A weak convergence approach to large deviations for stochastic approximations

Henrik Hult, Adam Lindhe, Pierre Nyquist, Guo-Jhen Wu

September 21, 2023

Abstract

Large deviations for stochastic approximations is a well-studied field that yields convergence properties for many useful algorithms in statistics, machine learning and statistical physics. In this article, we prove, under certain assumptions, a large deviation principle for a stochastic approximation with state-dependent Markovian noise and with decreasing step size. Common algorithms that satisfy these conditions include stochastic gradient descent, persistent contrastive divergence and the Wang-Landau algorithm. The proof is based on the weak convergence approach to the theory of large deviations and uses a representation formula to rewrite the problem into a stochastic control problem. The resulting rate function is an action potential over a local rate function that is the Fenchel-Legendre transform of a limiting Hamiltonian.

1 Introduction

Stochastic approximations with state-dependent noise provide a rich and useful family of stochastic recursive algorithms. It includes many popular learning algorithms in statistics, machine learning, and statistical physics. Examples include stochastic gradient descent, persistent contrastive divergence, reinforcement learning, adaptive Markov Chain Monte-Carlo and extended ensemble algorithms. The theory of stochastic approximations originates from the work of Robbins and Monro in the 1950s, see [18] and Kiefer-Wolfowitz[15], and has been thoroughly developed ever since. Monographs covering many of the fundamental results of the theory include [2, 7, 16].

The basic stochastic approximation algorithm with state-dependent noise considers a stochastic process $\{X_k\}_{k\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{F}, P)$, with an associated noise sequence $\{Y_k\}_{k\in\mathbb{N}}$. The process $\{X_k\}_{k\in\mathbb{N}}$ is assumed to satisfy the recursion,

$$X_{k+1} = X_k + \varepsilon_k g(X_k, Y_{k+1}), \qquad k \geq 0,$$

where $X_0 = x_0 \in \mathbb{R}^{d_1}$, $Y_0 = y_0 \in \mathbb{R}^{d_2}$, $g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$, and $\{\varepsilon_k\}$ is a sequence of step-sizes. The noise sequence $\{Y_k\}_{k\in\mathbb{N}}$ is state-dependent in such a way that

$$P(Y_{k+1} \in A | X_k, Y_k) = \rho_{X_k}(Y_k, A), \qquad A \in \mathcal{B}(\mathbb{R}^{d_2}),$$

with $\rho_x(y, \cdot)$ being a probability measure on the Borel sets of $\mathbb{R}^{d_2}$, for any $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$.

In this paper, we study the behaviour of the algorithm close to a point of convergence. The ODE method is a well-established method for studying convergence of stochastic approximations, which states that, for large k, the stochastic approximation essentially follows a limit ODE. The ODE method can be briefly explained as follows. By assuming that, for each $x \in \mathbb{R}^{d_1}$, the transition kernel admits a unique invariant distribution π_x we may rewrite the recursion as

$$\frac{X_{k+1} - X_k}{\varepsilon_k} = \bar{g}(X_k) + [g(X_k, Y_{k+1}) - \bar{g}(X_k)],$$

where $\bar{g}(x) = \int g(x, y)\pi_x(dy)$. Under appropriate conditions for small ε_k and large k, the effect of the noise $g(X_k, Y_{k+1}) - \bar{g}(X_k)$ will be small and the algorithm will approximately follow the solution to the limit ODE

$$\dot{x}(t) = \bar{g}(x(t)).$$

Consequently, points of convergence for the stochastic approximation may be described as the forward limit set of the limit ODE. Due to the inherent randomness of the stochastic approximation algorithm, it may, with a small probability, deviate from a neighbourhood of a point of convergence. Large deviations theory provides insights into the rate at which the algorithm deviates from such a neighbourhood and characterizes the most likely trajectories along which such deviations occur.

In popular language, we may say that as the stochastic approximation algorithm approaches a point of convergence it is learning, while as it starts to deviate from a point of convergence it is forgetting. The large deviations principle characterizes the rate at which the algorithm forgets and how the forgetting occurs.

A simple and useful method to exclude the possibility of divergence of a stochastic approximation algorithm, is to project the updates on a compact set C, by considering the projected recursion

$$X_{k+1} = \mathrm{proj}_C \left[X_k + \varepsilon_k g(X_k, Y_{k+1})\right], \qquad k \geq 0,$$

with proj_C denoting the projection onto C. However, in this paper we primarily study the behaviour of the algorithm close to a point of convergence and will therefore not be concerned with projected algorithms.

