$\sum_{i=1}^k a_{ii}$, the trace of a square $k \times k$ matrix $A = (a_{ij})_{i,j=1}^k$.

Spaces of functions (with S a metric space)

$\mathscr{AC}([0,T]:\mathbb{R}^d)$	the space of absolutely continuous functions from $[0,T]$ to $\mathbb{R}^d$, a subspace of $\mathscr{C}([0,T]:\mathbb{R}^d)$.
$\mathscr{AC}_x([0,T]:\mathbb{R}^d)$	the subset of $\mathscr{AC}([0,T]:\mathbb{R}^d)$ with initial condition $\phi(0)=x$.
$\mathscr{C}([0,T]:S)$	the space of continuous functions from $[0,T]$ to S with the supremum norm.
$\mathscr{C}_b(S)$	the space of bounded continuous functions from S to $\mathbb{R}$.
$\mathscr{C}_c(S)$	the space of continuous functions with compact support from S to $\mathbb{R}$.
$\mathscr{D}([0,T]:S)$	the space of functions that are right continuous with limits from the left for all $t\in(0,T]$, with the Skorohod metric.
$\mathscr{L}^1([0,T]:\mathbb{R}_+)$	the space of integrable functions from $[0,T]$ to $\mathbb{R}_+$.
$\mathscr{L}^2([0,T]:\mathbb{R}^d)$	the space of square integrable functions from $[0,T]$ to $\mathbb{R}^d$.
$\mathscr{L}^0([0,T]:\mathbb{R}_+)$	the space of Borel measurable functions from $[0,T]$ to $[0,\infty)$.
$\mathscr{M}_b(S)$	the space of bounded measurable functions from S to $\mathbb{R}$.

Controls and Spaces of Controls

In Chaps. 3 and 8–13, many different spaces of controls are used, and frequently several different spaces are given the same notation. In presenting representations for functionals of a finite dimensional Brownian motion in Sect. 3.2, spaces $\mathscr{A}$ and $\bar{\mathscr{A}}$ are introduced. These denote the collection of all $\mathscr{G}_t$-progressively [resp. $\mathscr{F}_t$-progressively] measurable processes $\{v(t)\}_{0\le t\le T}$ that satisfy the integrability condition $E[\int_0^T \|v(t)\|^2 dt]<\infty$. Here $\mathscr{F}_t$ is a general filtration, and $\mathscr{G}_t$ is the (augmentation of the) filtration generated by the Brownian motion. This section also introduces the subsets of $\mathscr{A}$ denoted by $\mathscr{A}_{b,M}$ and $\mathscr{A}_b$. The first consists of $v\in\mathscr{A}$ such that $\int_0^T \|v(t)\|^2 dt\le M$ a.s. and $\mathscr{A}_b=\cup_{M=1}^\infty \mathscr{A}_{b,M}$.

In Sect. 3.3, in the study of a process, the same notation is used for somewhat different spaces. Specifically, $\mathscr{A}$ is the collection of nonnegative predictable processes, while $\mathscr{A}_{b,M}$ is the subset of $\mathscr{A}$ consisting of φ such that $\int_0^T \ell(\varphi(s))ds\le M$ a.s. and for some $K\in(0,\infty)$ (possibly depending on φ), $K^{-1}\le\varphi\le K$ a.s. Once more, $\mathscr{A}_b=\cup_{M=1}^\infty\mathscr{A}_{b,M}$.

In Chap. 8, we begin with the general theory for continuous time processes. In Sect. 8.1, $\bar{\mathscr{A}}$ denotes the class of $\mathscr{H}_0$-valued $\mathscr{F}_t$-predictable processes v that satisfy

$$P\left\{\int_0^T \|v(s)\|_0^2 ds<\infty\right\}=1,$$

and $\mathscr{A}$ denotes the subset comprising those that are predictable with respect to $\{\mathscr{G}_t\}_{0\le t\le T}$, where $\mathscr{H}_0$ is a Hilbert space with norm $\|\cdot\|_0$ and $\mathscr{F}_t$ and $\mathscr{G}_t$ are similar to their counterparts in Chap. 3. Also, $\mathscr{A}_{b,M}$ consists of $v\in\mathscr{A}$ such that $\int_0^T \|v(s)\|_0^2 ds\le M$ and $\mathscr{A}_b\doteq\cup_{M\in\mathbb{N}}\mathscr{A}_{b,M}$. Also, $\mathscr{A}_s$ denotes the subset of $\mathscr{A}_b$ consisting of all simple processes. The spaces $\bar{\mathscr{A}}_{b,M}$ [resp. $\bar{\mathscr{A}}_b$] are defined exactly like $\mathscr{A}_{b,M}$ [resp. $\mathscr{A}_b$], except that $\{\mathscr{G}_t\}$ is replaced by $\{\mathscr{F}_t\}$.

In Sect. 8.2, we turn to a representation for a Poisson random measure (PRM). In this section, $\mathscr{A}$ is the class of all $(\mathscr{P}\mathscr{F} \otimes \mathscr{B}(\mathscr{X}))\backslash\mathscr{B}[0,\infty)$-measurable maps $\varphi : \mathscr{X}_T \times \bar{\mathbb{M}} \to [0,\infty)$. Here $\mathscr{X}$ is the point space associated with the PRM, and $\mathscr{P}\mathscr{F}$ is the predictable σ-field. Also,

$$\mathscr{A}_{b,M} \doteq \{\varphi \in \mathscr{A} : L_T(\varphi) \le M \text{ a.e. and for some } n \in \mathbb{N},\ n \ge \varphi(t,x,\omega) \ge 1/n \\ \text{and } \varphi(t,x,\omega) = 1 \text{ if } x \in K_n^c, \text{ for all } (t,\omega) \in [0,T] \times \bar{\mathbb{M}}\}, \tag{E.3}$$

where

$$L_T(\varphi)(\omega) \doteq \int_{\mathscr{X}_T} \ell(\varphi(t,x,\omega))\, \nu_T(dt \times dx),\ \ \omega \in \bar{\mathbb{M}}$$

and $\{K_n\}_{n\in\mathbb{N}}$ is an increasing sequence of compact subsets of $\mathscr{X}$ such that $\cup_{n=1}^{\infty} K_n = \mathscr{X}$. As before, $\mathscr{A}_b = \cup_{M=1}^{\infty} \mathscr{A}_{b,M}$. Once more, we let $\bar{\mathscr{A}}_{b,M}$, $\bar{\mathscr{A}}$, and $\bar{\mathscr{A}}_b$ denote the analogous spaces of controls when the canonical filtration $\{\mathscr{G}_t\}$ is replaced by $\{\mathscr{F}_t\}$. In this section, we also consider simple processes. A process $\varphi \in \mathscr{A}_{b,M}$ is in the set $\mathscr{A}_{s,M}$ if the following holds. There exist $n, \ell, n_1, \dots, n_\ell \in \mathbb{N}$; a partition $0 = t_0 < t_1 < \cdots < t_\ell = T$; for each $i = 1, \dots, \ell$, a disjoint measurable partition E_{ij} of K_n, $j = 1, \dots, n_i$; $\mathscr{G}_{t_{i-1}}$-measurable random variables X_{ij}, $i = 1, \dots, \ell$, $j = 1, \dots, n_i$, such that $1/n \le X_{ij} \le n$; and

$$\varphi(t,x,\bar{m}) = 1_{\{0\}}(t) + \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} 1_{(t_{i-1},t_i]}(t) X_{ij}(\bar{m}) 1_{E_{ij}}(x) + 1_{K_n^c}(x) 1_{(0,T]}(t). \tag{E.4}$$

We define $\mathscr{A}_s \doteq \cup_{M=1}^{\infty} \mathscr{A}_{s,M}$.

In Sect. 8.3, we consider a representation for functionals of both PRM and Brownian motion. This representation involves both types of controls appearing in Sects. 8.1 and 8.2, and therefore we need to modify the notation. We denote by $\bar{\mathscr{A}}^W$ and $\bar{\mathscr{A}}_b^W$ the collections of controls for the Wiener process that were denoted by $\bar{\mathscr{A}}$ and $\bar{\mathscr{A}}_b$ in Sect. 8.1, and by $\bar{\mathscr{A}}^N$, $\bar{\mathscr{A}}_b^N$ the controls for the PRM that were denoted by $\bar{\mathscr{A}}$ and $\bar{\mathscr{A}}_b$ in Sect. 8.2. Similarly, the classes $\bar{\mathscr{A}}_{b,M}^N$ and $\bar{\mathscr{A}}_{b,M}^W$, which give uniform (in ω) bounds, are defined as they were in Sects. 8.1 and 8.2, respectively. Also we let $\bar{\mathscr{A}}_{b,M} \doteq \bar{\mathscr{A}}_{b,M}^W \times \bar{\mathscr{A}}_{b,M}^N$, $\bar{\mathscr{A}}_b \doteq \bar{\mathscr{A}}_b^W \times \bar{\mathscr{A}}_b^N$ and $\bar{\mathscr{A}} \doteq \bar{\mathscr{A}}^W \times \bar{\mathscr{A}}^N$.