The existing literature on large deviations for stochastic approximations studies, on the one hand, the setting with fixed step size, where $\varepsilon_k = \varepsilon > 0$, does not depend on k, and on the other hand the setting of decreasing step size, where $\varepsilon_k \to 0$ as $k \to \infty$. For fixed step size the theory was first developed by Freidlin, see [12, 13], for dynamical systems in continuous time with noise that does not depend on the state. The results were generalized by Iscoe, Ney and Nummelin, see [14], who consider Markov-additive processes in continuous and discrete time. The most general results are obtained by Dupuis, see [8], for discrete time systems by providing results for state-dependent noise. The results rely on the existence of an appropriate limiting Hamiltonian and the rate function is given by an action functional where the local rate function is the Fenchel-Legendre transform of the limiting Hamiltonian. See Section 4.1 for additional details on the development of large deviations principles for stochastic approximations with constant step size.

For stochastic approximations with decreasing step size the first results are obtained by Kushner [17] who considers step size sequences of the form $\varepsilon_n = (n+1)^{-\rho}$, $\rho \in (0, 1]$, and update functions $g(x, y) = b(x) + y$, with $b(\cdot)$ Lipschitz continuous and $\{Y_n\}$ a sequence of iid centered Gaussian variables. Kushner assumes the existence of an appropriate limiting time-dependent Hamiltonian and identifies the appropriate scaling sequence. A generalization is provided by Dupuis and Kushner [11] who consider step size sequences where $\varepsilon_n \geq 0$, $\sum_n \varepsilon_n = \infty$, $\varepsilon_n \to 0$, update functions of the form $g(x, y) = \bar{b}(x) + b(x, y)$, and $E[b(x, Y_n)] = 0$. They assume further that the noise sequence satisfies $Y_n = (\tilde{Y}_n, \hat{Y}_n)$, where $\{\tilde{Y}_n\}$ and $\{\hat{Y}_n\}$ are mutually independent, $\{\tilde{Y}_n\}$ is stationary and bounded and $\{\hat{Y}_n\}$ is stationary centered Gaussian process with summable correlation function. Moreover, $b(x, Y_n) = b_1(x, \tilde{Y}_n) + b_0(x)\hat{Y}_n$, where $b_1(\cdot, \tilde{y})$, b_0 and $\bar{b}$ are uniformly (in $\tilde{y}, x$) Lipschitz and bounded.

In the existing literature, the large deviations principle is obtained by identifying a Hamiltonian $H(x, \alpha)$, that sometimes can be interpreted as a limiting log-moment generating function and defining the local rate function $L(x, \beta)$ as the convex conjugate of $H(x, \alpha)$. A problem with this approach is that the Hamiltonian is implicitly defined as a limit and its relation to the underlying dynamics such as the transition kernel $\rho_x(y, dz)$ can only be established in some special cases. In this paper, the results of Dupuis and Kushner, see [11], for a decreasing step size sequence, are generalized to include state-dependent noise and the local rate function is expressed in terms of the family of transition kernels. In addition, the conditions are somewhat more general as the update function g need not be bounded in x. We also remark that from a technical point of view, the setting of fixed step size, is somewhat easier and analogous results can be obtained with minor modification of the weak-convergence techniques used in this paper. However, due to space considerations we do not pursue the results for fixed step size in this paper.

2 Stochastic approximation

In this section, the stochastic approximation algorithm will be introduced and the notation and assumptions are stated. Some preliminaries on Laplace principles and a heuristic derivation of the rate function is also be provided.