The notation in Chaps. 9 and 10 for control spaces is same as that in Sect. 8.3, since we work here with systems that have both types of noise terms. Section 9.2.2, which studies a moderate deviation principle, introduces also a new specialized type of control space $\mathscr{U}_{n,+}^{\varepsilon}$ (see (9.8)) that is the class of controls for both types of noise for which the cost scales proportionally with $a(\varepsilon)^2$.

In Chap. 11, we consider systems with different types of infinite dimensional Brownian motions, and therefore the superscript W in the notation of control spaces is dropped. In considering Brownian sheet-driven systems, we consider the class of control $\bar{\mathscr{A}}$ analogous to those in Sect. 8.1 as the class of all $\{\mathscr{F}_t\}$-predictable processes f such that $\int_{[0,T]\times O} f^2(s,x)ds\,dx < \infty$ a.s. Here predictable processes are functions of (t,x,ω) (see Definition 11.7). Classes $\mathscr{A}_b$, $\mathscr{A}$ and $\bar{\mathscr{A}}_b$ are defined

similarly. In this chapter, we also use controls associated with a Hilbert space valued Brownian motion with $\mathscr{H}_0 = l_2$. The spaces of these controls, as in Sect. 8.1, are denoted once more by $\mathscr{A}_b$, $\mathscr{A}$, $\bar{\mathscr{A}}$, and $\bar{\mathscr{A}}_b$. It is made clear at each place they appear which space is intended.

Chapter 12 uses the notation $\bar{\mathscr{A}}$ and $\bar{\mathscr{A}}_b$ (and $\bar{\mathscr{A}}_{b,M}$) for controls as in Sect. 8.1 with $\mathscr{H}_0 = l_2$.

In Chap. 13, we consider systems driven by a PRM, and therefore in denoting spaces of controls, we drop the superscript N. Thus the space $\mathscr{A}$ and its variants are as in Sect. 8.2. In studying a moderate deviation principle, the space $\mathscr{U}^{\varepsilon}_{n,+}$, which was introduced in Sect. 9.2.2, also makes an appearance in Sect. 13.3.2.

Together with the spaces of random controls such as $\mathscr{A}$, we also use many spaces of deterministic controls, employing once again the same notation for several different spaces. In Sect. 3.2, where we consider representations for a k-dimensional Brownian motion, the notation S_M is used for the space

$$\left\{\phi \in \mathscr{L}^2([0,T] : \mathbb{R}^k) : \int_0^T \|\phi(s)\|^2 \, ds \leq M\right\},$$

while in Sect. 3.3, in the study of a process, we have

$$S_M = \left\{\phi \in \mathscr{L}^0([0,T] : \mathbb{R}_+) : \int_0^T \ell(\phi(s)) ds \leq M\right\}.$$

In Sect. 8.1, where we consider a Hilbert space valued Brownian motion, we use the notation S_M for the space

$$\left\{u \in \mathscr{L}^2([0,T] : \mathscr{H}_0) : \int_0^T \|u(s)\|_0^2 ds \leq M\right\}.$$

In Chap. 9, where we consider systems that have both types of noise terms, we need to distinguish the two types of control spaces. The space S_n from Sect. 8.1 is denoted here by S_n^W, and we define

$$S_n^N \doteq \left\{g : \mathscr{X}_T \to [0,\infty) : L_T^N(g) \leq n\right\},$$

where L_T^N is as in (9.1). We define $S_n \doteq S_n^W \times S_n^N$ and $S \doteq \cup_{n\in\mathbb{N}} S_n$. This chapter also uses two other specialized deterministic control spaces. The first corresponds to controls that hit a target ϕ for a given z in the abstract large deviation principle of Sect. 9.2.1, namely

$$S^{\mathscr{G}}_{z,\phi} \doteq \left\{(f,g) \in S : \phi = \mathscr{G}^0(z, \int_0^{\cdot} f(s)ds, \nu_T^g)\right\},$$

while the second is a similar space in the abstract moderate deviation principle of Sect. 9.2.2,

$$S^{\mathscr{K}}_{z,\eta} \doteq \{q = (f_1, f_2) \in \mathscr{L}^2 : \eta = \mathscr{K}^0(z, q)\},$$

where $\mathscr{L}^2$ is introduced below (9.8).

Section 9.2.2 introduces three additional spaces of controls that are needed for the proof of the moderate deviation principle. These are

$$S^{N,\varepsilon}_{n,+} \doteq \{g : \mathscr{X}_T \to \mathbb{R}_+ \text{ such that } L^N_T(g) \le na^2(\varepsilon)\},$$
$$S^{N,\varepsilon}_{n} \doteq \{f : \mathscr{X}_T \to \mathbb{R} \text{ such that } f = (g-1)/a(\varepsilon), \text{ with } g \in S^{N,\varepsilon}_{n,+}\},$$

and

$$\hat{S}_n \doteq \{(f_1, f_2) \in \mathscr{L}^2 : \|f_1\|^2_{W,2} + \|f_2\|^2_{N,2} \le n\},$$

where $a(\varepsilon)$ is the scaling sequence in (9.6), and the norms $\|\cdot\|_{W,2}$ and $\|\cdot\|_{N,2}$ are introduced below (9.8).

Chapter 10 uses the same notation as Chap. 9 for the various spaces of deterministic controls.

In Chap. 11, where we consider systems with different types of infinite dimensional Brownian motions, the superscript W in the notation S^W_n is dropped. In particular, either, the space S_n denotes the space in Sect. 8.1 with $\mathscr{H}_0 = l_2$, or it is the space

$$\left\{\phi \in \mathscr{L}^2([0,T] \times O) : \int_{[0,T]\times O} \phi^2(s,r)dsdr \le n\right\},$$

where O is the open set from Sect. 11.1. The precise space that is being referred to is clear from the context.

In Chap. 12, S_n is the space in Sect. 8.1 with $\mathscr{H}_0 = l_2$.

The spaces of Sect. 9.2.2 appear in Sect. 13.3.2 once more. However, since there is no Brownian motion in the dynamics, the spaces and notations are slightly different. Specifically,

$$\hat{S}_n \doteq \left\{f = \{f_i\}_{i=1}^K : f_i \in \mathscr{L}^2([0,1] \times \mathbb{R}_+) \text{ and } \sum_{i=1}^K \int_{[0,1]\times\mathbb{R}_+} f_i^2(s,y)dyds \le n\right\}$$

and

$$S^{\varepsilon}_{n,+} \doteq \left\{g = \{g_i\}_{i=1}^K : g_i : [0,1] \times \mathbb{R}_+ \to \mathbb{R}_+ \right.$$
$$\left. \text{and} \sum_{i=1}^K \int_{[0,1]\times\mathbb{R}_+} \ell(g_i(s,y))dyds \le na^2(\varepsilon)\right\}.$$

Norms and Distances

$\|x\|$	$(\sum_{i=1}^d x_i^2)^{1/2}$ for $x \in \mathbb{R}^d$.
$\|F\|_\infty$	$\sup_{x\in S} \|F(x)\|$ for $F : S \to \mathbb{R}^d$.

$\|F\|_{\infty,T}$	$\sup_{0\le t\le T}\|F(s)\|$ for $F:[0,T]\to\mathbb{R}^d$.
$\|\gamma\|_{\mathrm{TV}}$	for a signed measure γ on a measurable space $(S,\mathscr{F})$, $\sup\left\|\int_S f(x)\,\gamma(dx)\right|$, where the supremum is over $f\in\mathscr{M}_b(S)$ with $\|f\|_\infty\le 1$, called the total variation norm of γ.
$d_{\mathrm{BL}}(\nu_1,\nu_2)$	for probability measures ν_1 and ν_2 on a Polish space $(\mathscr{X},d)$, $\sup\left\|\int_{\mathscr{X}} f(x)\nu_1(dx)-\int_{\mathscr{X}} f(x)\nu_2(dx)\right|$, where the supremum is over f with $\|f\|_\infty\le 1$ and $\|f(x)-f(y)\|\le d(x,y)$ for all $x,y\in\mathscr{X}$, called the Dudley metric or bounded-Lipschitz metric.
$\|f\|_1$	for measurable f on a measure space $(S,\mathscr{F},\lambda)$, $\int_S \|f\|\,d\lambda$, called the $\mathscr{L}^1$-norm.
$\|\Lambda\|$	for a bounded linear operator Λ on a Hilbert space $\mathscr{H}$, $\sup_{h\in\mathscr{H}:\|h\|=1}\|\Lambda h\|$, called the operator norm.
$\|\psi\|_\alpha$	for $\alpha\in(0,1)$ and $\psi:S\to\mathbb{R}$, (S,d) a metric space, $\sup\{\|\psi(x)-\psi(y)\|^\alpha/d(x,y), x,y\in S\}$, called the α-Hölder norm.

Abbreviations

a.s.	almost surely
CONS	complete orthonormal sequence
DPR	direct probability redistribution
iid	independent and identically distributed
HJB	Hamilton–Jacobi–Bellman
IS	importance sampling
LLN	law of large numbers
LDP	large deviation principle
MDP	moderate deviation principle
PDE	partial differential equation
PRM	Poisson random measure
RCLL	right continuous with left limits
RESTART	repetitive simulation trials after reaching threshold
r.c.p.d.	regular conditional probability distribution
SDE	stochastic differential equation
SPDE	stochastic partial differential equation
w.p.1	with probability 1
WSLQ	weighted serve the longer queue

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0

Specialized Symbols

$\mathscr{D}^m$	group of $\mathscr{C}^m$ diffeomorphisms, page 311
$(-\infty, x]$	$\{y : y_i \le x_i \text{ for all } i = 1, \ldots, d\}$, page 289
$(\mathscr{H}, \langle \cdot, \cdot \rangle)$	a real separable Hilbert space, page 202
$\bar{V}^\delta$	mollification of $\bar{V}$, page 389
$\bar{\delta}_{Z_i}$	a random measure associated with branching processes, page 427
$\bar{\nu}_T$	$\lambda_T \times \nu \times \lambda_\infty$, page 215
$\bar{L}^n$	controlled empirical measure on $\bar{X}_i^n, i = 1, \ldots, n$, page 47
$\bar{l}_2$	a weighted l_2 space, page 283
$\bar{L}_T(u)$	$L_T^W(\psi) + L_T^N(\varphi)$, page 231
$\bar{M}^n(dw \times dt)$	random measures used in the analysis of a MDP, page 121
$\bar{N}$	an augmented PRM, page 215
$\bar{r}(x, t; v)$	controlled feedback jump rates in WSLQ, page 343
$\bar{U}(x)$	$j\Delta$ if $x \in C_j \backslash C_{j-1}$, page 428
$\bar{U}_k$	$\bar{U}(x)$ if $\sigma(x) = k$, page 428
$\bar{w}^n(t)$	mean of control measure $\bar{\mu}_i^n$ for $t \in [1/n, 1/n + 1/n)$, page 121
$\bar{X}_i^n$	controlled random variables whose conditional distribution is $\bar{\mu}_i^n$, page 46
$\kappa_1(\beta), \bar{\kappa}_1(\beta), \kappa_2(\beta), \kappa_3$	quantities used to describe properties of ℓ, page 238
$\Delta(\mathscr{X})$	$\{\gamma \in \mathscr{P}(\mathscr{X}) : R(\gamma \| \theta) < \infty\}$, page 31
Δ_ψ	the mapping from $\mathscr{P}(\mathscr{X})$ into $\mathscr{P}(\mathscr{Y})$ defined by $\beta = \Delta_\psi \alpha$ when $\beta(A) = \alpha(x : \psi(x) \in A)$ for Borel sets A, page 34
ℓ	the function $\ell(b) = b \log b - b + 1, b \ge 0$, page 54
$\gamma(dy \,\|x\, , \beta)$	exponentially tilted version of $\theta(dy \,\|x\,)$, page 93
$\hat{\mathscr{W}}_m$	$\mathscr{C}\big([0, T] : \mathscr{D}^m\big)$, page 311
$\hat{A}_b$	a collection of bounded predictable processes, page 219

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0

$\kappa(t)$	$\max\{0 \le k \le K : t_k \le t\}$, page 344
Λ	urn type index for occupancy models, page 177
$\lambda_k(l)$	initializing distribution for splitting, page 433
$\langle h, k\rangle_0$	inner product on $\mathscr{H}_0$, page 202
$\|\alpha\|_A^2$	$\langle \alpha, A\alpha\rangle$, page 117
$\{\bar{\mu}_i^n\}$	random control probability measures, page 46
$\{v_i(x)\}$	iid random vector fields, page 77
$\mathbb{M}$	$\Sigma(\mathscr{Y}_T)$, sample space for an augmented PRM, page 215
$\bar{\mathbb{V}}$	$\mathbb{W} \times \bar{\mathbb{M}}$, page 234
$\mathbb{B}_0^T$	space of all continuous maps from $[0, T] \times \bar{O}$ to $\mathbb{R}$ endowed with the sup–norm, page 293
$\mathbb{B}_\alpha$	Banach space of α-Hölder functions on O, page 293
$\mathbb{B}_\alpha([0, T] \times O)$, $\mathbb{B}_\alpha^T$	Banach space of α-Hölder functions on $[0, T] \times O$, page 293
$\mathbb{H}(p)$	$-H(-p)$, page 382
$\mathbb{H}(x, p)$	$-H(x, -p)$, page 384
$\mathbb{M}$	$\Sigma(\mathscr{X}_T)$, sample space for a PRM, page 215
$\mathbb{R}^\infty$	space of real valued sequences equipped with product topology, page 282
$\mathbb{U}$	range space for the abstract LD and MD results of Chap. 9, page 236
$\mathbb{V}$	$\mathbb{W} \times \mathbb{M}$, page 234
$\mathbb{W}$	$\mathscr{C}([0, T] : \mathscr{H}_0)$, page 234
$\mathscr{D}_a$	feasible domain for an occupancy problem, page 190
$\mathscr{E}^\varphi(t)$	an exponential martingale associated with PRM, page 218
$\mathscr{F}(x, t; \omega, T)$	a collection of probability vectors with certain properties, page 192
$\mathscr{G}^\varepsilon$	measurable maps used in the abstract LDP of Chap. 9, page 236
$\mathscr{H}_0$	$\Lambda^{1/2}\mathscr{H}$, with Λ a symmetric, strictly positive, trace class operator, page 202
$\mathscr{K}^\varepsilon$	measurable maps used in the abstract MDP of Chap. 9, page 238
$\mathscr{L}_{\exp}$	$\cap_{\rho\in(0,\infty)}\mathscr{L}_{\exp}^\rho$, page 251
$\mathscr{L}_{\exp}^\rho$	functions that satisfy an exponential integrability assumption, page 251
$\mathscr{N}$	a collection of simple form absolutely continuous paths, page 341
$\mathscr{P}(\Lambda)$	the probabilities on $0, 1, \ldots, J+1$, identified with the simplex in $\mathbb{R}^{J+2}$, page 176
$\mathscr{P}\mathscr{F}$	predictable σ-field, page 202
$\mathscr{V}$	set of possible jump vectors for the WSLQ model, page 332
$\mathscr{W}_m$	$\mathscr{C}\big([0, T] : \mathscr{C}^m(\mathbb{R}^d)\big)$, page 311
$\mathscr{X}$	space of types for a PRM, page 214

$\mathscr{X}_T$	$[0, T] \times \mathscr{X}$, page 214
$\mathscr{Y}$	augmented space of types for a controlled PRM, page 215
$\mathscr{Y}_T$	$[0, T] \times \mathscr{Y}$, page 215
$\mathfrak{S}^n(\bar{V})$	second moment of an estimator based on $\bar{V}$, page 388
$\mathrm{Int}(u)$	integrated version of a control u, page 289
ν	measure on the space of types for a PRM, page 214
ν_T	$\lambda_T \times \nu$, page 214
ν_T^g	measure defined by $\int_A g(s,x)\, \nu_T(ds \times dx)$, $A \in \mathscr{B}(\mathscr{X}_T)$, page 235
$\omega(x, \delta)$	modulus of continuity, page 300
$\pi(x)$	indices that maximize the weighted queue length, page 332
ρ_j^δ	weights in the implementation of schemes based on mollified piecewise smooth subsolutions, page 389
$\rho_k(x, t)$	probabilities for ball placement in occupancy models, page 176
$\Sigma(\mathscr{S})$	measures ν on $(\mathscr{S}, \mathscr{B}(\mathscr{S}))$ satisfying $\nu(K) < \infty$ for every compact $K \subset \mathscr{S}$, page 214
$\sigma(x)$	unique integer j such that $x \in C_j \backslash C_{j-1}$, page 428
$\tau(dy \,\vert x)$	a stochastic kernel in dy given x, page 32
$\theta(\cdot \,\vert x)$	distribution of iid random vector fields $\{v_i(x)\}$, page 77
$\theta \otimes \sigma(A \times B)$	$\int_{A\times B} \theta(dx)\, \sigma(dy \,\vert x) = \int_A \sigma(B \,\vert x)\, \theta(dx)$, page 38
$\Vert f \Vert_{\infty,t}$	$\sup_{0\le s\le t} \Vert f(s) \Vert$, page 252
$\Vert \cdot \Vert_0$	norm on the Hilbert space $\mathscr{H}_0$, page 202
$\Vert \cdot \Vert_{N,2}$	norm in the Hilbert space $\mathscr{L}^2(\nu_T)$, page 239
$\Vert \cdot \Vert_{W,2}$	norm in $\mathscr{L}^2([0, T] : \mathscr{H}_0)$, page 239
$\{\mathscr{F}_t\}$	a general filtration, page 56
$\{\mathscr{G}_t\}$	a filtration generated by driving noises, page 56
$\{U(t, s)\}$	a two parameter semigroup, page 292
$A(\mu)$	the probability measures on $S \times S$ with both marginals μ, page 149
$a(\varepsilon)$ and $\varkappa(\varepsilon)$	functions used in the statement of the abstract MDP of Chap. 9, page 238
$a(n)$	scaling sequence used in an MDP, page 115
$A_\kappa^{-1}(x)$	matrix obtained by truncating the eigenvalues of $A^{-1}(x)$ at κ^2, page 139
C_j	splitting thresholds, page 425
$C_{x,T}$	collection of paths starting from x and reaching a set B before reaching A, by time T, page 373
C_x	$\cup_{T\in(0,\infty)} C_{x,T}$, page 373
$Db(x)$	matrix of first order partial derivatives, page 119
$G(t, s, r, q)$	kernel of a two parameter semigroup, page 292
$H(\alpha)$	a log moment generating function, page 52
$H(x, \alpha)$	log moment generating function of iid random vector fields $\{v_i(x)\}$, page 78