2.1 Notation

The following notation will be used. Let $\mathbb{N} = \{1, 2, \dots\}$, $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ and $\{\varepsilon_k\}_{k\in\mathbb{N}_0} \doteq \{1/k\}$ denote the sequence of step-sizes (learning rate) of our stochastic approximation algorithm. Define the intermediate times $t_0 = 0$, $t_n = \sum_{k=1}^n \varepsilon_k$, and let $\mathbf{m}(t) = \max\{n : t_n \leq t\}$ be the maximum number of iterations that occurs before time t. Note that $\mathbf{m}(t_n) = n$. The space $C([0, T] : \mathbb{R}^d)$ consists of $\mathbb{R}^d$-valued continuous functions defined on $[0, T]$ and $C_x([0, T] : \mathbb{R}^d)$ is the subspace of continuous functions starting at x at time 0. The space $C([0, T] : \mathbb{R}^d)$ is equipped with the sup norm $\|f\|_\infty = \sup_{s,t\in[0,T]} \|f(s) - f(t)\|$ for $f \in C([0, T] : \mathbb{R}^d)$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$. For $x, y \in \mathbb{R}^d$, their inner product is denoted $\langle x, y\rangle$. Given a Polish space $\mathcal{X}$, with Borel σ-algebra $\mathcal{B}(\mathcal{X})$, the space of probability measures on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$, equipped with the topology of weak convergence. For $\theta \in \mathcal{P}(\mathcal{X})$, the relative entropy $R(\cdot\|\theta)$ is a map from $\mathcal{P}(\mathcal{X})$ into the extended real numbers, defined by

$$R(\gamma\|\theta) \doteq \begin{cases} \int_{\mathcal{X}} \left(\log \frac{d\gamma}{d\theta}\right) d\gamma, & \gamma \ll \theta, \\ +\infty, & \text{otherwise} \end{cases}$$

We refer to $R(\gamma\|\theta)$ as the relative entropy of γ with respect to θ. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, $\sigma(dy|x)$ a stochastic kernel on $\mathcal{Y}$ given $\mathcal{X}$, and $\theta \in \mathcal{P}(\mathcal{X})$. Then $\theta \otimes \sigma$ is defined to be the unique probability measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$ with the property that for $A \in \mathcal{B}(\mathcal{X})$ and $B \in \mathcal{B}(\mathcal{Y})$,

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

The formula is summarized by the notation $\theta \otimes \sigma(dx \times dy) = \theta(dx) \otimes \sigma(dy|x)$. Given a transition kernel $p(x, dy)$ on $\mathcal{X}$ and $k \in \mathbb{N}$, let $p^{(1)}(x, dy) = p(x, dy)$ and, for $k \geq 1$, $p^{(k)}(x, dy)$ denote the k-step transition probability function defined recursively by

$$p^{(k+1)}(x, A) = \int_{\mathcal{X}} p(y, A)p^{(k)}(x, dy), \quad A \in \mathcal{B}(\mathcal{X}).$$

Given $\mu \in \mathcal{P}(\mathcal{X})$, let $\mathcal{A}(\mu) \doteq \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : [\gamma]_1 = [\gamma]_2 = \mu\}$, where $[\gamma]_1$ and $[\gamma]_2$ denote the first and second marginals of γ.

2.2 The model

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Consider a stochastic approximation algorithm $\{X_k\}_{n\in\mathbb{N}_0}$ of the Robbins-Munro type, with state-dependent noise sequence $\{Y_k\}_{k\in\mathbb{N}}$, starting from X_0 and satisfying the recursion,

$$X_{k+1} = X_k + \varepsilon_{k+1} g(X_k, Y_{k+1}), \quad k \geq 0,$$

where $g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$, and $\{Y_n\}_{n\in\mathbb{N}_0}$ starting from Y_0, and, for every $k \in \mathbb{N}_0$ and $A \in \mathcal{B}(\mathbb{R}^{d_2})$

$$P(Y_{k+1} \in A|X_k, Y_k) = \rho_{X_k}(Y_k, A)$$

with $\rho_x(y, \cdot) \in \mathcal{P}(\mathbb{R}^{d_2})$ for any $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$.

We are interested in analyzing the asymptotic behavior of the stochastic approximation $\{X_n\}_{n\in\mathbb{N}}$ for large values of n. Therefore, for each $n \in \mathbb{N}$ and $x_0 \in \mathbb{R}^{d_1}$, define a process $\{X^n_k\}_{k\geq n}$ that follows the same recursive iterations but starts from the n-th step recursion. To be more precise, let $X^n_n = x_0$ and for $k \geq n$

$$X^n_{k+1} = X^n_k + \varepsilon_{n+k+1} g(X^n_k, Y_{n+k+1}). \tag{2.1}$$

We consider a family of continuous interpolations of $\{X^n_k\}_{k\geq n}$: for each n, $X^n = \{X^n(t) : t \in [0, T]\}$ is given by $X^n(t_{n+k} - t_n) = X^n_{n+k}$ for $k = 0, 1, \ldots,$ and for intermediate time points t, $X^n(t)$ is defined by a piece-wise linear interpolation. Note that, for each n, $X^n \in C_{x_0}([0, T] : \mathbb{R}^{d_1})$.