$H^{(i)}(\alpha)$	$\mu_i(e^{-\alpha_i} - 1) + \sum_{j=1}^{d} \lambda_j(e^{\alpha_j} - 1)$, page 333
$H^*(x, \beta)$	the Legendre-Fenchel transform of $H(x, \alpha)$, also sometimes denoted as $L(x, \beta)$, page 90
H^A	$\max_{i \in A} H^{(i)}$, page 334
$H_c(x, \alpha)$	the centered log cumulant generating function, page 117
$I(A)$	$\inf_{x \in A} I(x)$, page 3
I_M	rate function in a moderate deviation principle, page 119
$L(\beta)$	the Legendre-Fenchel transform of $H(\alpha)$, page 55
$L(x, \beta)$	the Legendre-Fenchel transform of $H(x, \alpha)$, also sometimes denoted as $H^*(x, \beta)$, page 81
$L^{(i)}$	L^A when $A = \{i\}$, page 334
L^A	Legendre transform of H^A, page 334
L^n	the empirical measure on X_i, $i = 1, \ldots, n$, page 47
l_2	Hilbert space of square summable sequences, page 283
$L_c(x, \beta)$	the Legendre-Fenchel transform of $H_c(x, \alpha)$, page 119
$L_{i,m}$	support threshold of particle m at time i, page 433
N^φ	Poisson random measure with controlled intensity governed by φ, page 215
N_c^1	the compensated version of N^1, page 218
$Q(j, k)$	random vector defined in terms of $q_l(j, k)$, page 439
$Q(t)$	vector of queue lengths at time t, page 332
$q^{(k)}(x, dy)$	k-step transition probability kernel for $q(x, dy)$, page 150
$q_l(j, k)$	splitting vectors, page 430
$R(\gamma \| \theta)$	relative entropy of γ with respect to θ, page 29
$r(x, v)$	jump rate for the WSLQ model, page 332
$R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,Tn-1})$	likelihood ratio in an importance sampling estimator, page 389
R_j	splitting rates, page 425
$s^n(t)$	$\lfloor nt \rfloor / n$, page 124
$W^\psi(t)$	controlled Brownian motion $W(t) + \int_0^t \psi(s)ds$, page 231
$w^n(\delta)$	modulus of continuity, page 85
$w^n(t)$	rescaled mean of control measure $\bar{\mu}_i^n$ for $t \in [1/n, 1/n + 1/n)$, page 121
$L_T^N(\varphi)$	cost function for a PRM, page 231
$L_T^W(\psi)$	cost function for a Hilbert space valued Brownian motion, page 231

References

1. R.A. Adams, J.J.F. Fournier, *Sobolev Spaces*, 2nd edn. (Academic, New York, 2003)
2. M. Alanyali, B. Hajek, On large deviations of Markov processes with discontinuous statistics. Ann. Appl. Probab. **8**, 45–66 (1998)
3. D. Aldous, Stopping times and tightness. Ann. Prob. **6**, 335–340 (1978)
4. D.F. Anderson, T. Kurtz, *Stochastic Analysis of Biochemical Systems*, Mathematical Biosciences Institute Graduate Lecture (Springer, New York, 2015)
5. G. Ben Arous, F. Castell, Flow decomposition and large deviations. J. Funct. Anal. **140**, 23–67 (1995)
6. S. Asmussen, P.W. Glynn, *Stochastic Simulation: Algorithms and Analysis*. Applications of Mathematics. (Springer Science+Business Media, LLC, Berlin, 2007)
7. R. Azencott, Petits perturbations aléatoires des systems dynamiques: Développments asymptotiques. Bull. des Sci. Math. **109**, 253–308 (1985)
8. R. Azencott, G. Ruget, Mélanges d'équations differentielles et grand écart à la loi des grandes nombres. Z. Wahrsch. Verw. Gebiete **38**, 1–54 (1977)
9. R.R. Bahadur, R. Ranga Rao, On deviations of the sample mean. Ann. Math. Stat. **31**(4), 1015–1027 (1960)
10. P. Baldi, M. Sanz-Solé, Modulus of continuity for stochastic flows, in *Progress in Probability*, vol. 32 (Birkhauser, Basel, 1993)
11. S. Banerjee, A. Budhiraja, M. Perlmutter, Large deviations from the hydrodynamic limit for a system with nearest neighbor interactions. Math. arXiv:1803.09344
12. J. Bao, C. Yuan, Large deviations for neutral functional SDEs with jumps. Stoch. Int. J. Probab. Stoch. Process. **87**(1), 48–70 (2015)
13. V. Barbu, T. Precupanu, *Convexity and Optimization in Banach Spaces* (Springer, Berlin, 2012)
14. M. Bardi, I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations* (Birkhäuser, Basel, 1997)
15. G. Barles, An approach of deterministic control problems with unbounded data. Ann. Inst. Henri Poincaré: Anal. Non linéaire **7**, 235–258 (1990)
16. P. Baxendale, Brownian motions in the diffeomorphism group. I. Compos. Math. **53**(1), 19–50 (1984)
17. P. Baxendale, Asymptotic behaviour of stochastic flows of diffeomorphisms. In *Stochastic Processes and Their Applications* (Springer, Berlin, 1986), pp. 1–19
18. A. Benveniste, M. Metivier, P. Priouret, *Adaptive Algorithms and Stochastic Approximation* (Springer, Berlin, 1990)
19. D. Bertsekas, S. Shreve, *Stochastic Optimal Control: The Discrete Time Case* (Academic, San Diego, 1978)

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0

20. H. Bessaih, A. Millet, Large deviation principle and inviscid shell models. Electron. J. Probab. **14**(89), 2551–2579 (2009)
21. H. Bessaih, A. Millet, Large deviations and the zero viscosity limit for 2D stochastic Navier-Stokes equations with free boundary. SIAM J. Math. Anal. **44**(3), 1861–1893 (2012)
22. H. Bessaih, A. Millet, On stochastic modified 3d Navier-Stokes equations with anisotropic viscosity. J. Math. Anal. Appl. **462**(1), 915–956 (2018)
23. S. Bhamidi, A. Budhiraja, P. Dupuis, R. Wu, Large deviation principle for the exploration process of the configuration model (2017), arXiv:1708.01832
24. P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968)
25. J.M. Bismut, Mécanique Aléatoire. Lecture Notes in Mathematics, vol. 866 (1981)
26. J. Blanchet, H. Lam, State-dependent importance sampling for rare-event simulation: an overview and recent advances. Surv. Oper. Res. Manag. Sci. **17**(1), 38–59 (2012)
27. J.H. Blanchet, P. Glynn, Efficient rare-event simulation for the maximum of heavy-tailed random walks. Ann. Appl. Prob. **18**, 1351–1378 (2008)
28. J.H. Blanchet, P. Glynn, K. Leder, On Lyapunov inequalities and subsolutions for efficient importance sampling. ACM Trans. Model. Comput. Simul. **22**(3), 13:1–13:27 (2012)
29. E. Bolthausen, Markov process large deviations in τ- topology. Stoch. Proc. Appl. **25**, 95–108 (1987)
30. T.E. Booth, J. Hendricks, Importance estimation in forward Monte Carlo calculations. Nucl. Tech./Fusion **6**, 90–100 (1984)
31. C. Borell, Diffusion equations and geometric inequalities. Potential Anal. **12**(1), 49–71 (2000)
32. M. Boué, P. Dupuis, A variational representation for certain functionals of Brownian motion. Ann. Prob. **26**, 1641–1659 (1998)
33. M. Boué, P. Dupuis, R.S. Ellis, Large deviations for small noise diffusions with discontinuous statistics. Prob. Theor. Rel. Fields **116**, 125–148 (2000)
34. L. Breiman, *Probability Theory* (Addison-Wesley, Reading, 1968)
35. P. Brémaud, *Point Processes and Queues: Martingale Dynamics* (Springer, Berlin, 1981)
36. W. Bryc, Large deviations by the asymptotic value method, in *Diffusion Processes and Related Problems in Analysis*. Progress in Probability, vol. I (Evanston, IL, 1989), pp. 447–472
37. Z. Brzeźniak, B. Goldys, T. Jegaraj, Large deviations and transitions between equilibria for stochastic Landau-Lifshitz-Gilbert equation. Arch. Ration. Mech. Anal. **226**(2), 497–558 (2017)
38. A. Budhiraja, J. Chen, P. Dupuis, Large deviations for stochastic partial differential equations driven by a Poisson random measure. Stoch. Proc. Appl. **123**, 523–560 (2013)
39. A. Budhiraja, P. Dupuis, A variational representation for positive functionals of infinite dimensional Brownian motion. Prob. Math. Stat. **20**, 39–61 (2000)
40. A. Budhiraja, P. Dupuis, Large deviations for the empirical measures of reflecting Brownian motion and related constrained processes in $\mathbb{R}_+$. Elec. J. Probab. **8**, 1–46 (2003)
41. A. Budhiraja, P. Dupuis, A. Ganguly, Moderate deviation principles for stochastic differential equations with jumps. Ann. Probab. **44**, 1723–1775 (2016)
42. A. Budhiraja, P. Dupuis, A. Ganguly. Large deviations for small noise diffusions in a fast Markovian environment. Electron. J. Probab. **23** (2018)
43. A. Budhiraja, P. Dupuis, V. Maroulas, Large deviations for infinite dimensional stochastic dynamical systems. Ann. Probab. **36**, 1390–1420 (2008)
44. A. Budhiraja, P. Dupuis, V. Maroulas, Large deviations for stochastic flows of diffeomorphisms. Bernoulli J. **16**, 234–257 (2010)
45. A. Budhiraja, P. Dupuis, V. Maroulas, Variational representations for continuous time processes. Ann. de l'Inst. H. Poincaré **47**, 725–747 (2011)
46. A. Budhiraja, W-T. Fan, R. Wu, Large deviations for Brownian particle systems with killing. J. Theor. Probab. 1–40 (2017)
47. A. Budhiraja, P. Nyquist, Large deviations for multidimensional state-dependent shot-noise processes. J. Appl. Probab. **52**(4), 1097–1114 (2015)

48. A. Budhiraja, R. Wu, Moderate deviation principles for weakly interacting particle systems. Probab. Theory Relat. Fields **168**(3–4), 721–771 (2017)
49. A. Buijsrogge, P. Dupuis, M. Snarski, Splitting algorithms for rare event simulation over long time intervals. To appear in Ann. App. Prob.
50. C. Cardon-Weber, Large deviations for a Burgers'-type SPDE. Stoch. Process. Appl. **84**(1), 53–70 (1999)
51. F. Cerou, P. Del Moral, F. LeGland, P. Lezaud, Genetic genealogical models in rare event analysis. ALEA Lat. Am. J. Probab. Math. Stat. **1**, 181–203 (2006)
52. S. Cerrai, M. Röckner, Large deviations for stochastic reaction-diffusion systems with multiplicative noise and non-Lipschitz reaction term. Ann. Probab. **32**(1B), 1100–1139 (2004)
53. A. Charalambides, A unified derivation of occupancy and sequential occupancy distributions, in *Advances in Combinatorial Methods and Applications to Probability and Statistics* (1997), pp. 259–273
54. B. Chen, J. Blanchet, C.-H. Rhee, B. Zwart, Efficient rare-event simulation for multiple jump events in regularly varying random walks and compound Poisson processes. Math. arxiv: https://arxiv.org/abs/1706.03981
55. Y. Chen, H. Gao, Well-posedness and large deviations for a class of SPDEs with Lévy noise. J. Differ. Equ. **263**(9), 5216–5252 (2017)
56. F. Chenal, A. Millet, Uniform large deviations for parabolic SPDEs and applications. Stoch. Process. Appl. **72**(2), 161–186 (1997)
57. L. Cheng, R. Li, W. Liu, Moderate deviations for the Langevin equation with strong damping. J. Stat. Phys. **170**(5), 845–861 (2018)
58. S. Chevet, Gaussian measures and large deviations, in *Probability in Banach Spaces IV* (Springer, Berlin, 1983), pp. 30–46
59. T.-S. Chiang, A lower bound of the asymptotic behavior of some Markov processes. Ann. Probab. **10**(4), 955–967 (1982)
60. P.-L. Chow, Large deviation problem for some parabolic Itô equations. Commun. Pure Appl. Math. **45**(1), 97–120 (1992)
61. G. Christensen, Deformable shape models for anatomy. Ph.D. thesis (1994)
62. G.E. Christensen, R.D. Rabbitt, M.I. Miller, Deformable templates using large deformation kinematics. IEEE Trans. Image Process. **5**(10), 1435–1447 (1996)
63. I. Chueshov, A. Millet, Stochastic 2D hydrodynamical type systems: well posedness and large deviations. Appl. Math. Optim. **61**(3), 379–420 (2010)
64. F. Cipriano, T. Costa, A large deviations principle for stochastic flows of viscous fluids. J. Differ. Equ. **264**(8), 5070–5108 (2018)
65. J.F. Collamore, Hitting probabilities and large deviations. Ann. Probab. **24**(4), 2065–2078 (1996)
66. J.B. Conway, *A Course in Functional Analysis*, vol. 96 (Springer Science & Business Media, Berlin, 2013)
67. T.M. Cover, J.A. Thomas, *Elements of Information Theory*, 2nd edn. (Wiley, New York, 2006)
68. H. Cramér, Sur un nouveau théorème-limite de la théorie des probabilités. Actualités Scientifiques et Industrielles, 736:2–23 1938; Colloque consacré à la théorie des probabilités, vol. 3 (Hermann, Paris)
69. G. Da Prato, J. Zabczyk, *Stochastic Equations in Infinite Dimensions*. Encyclopedia of Mathematics and its Applications, vol. 44 (Cambridge University Press, Cambridge, 1992)
70. H. Dadashi, Large deviation principle for semilinear stochastic evolution equations with Poisson noise. Infin. Dimens. Anal. Quantum Probab. Relat. Top. **20**(02), 1750009 (2017)
71. H. Dadashi-Arani, B.Z. Zangeneh, Large deviation principle for semilinear stochastic evolution equations with monotone nonlinearity and multiplicative noise. Differ. Integral Equ. **23**(7–8), 747–772 (2010)
72. A. de Acosta, Large deviations for empirical measures of Markov chains. J. Theor. Proba. **3** (1990)
73. A. de Acosta, On large deviations of empirical measures in the τ-topology. J. Appl. Proba. **31**, 41–47 (1994)

74. E. De Giorgi, Sulla convergenza di alcune successioni dintegrali del tipo dellarea. Ennio De Giorgi **414** (1975)
75. A. de Oliveira Gomes, Asymptotics for FBSDEs with jumps and connections with partial integral differential equations, in *From Particle Systems to Partial Differential Equations III* (Springer, Berlin, 2016), pp. 99–120
76. T. Dean, P. Dupuis, Splitting for rare event simulation: a large deviations approach to design and analysis. Stoch. Proc. Appl. **119**, 562–587 (2009)
77. T. Dean, P. Dupuis, The design and analysis of a generalized RESTART/DPR algorithm for rare event simulation. Ann. OR **189**, 63–102 (2011)
78. P. Del Moral, J. Garnier, Genealogical particle analysis of rare events. Ann. Appl. Probab. **15**, 2496–2534 (2005)
79. C. Dellacherie, P.A. Meyer, *Probabilities and Potential B: Theory of Martingales*. North-Holland Mathematics Studies, vol. 2 (North-Holland Publishing Company, Amsterdam, 1982)
80. A. Dembo, O. Zeitouni, *Large Deviations Techniques and Applications* (Springer, New York, 1998)
81. F. den Hollander, *Large Deviations*, Fields Institute Monographs (AMS, Providence, 2000)
82. J.-D. Deuschel, D.W. Stroock, *Large Deviations* (Academic, San Diego, 1989)
83. J. Dieudonné, *Foundations of Modern Analysis*. Number v. 10, pt. 1 in Dieudonné, Jean: Treatise on analysis (Academic, New York, 1960)
84. I.H. Dinwoodie, P. Ney, Occupation measures for Markov chains. J. Theor. Probab. **8**(3), 679–691 (1995)
85. J. Doll, P. Dupuis, On performance measures for infinite swapping Monte Carlo methods. J. Chem. Phys. **142**, 024111 (2015)
86. J. Doll, P. Dupuis, P. Nyquist, A large deviations analysis of certain qualitative properties of parallel tempering and infinite swapping algorithms. Appl. Math. Opt. 1–42 (2017)
87. M.D. Donsker, S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time, I. Commun. Pure Appl. Math. **28**, 1–47 (1975)
88. M.D. Donsker, S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time, III. Commun. Pure Appl. Math. **29**, 389–461 (1976)
89. M.D. Donsker, S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time, IV. Commun. Pure Appl. Math. **36**, 183–212 (1983)
90. J. Doob, *Stochastic Processes* (Wiley, New York, 1953)
91. J. Duan, A. Millet, Large deviations for the Boussinesq equations under random influences. Stoch. Process. Appl. **119**(6), 2052–2081 (2009)
92. R.M. Dudley, *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics, vol. 74 (Cambridge University Press, Cambridge, 2002). Revised Reprint of the 1989 Original
93. N. Dunford, J.T. Schwartz, *Linear Operators Parts I, II, III* (Interscience Publishers, Geneva, 1963)
94. P. Dupuis, Large deviations analysis of some recursive algorithms with state dependent noise. Ann. Probab. **16**, 1509–1536 (1988)
95. P. Dupuis, R.S. Ellis, Large deviations for Markov processes with discontinuous statistics, II: Random walks. Probab. Theory Rel. Fields **91**, 153–194 (1992)
96. P. Dupuis, R.S. Ellis, The large deviation principle for a general class of queueing systems. I. Trans. Am. Math. Soc. **347**, 2689–2751 (1996)
97. P. Dupuis, R.S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations* (Wiley, New York, 1997)
98. P. Dupuis, R.S. Ellis, A. Weiss, Large deviations for Markov processes with discontinuous statistics, I: General upper bounds. Ann. Probab. **19**, 1280–1297 (1991)
99. P. Dupuis, U. Grenander, M. Miller, A variational formulation of a problem in image matching. Q. Appl. Math. **56**, 587–600 (1998)
100. P. Dupuis, D. Johnson, Moderate deviations for recursive stochastic algorithms. Stoch. Syst. **1**, 1–33 (2015)
101. P. Dupuis, D. Johnson, Moderate deviations based importance sampling for recursive stochastic equations. J. Appl. Probab. **49**, 981–1010 (2017)

102. P. Dupuis, H.J. Kushner, Stochastic approximation and large deviations: upper bounds and w.p.1 convergence. SIAM J. Control Optim. **27**, 1108–1135 (1989)
103. P. Dupuis, K. Leder, H. Wang, Large deviations and importance sampling for a tandem network with slow-down. QUESTA **57**, 71–83 (2007)
104. P. Dupuis, K. Leder, H. Wang, On the large deviations properties of the weighted-serve-the-longest-queue policy, in *In and Out of Equilibrium 2*, ed. by V. Sidoravicius, M.E. Vares (Birkhauser, New York, 2008)
105. P. Dupuis, K. Leder, H. Wang, Importance sampling for weighted serve-the-longest-queue. Math. Oper. Res. **34**, 642–660 (2009)
106. P. Dupuis, D. Lipshutz, Large deviations for the empirical measure of a diffusion via weak convergence methods. Stoch. Proc. Appl. **128**, 2581–2604 (2018)
107. P. Dupuis, Y. Liu, On the large deviation rate for the empirical measure of reversible pure jump Markov processes. Ann. Probab. **43**, 1121–1156 (2015)
108. P. Dupuis, Y. Liu, N. Plattner, J.D. Doll, On the infinite swapping limit for parallel tempering. SIAM J. Multiscale Model. Simul. **10**, 986–1022 (2012)
109. P. Dupuis, C. Nuzman, P. Whiting, Large deviation asymptotics for occupancy problems. Ann. Probab. **32**, 2765–2818 (2004)
110. P. Dupuis, A. Sezer, H. Wang, Dynamic importance sampling for queueing networks. Ann. Appl. Prob. **17**(4), 1306–1346 (2007)
111. P. Dupuis, K. Spiliopoulos, Large deviations for multiscale diffusions via weak convergence methods. Stoch. Proc. Appl. **122**, 1947–1987 (2012)
112. P. Dupuis, K. Spiliopoulos, H. Wang, Importance sampling for multiscale diffusions. SIAM J. Multiscale Model. Simul. **10**, 1–27 (2012)
113. P. Dupuis, K. Spiliopoulos, X. Zhou, Escaping from an attractor: importance sampling and rest points I. Ann. Appl. Probab. **25**, 2909–2958 (2015)
114. P. Dupuis, H. Wang, Importance sampling, large deviations, and differential games. Stoch. Stoch. Rep. **76**, 481–508 (2004)
115. P. Dupuis, H. Wang, Dynamic importance sampling for uniformly recurrent Markov chains. Ann. Appl. Probab. **15**, 1–38 (2005)
116. P. Dupuis, H. Wang, Subsolutions of an Isaacs equation and efficient schemes for importance sampling. Math. Oper. Res. **32**, 1–35 (2007)
117. P. Dupuis, H. Wang, Importance sampling for Jackson networks. Queueing Syst. **62**, 113–157 (2009)
118. P. Dupuis, O. Zeitouni, A nonstandard form of the rate function for the occupation measure of a Markov chain. Stoch. Proc. Appl. **61**, 249–261 (1996)
119. P. Dupuis, J. Zhang, Explicit solutions for a class of nonlinear PDE that arise in allocation problems. SIAM J. Math. Anal. **39**(5), 1627–1667 (2008)
120. R.S. Ellis, Large deviations for a general class of random vectors. Ann. Probab. **12**, 1–12 (1984)
121. R.S. Ellis, *Entropy, Large Deviations and Statistical Mechanics* (Springer, New York, 1985)
122. R.S. Ellis, Large deviations for the empirical measure of a Markov chain with an application to the multivariate empirical measure. Ann. Probab. **16**, 1496–1508 (1988)
123. R.S. Ellis, A.D. Wyner, Uniform large deviation property of the empirical process of a Markov chain. Ann. Probab. **17**, 1147–1151 (1989)
124. K.D. Elworthy, Stochastic dynamical systems and their flows, in *Stochastic Analysis* (Academic Press, New York, 1978), pp. 79–95
125. K.D. Elworthy, Stochastic flows on Riemannian manifolds, in *Diffusion Processes and Related Problems in Analysis*, vol. II (Springer, Berlin, 1992), pp. 37–72
126. S.N. Ethier, T.G. Kurtz, *Markov Processes: Characterization and Convergence* (Wiley, New York, 1986)
127. W.G. Faris, G. Jona-Lasinio, Large fluctuations for a nonlinear heat equation with noise. J. Phys. A **15**(10), 3025–3055 (1982)
128. H. Federer, *Geometric Measure Theory* (Springer, Berlin, 1996)

129. W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1 (Wiley, New York, 1968)
130. W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 2 (Wiley, New York, 1971)
131. J. Feng, Large deviation for diffusions and Hamilton-Jacobi equation in Hilbert spaces. Ann. Probab. **34**(1), 321–385 (2006)
132. J. Feng, T.G. Kurtz, *Large Deviations for Stochastic Processes*. Mathematical Surveys and Monographs, vol. 131 (American Mathematical Society, Providence, 2006)
133. W.H. Fleming, Exit probabilities and optimal stochastic control. Appl. Math. Optim. **4**, 329–346 (1978)
134. W.H. Fleming, H.M. Soner, Asymptotic expansions for Markov processes with Lévy generators. Appl. Math. Optim. **19**, 203–223 (1989)
135. W.H. Fleming, H.M. Soner, *Controlled Markov Processes and Viscosity Solutions* (Springer, New York, 1992)
136. W.H. Fleming, P.E. Souganidis, Asymptotic series and the method of vanishing viscosity. Indiana Univ. Math. J. **35**, 425–447 (1987)
137. R. Foley, D. McDonald, Join the shortest queue: stability and exact asymptotics. Ann. Appl. Probab. **11**, 569–607 (2001)
138. M.I. Freidlin, The averaging principle and theorems on large deviations. Russian Math. Surv. **33**, 117–176 (1978)
139. M.I. Freidlin, Random perturbations of reaction-diffusion equations: the quasideterministic approximation. Trans. Am. Math. Soc. **305**(2), 665–697 (1988)
140. M.I. Freidlin, A.D. Wentzell, *Random Perturbations of Dynamical Systems* (Springer, New York, 1984)
141. A.J. Ganesh, N. O'Connell, D.J. Wischik, *Big Queues* (Springer, Berlin, 2004)
142. H. Gao, C. Sun, Well-posedness and large deviations for the stochastic primitive equations in two space dimensions. Commun. Math. Sci. **10**(2), 575–593 (2012)
143. P. Gao, The stochastic Swift-Hohenberg equation. Nonlinearity **30**(9), 3516 (2017)
144. J. Gärtner, On large deviations from the invariant measure. Theory Probab. Appl. **22**, 24–39 (1977)
145. M.J.J. Garvels, The splitting method in rare event simulation. Ph.D. thesis, University of Twente, The Netherlands (2000)
146. M. Ghosh, Probabilities of moderate deviations under m-dependence. Can. J. Stat. **2**(1–2), 157–168 (1974)
147. D. Gilbarg, N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd edn. (Springer, Berlin, 1983)
148. P. Glasserman, P. Heidelberger, P. Shahabuddin, T. Zajic, A large deviations perspective on the efficiency of multilevel splitting. IEEE Trans. Automat. Control **43**, 1666–1679 (1998)
149. P. Glasserman, S. Kou, Analysis of an importance sampling estimator for tandem queues. ACM Trans. Model. Comput. Simul. **4**, 22–42 (1995)
150. P. Glasserman, Y. Wang, Counter examples in importance sampling for large deviations probabilities. Ann. Appl. Probab. **7**, 731–746 (1997)
151. U. Grenander, M.I. Miller, Representations of knowledge in complex systems. J. R. Stat. Soc. B **56**(3), 549–603 (1994)
152. Z. Haraszti, J.K. Townsend, The theory of direct probability redistribution and its application to rare event simulation, in *Proceedings of the IEEE International Conference on Communications* (1998), pp. 1443–1450
153. Z. Haraszti, J.K. Townsend, The theory of direct probability redistribution and its application to rare event simulation. ACM Trans. Model. Comput. Simul. **9**, 105–140 (1999)
154. L. Holst, On the coupon collectors and other urn problems. Int. Stat. Rev. **54**(1), 15–27 (1986)
155. R.A. Horn, C.R. Johnson, *Topics in Matrix Analysis* (Cambridge University Press, New York, 1991)
156. Y. Hu, D. Nualart, T. Zhang, Large deviations for stochastic heat equation with rough dependence in space. Bernoulli **24**(1), 354–385 (2018)

157. I.A. Ibragimov, Conditions for smoothness of trajectories of random functions. Teor. Veroyatnost. i Primenen. **28**(2), 229–250 (1983)
158. I. Ignatiouk-Robert, Large deviations for processes with discontinuous statistics. Ann. Probab. **33**, 1479–1508 (2005)
159. N. Ikeda, S. Watanabe. *Stochastic Differential Equations and Diffusion Processes*. North-Holland Mathematical Library, vol. 24 (North-Holland Publishing Co., Amsterdam; 2nd edn., Kodansha, Ltd., Tokyo, 1989)
160. V.M. Imaĭkin, A.I. Komech, Large deviations of solutions of nonlinear stochastic equations. Trudy Sem. Petrovsk. 258(13), 177–196 (1988)
161. J. Jacod, A.N. Shiryaev, *Limit Theorems for Stochastic Processes* (Springer, Berlin, 1987)
162. N. Jain, Large deviation lower bounds for additive functionals of Markov processes. Ann. Probab. **18**, 1071–1098 (1990)
163. Y. Le Jan, Flots de diffusions dans $\mathbb{R}^d$. C.R. Acad. Sci., Paris, Ser. I **294**, 687–689 (1982)
164. Y. Le Jan, S. Watanabe, Stochastic flows of diffeomorphisms, in *North-Holland Mathematical Library*, vol. 32 (Elsevier, Amsterdam, 1984), pp. 307–332
165. N.L. Johnson, S. Kotz, *Urn Models and Their Applications* (Wiley, New York, 1977)
166. H. Kahn, T.E. Harris, Estimation of particle transmission by random sampling. Natl. Bur. Stand. Appl. Math. Ser. **12**, 27–30 (1951)
167. O. Kallenberg, *Foundations of Modern Probability*. Probability and its Applications (New York), 2nd edn. (Springer, New York, 2002)
168. G. Kallianpur, *Stochastic Filtering Theory*, vol. 13 (Springer Science & Business Media, Berlin, 2013)
169. G. Kallianpur, J. Xiong, *Stochastic Differential Equations in Infinite-Dimensional Spaces*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, vol. 26 (Institute of Mathematical Statistics, Hayward, 1995)
170. G. Kallianpur, J. Xiong, Large deviations for a class of stochastic partial differential equations. Ann. Probab. **24**(1), 320–345 (1996)
171. H.-W. Kang, T.G. Kurtz, Separation of time-scales and model reduction for stochastic reaction networks. Ann. Appl. Probab. **23**(2), 529–583 (2013)
172. I. Karatzas, S.E. Shreve, *Brownian Motion and Stochastic Calculus* (Springer, New York, 1988)
173. A. Klenke, *Probability Theory: A Comprehensive Course* (Springer Science & Business Media, Berlin, 2013)
174. P. Kotelenez, Existence, uniqueness and smoothness for a class of function valued stochastic partial differential equations. Stoch.: Int. J. Probab. Stoch. Process. **41**(3), 177–199 (1992)
175. P. Kotelenez, *Stochastic Ordinary and Stochastic Partial Differential Equations: Transition from Microscopic to Macroscopic Equations*, vol. 58 (Springer Science & Business Media, Berlin, 2007)
176. S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959)
177. S. Kullback, R.A. Leibler, On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
178. H. Kunita, *Stochastic Flows and Stochastic Differential Equations* (Cambridge University Press, Cambridge, 1990)
179. T.G. Kurtz, *Approximation of Population Processes*, CBMS-NSF Regional Conference, vol. 36 (Series in Applied Mathematics) (SIAM, Philadelphia, 1981)
180. T.G. Kurtz, P.E. Protter, Weak convergence of stochastic integrals and differential equations II: Infinite dimensional case, in *Probabilistic Models for Nonlinear Partial Differential Equations* (Springer, Berlin, 1996), pp. 197–285
181. H.J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations* (Academic, New York, 1977)
182. H.J. Kushner, G.G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Stochastic Modelling and Applied Probability (Springer, New York, 2003)
183. P. L'Ecuyer, J.H. Blanchet, B. Tuffin, P.W. Glynn, Asymptotic robustness of estimators in rare-event simulation. ACM Trans. Model. Comput. Simul. **20**(1), 1–41 (2010)

184. J. Lehec, Representation formula for the entropy and functional inequalities. Ann. Inst. H. Poincar Probab. Stat. **49**(3), 885–899 (2013)
185. C. Léonard, Large deviations for Poisson random measures and processes with independent increments. Stoch. Process. Appl. **85**(1), 93–121 (2000)
186. Y. Li, R. Wang, N. Yao, S. Zhang, A moderate deviation principle for stochastic Volterra equation. Stat. Probab. Lett. **122**, 79–85 (2017)
187. Y. Li, R. Wang, N. Yao, S. Zhang, Moderate deviations for a fractional stochastic heat equation with spatially correlated noise. Stoch. Dyn. **17**(04), 1750025 (2017)
188. Y. Li, S. Zhang, Moderate deviations and central limit theorem for positive diffusions. J. Inequalities Appl. **2016**(1), 87 (2016)
189. D. Lipshutz, Exit time asymptotics for small noise stochastic delay differential equations (2017). arXiv:1710.09771
190. J.S. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2004)
191. W. Liu, Large deviations for stochastic evolution equations with small multiplicative noise. Appl. Math. Optim. **61**(1), 27–56 (2010)
192. W. Liu, M. Röckner, X.-C. Zhu, Large deviation principles for the stochastic quasi-geostrophic equations. Stoch. Process. Appl. **123**(8), 3299–3327 (2013)
193. L. Ljung, T. Söderström, *Theory and Practice of Recursive Identification*. Series in Signal Processing, Optimization, and Control (MIT Press, Cambridge, 1985)
194. X. Ma, F. Xi, Moderate deviations for neutral stochastic differential delay equations with jumps. Stat. Probab. Lett. **126**, 97–107 (2017)
195. K. Majewski, Large deviations of the steady-state distribution of reflected processes with applications to queueing systems. Queueing Syst. **29**, 351–381 (1998)
196. U. Manna, S.S. Sritharan, P. Sundar, Large deviations for the stochastic shell model of turbulence. NoDEA Nonlinear Differ. Equ. Appl. **16**(4), 493–521 (2009)
197. M. Métivier, *Semimartingales: A Course on Stochastic Processes*, vol. 2 (Walter de Gruyter, Berlin, 1982)
198. M. Métivier, J. Pellaumail, *Stochastic Integration* (Academic, New York, 2014)
199. Paul-André Meyer, Les inégalités de Burkholder en théorie des martingales, d'après Gundy. Séminaire de probabilités de Strasbourg **3**, 163–174 (1969)
200. R. Michel, Results on probabilities of moderate deviations. Ann. Probab. **2**(2), 349–353 (1974)
201. M.I. Miller, G.E. Christensen, Y. Amit, U. Grenander, Mathematical textbook of deformable neuroanatomies. Proc. Natl. Acad. Sci. **90**(24), 11944–11948 (1993)
202. A. Millet, D. Nualart, M. Sanz-Solé, Large deviations for a class of anticipating stochastic differential equations. Ann. Probab. **20**(4), 1902–1931 (1992)
203. C. Mo, J. Luo, Large deviations for stochastic differential delay equations. Nonlinear Anal. **80**, 202–210 (2013)
204. P. Ney, E. Nummelin, Markov additive processes I: eigenvalue properties and limit theorems. Ann. Prob. **16**, 561–592 (1987)
205. P. Ney, E. Nummelin, Markov additive processes II: large deviations. Ann. Prob. **15**, 593–609 (1987)
206. D. Nualart, E. Pardoux, Markov field properties of solutions of white noise driven quasi-linear parabolic PDEs. Stoch.: Int. J. Probab. Stoch. Process. **48**(1–2), 17–44 (1994)
207. V. Ortiz-López, M. Sanz-Solé, A Laplace principle for a stochastic wave equation in spatial dimension three, in *Stochastic Analysis 2010* (Springer, Berlin, 2011), pp. 31–49
208. V. Demers, P. L'Ecuyer, B. Tuffin, Rare events, splitting, and quasi-Monte Carlo. ACM Trans. Model. Comput. Simul. **17**(2) (2007)
209. S. Peszat, Large deviation principle for stochastic evolution equations. Probab. Theory Related Fields **98**(1), 113–136 (1994)
210. P.E. Protter, *Stochastic Integration and Differential Equations*. Stochastic Modelling and Applied Probability, vol. 21 (Springer, Berlin, 2005). 2nd edn. Version 2.1, Corrected Third Printing
211. F. Rassoul-Agha, T. Seppäläinen, *A Course on Large Deviations with an Introduction to Gibbs Measures*, vol. 162 (American Mathematical Society, Providence, 2015)

212. P.A. Razafimandimby, Viscosity limit and deviations principles for a grade-two fluid driven by multiplicative noise. Annali di Matematica Pura ed Applicata **197**(5), 1547–1583 (2018)
213. J. Ren, S. Xu, X. Zhang, Large deviations for multivalued stochastic differential equations. J. Theoret. Probab. **23**(4), 1142–1156 (2010)
214. J. Ren, X. Zhang, Freidlin-Wentzell's large deviations for homeomorphism flows of non-Lipschitz SDEs. Bull. Sci. Math. **129**(8), 643–655 (2005)
215. J. Ren, X. Zhang, Schilder theorem for the Brownian motion on the diffeomorphism group of the circle. J. Funct. Anal. **224**(1), 107–133 (2005)
216. J. Ren, X. Zhang, Freidlin-Wentzell's large deviations for stochastic evolution equations. J. Funct. Anal. **254**(12), 3148–3172 (2008)
217. R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970)
218. M. Röckner, T. Zhang, X. Zhang, Large deviations for stochastic tamed 3D Navier-Stokes equations. Appl. Math. Optim. **61**(2), 267–285 (2010)
219. S. Ross, *Introduction to Probability Models*, 8th edn. (Academic, New York, 2002)
220. H.L. Royden, P. Fitzpatrick, *Real Analysis*, vol. 198 (Macmillan, New York, 1988)
221. B.L. Rozovskii, *Stochastic Evolution Systems: Linear Theory and Applications to Non-Linear Filtering*, vol. 35 (Springer Science & Business Media, Berlin, 2012)
222. H. Rubin, J. Sethuraman, Probabilities of moderate deviations. Sankhyā Ser. A **27**, 325–346 (1965)
223. G. Rubino, B. Tuffin (eds.), *Rare Event Simulation Using Monte Carlo Methods* (Wiley, New York, 2009)
224. R.Y. Rubinstein, D.P. Kroese, *Simulation and the Monte Carlo Method*, 3rd edn. (Wiley, New York, 2016)
225. R.Y. Rubinstein, A. Ridder, R. Vaisman, *Fast Sequential Monte Carlo Methods for Counting and Optimization*, 1st edn. (Wiley Publishing, New York, 2013)
226. W. Rudin, *Functional Analysis* (McGraw-Hill, New York, 1991)
227. M. Salins, A. Budhiraja, P. Dupuis, Uniform large deviation principles for Banach space valued stochastic differential equations (2017), arXiv:1803.00648 (To appear in Trans. Am. Math. Soc.)
228. I.N. Sanov, On the probability of large deviations of random variables. Mat. Sbornik **42**(84), 11–44 (1957)
229. M. Schilder, Some asymptotic formulas for Wiener integrals. Trans. Am. Math. Soc. **125**(1), 63–85 (1966)
230. Z. Schuss, *Theory and Applications of Stochastic Differential Equations* (Wiley, New York, 1988)
231. A. Shwartz, A. Weiss, *Large Deviations for Performance Analysis: Queues, Communication and Computing* (Chapman and Hall, New York, 1995)
232. D. Siegmund, Importance sampling in the Monte Carlo study of sequential tests. Ann. Stat. **4**, 673–684 (1976)
233. M. Sion, On general minimax theorems. Pac. J. Math. **8**, 171–176 (1958)
234. H. Soner, Optimal control with state-space constraint I. SIAM J. Control. Optim. **24**, 05 (1986)
235. R.B. Sowers, Large deviations for a reaction-diffusion equation with non-Gaussian perturbations. Ann. Probab. **20**(1), 504–537 (1992)
236. S.S. Sritharan, P. Sundar, Large deviations for the two-dimensional Navier-Stokes equations with multiplicative noise. Stoch. Process. Appl. **116**(11), 1636–1659 (2006)
237. D.W. Stroock, *An Introduction to the Theory of Large Deviations* (Springer, New York, 1984)
238. S.R.S. Varadhan, Asymptotic probabilities and differential equations. Comm. Pure Appl. Math. **19**(3), 261–286 (1966)
239. S.R.S. Varadhan, *Large Deviations and Applications*. CBMS-NSF Regional Conference Series in Mathematics. (SIAM, Philadelphia, 1984)
240. C. Villani, *Optimal Transport: Old and New* (Springer, Berlin, 2009)
241. M. Villen-Altamirano, J. Villen-Altamirano, RESTART: A method for accelerating rare event simulations, in *Proceedings of the 13th International Teletraffic Congress, Queueing, Performance and Control in ATM* (Elsevier, Amsterdam, 1991), pp. 71–76

242. M. Villen-Altamirano, J. Villen-Altamirano, RESTART: A straightforward method for fast simulation of rare events, in *Proceedings of the 1994 Winter Simulation Conference* (1994), pp. 282–289
243. J.B. Walsh, An introduction to stochastic partial differential equations, in *École d'été de probabilités de Saint-Flour, XIV—1984*. Lecture Notes in Mathematics, vol. 1180 (Springer, Berlin, 1986), pp. 265–439
244. W. Wang, J. Duan, Reductions and deviations for stochastic partial differential equations under fast dynamical boundary conditions. Stoch. Anal. Appl. **27**(3), 431–459 (2009)
245. A.D. Wentzell, Rough limit theorems on large deviations for Markov stochastic processes. I. Theory Probab. Appl. **21**, 227–242 (1976)
246. A.D. Wentzell, Rough limit theorems on large deviations for Markov stochastic processes. II. Theory Probab. Appl. **21**, 499–512 (1976)
247. A.D. Wentzell, Rough limit theorems on large deviations for Markov stochastic processes. III. Theory Probab. Appl. **24**, 675–692 (1979)
248. A.D. Wentzell, Rough limit theorems on large deviations for Markov stochastic processes. IV. Theory Probab. Appl. **27**, 215–234 (1982)
249. R.J. Williams, Recurrence classification and invariant measure for reflected Brownian motion in a wedge. Ann. Probab. **13**, 758–778 (1985)
250. M. Winter, L. Xu, J. Zhai, T. Zhang, The dynamics of the stochastic shadow Gierer-Meinhardt system. J. Differ. Equ. **260**(1), 84–114 (2016)
251. J. Wu, Uniform large deviations for multivalued stochastic differential equations with Poisson jumps. Kyoto J. Math. **51**(3), 535–559 (2011)
252. J. Xiong, Large deviations for diffusion processes in duals of nuclear spaces. Appl. Math. Optim. **34**(1), 1–27 (1996)
253. J. Xiong, J. Zhai, Large deviations for locally monotone stochastic partial differential equations driven by Lévy noise. Bernoulli **24**(4A), 2842–2874 (2018)
254. T. Xu, T. Zhang, White noise driven SPDEs with reflection: existence, uniqueness and large deviation principles. Stoch. Process. Appl. **119**(10), 3453–3470 (2009)
255. D. Yang, J. Duan, Large deviations for the stochastic quasigeostrophic equation with multiplicative noise. J. Math. Phys. **51**(5), 053301 (2010)
256. D. Yang, Z. Hou, Large deviations for the stochastic derivative Ginzburg-Landau equation with multiplicative noise. Phys. D **237**(1), 82–91 (2008)
257. J. Yang, J. Zhai, Asymptotics of stochastic 2d hydrodynamical type systems in unbounded domains. Infin. Dimens. Anal., Quantum Probab. Relat. Top. **20**(03), 1750017 (2017)
258. X. Yang, J. Zhai, T. Zhang, Large deviations for SPDEs of jump type. Stoch. Dyn. **15**(04), 1550026 (2015)
259. G. Yin, C. Zhu, *Hybrid Switching Siffusions: Properties and Applications*, vol. 63 (Springer, New York, 2010)
260. L. Ying, R. Srikant, A. Eryilmaz, G.E. Dullerud, A large deviation analysis of scheduling in wireless networks. IEEE Trans. Inf. Theory **52**(11), 5088–5098 (2006)
261. J. Zabczyk, On large deviations for stochastic evolution equations, in *Stochastic Systems and Optimization (Warsaw, 1988)*. Lecture Notes in Control and Information Sciences, vol. 136 (Springer, Berlin, 1989), pp. 240–253
262. J. Zhai, T. Zhang, Large deviations for 2-d stochastic Navier-Stokes equations driven by multiplicative Lévy noises. Bernoulli **21**(4), 2351–2392 (2015)
263. J. Zhai, T. Zhang, Large deviations for stochastic models of two-dimensional second grade fluids. Appl. Math. Optim. **75**(3), 471–498 (2017)
264. J. Zhai, T. Zhang, W. Zheng, Large deviations for stochastic models of two-dimensional second grade fluids driven by Lévy noise (2017). arXiv preprint arXiv:1706.08862
265. J. Zhai, T. Zhang, W. Zheng, Moderate deviations for stochastic models of two-dimensional second grade fluids. Stoch. Dyn. **18**(03), 1850026 (2018)
266. J. Zhang, P. Dupuis, Large deviation principle for general occupancy models. Comb., Probab., Comput. **17**, 437–470 (2008)

267. R. Zhang, G. Zhou, Large deviations for nematic liquid crystals driven by pure jump noise. Math. Methods Appl. Sci. **41**(14), 5552–5581 (2018)
268. X. Zhang, Euler schemes and large deviations for stochastic Volterra equations with singular kernels. J. Differ. Equ. **244**(9), 2226–2250 (2008)
269. X. Zhang, Clark-Ocone formula and variational representation for Poisson functionals. Ann. Probab. **37**(2), 506–529 (2009)
270. X. Zhang, Stochastic Volterra equations in Banach spaces and stochastic partial differential equation. J. Funct. Anal. **258**(4), 1361–1425 (2010)
271. X. Zhang, Well-posedness and large deviation for degenerate SDEs with Sobolev coefficients. Rev. Mat. Iberoam. **29**(1), 25–52 (2013)
272. H. Zhao, S. Xu, Freidlin-Wentzells large deviations for stochastic evolution equations with Poisson jumps. Adv. Pure Math. **6**(10), 676 (2016)
273. W. Zheng, J. Zhai, T. Zhang, Moderate deviations for stochastic models of two-dimensional second grade fluids driven by Lévy noise (2018). arXiv preprint arXiv:1801.08429

Index

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0