2.3 Large deviations

In this section the definition of a Laplace principle is stated, which in the setting of a Polish space is equivalent to a large deviation principle (LDP) for a general class of random objects, including those considered in this paper. For general background on large deviation theory and the connection between the large deviation principle and the Laplace principle, see [4, Chapter 1]. Due to the equivalence of the large deviations principle and the Laplace principle, we will use the terminology of LDP and Laplace principle interchangeably throughout the paper.

A function $I : \mathcal{X} \to [0, \infty]$ is called a rate function on $\mathcal{X}$ if, for each $M < \infty$, the level set $\{x \in \mathcal{X} : I(x) \leq M\}$ is a compact subset of $\mathcal{X}$.

Definition 2.1 (Laplace principle) *Let I be a rate function on $\mathcal{X}$. The sequence $\{X^n\}$ is said to satisfy the Laplace principle on $\mathcal{X}$ with rate function I and scaling sequence $\{\beta_n\}$ if $\beta_n \to \infty$ as $n \to \infty$, and for all bounded continuous functions $F : \mathcal{X} \to \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{\beta_n} \log E e^{-\beta_n F(X^n)} = - \inf_{x\in\mathcal{X}} [F(x) + I(x)].$$

The term Laplace principle upper bound refers to the validity of

$$\limsup_{n\to\infty} \frac{1}{\beta_n} \log E e^{-\beta_n F(X^n)} \leq - \inf_{x\in\mathcal{X}} [F(x) + I(x)],$$

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

$$\liminf_{n\to\infty} \frac{1}{\beta_n} \log E e^{-\beta_n F(X^n)} \geq - \inf_{x\in\mathcal{X}} [F(x) + I(x)],$$

for all bounded continuous functions F.

Henceforth, when there is no ambiguity, we will refer to these only as upper and lower bounds, dropping the term "Laplace principle".

2.4 Assumptions

Recall that $\{X^n\} \subset C_{x_0}([0, T] : \mathbb{R}^{d_1})$ is the family of continuous interpolations of the stochastic approximation $\{X^n_k\}_{k\geq n}$. We end this section by listing the assumptions we use in this paper to establish an LDP for $\{X^n\}$.

Assumption 2.2

(A.1) *The function g is a measurable function, and for any $z \in \mathbb{R}^{d_2}$, $x \mapsto g(x, z)$ is Lipschitz continuous.*

(A.2) *The transition kernel $\rho_x(y, dz)$ is of the form $\rho_x(y, dz) = \eta_x(y, z)\lambda(dz)$, for some reference measure $\lambda \in \mathcal{P}(\mathbb{R}^{d_2})$. Moreover, $x \mapsto \eta_x(y, z)$ is uniformly continuous, in (z, y), and for any x, $(y, z) \mapsto \eta_x(y, z)$ is continuous.*

(A.3) *The function*

$$\Lambda(x,\alpha,y) = \log \int \exp\{\langle \alpha, g(x,z)\rangle\}\rho_x(y,dz),$$

is continuous in (x,α)*, uniformly in* y*.*

(A.4) *For every compact set* K*, there is a constant* $C(K)$*, such that for all* $y,z \in \mathbb{R}^{d_2}$

$$\sup_{x,w\in K} \frac{\eta_x(y,z)}{\eta_w(y,z)} < C(K).$$

(A.5) *For any* $x \in \mathbb{R}^{d_1}$*,* $\rho_x(y,dz)$ *satisfies the Feller property: for any* $\{y_n\}_{n\in\mathbb{N}} \subset \mathbb{R}^{d_2}$ *and* $y \in \mathbb{R}^{d_2}$ *such that* $y_n \to y$*,* $\rho_x(y_n,dz)$ *converges weakly to* $\rho_x(y,dz)$*.*

(A.6) *For any* $x \in \mathbb{R}^{d_1}$*, there exist positive integers* l_0 *and* n_0 *such that for all* y *and* w*,*

$$\sum_{i=l_0}^{\infty} \frac{1}{2^i}\rho_x^{(i)}(y,dz) \ll \sum_{j=n_0}^{\infty} \frac{1}{2^j}\rho_x^{(j)}(w,dz),$$

where $\rho_x^{(i)}$ *denotes the i-step transition probability.*

(A.7) *For every* $\alpha \in \mathbb{R}^{d_2}$*,*

$$\sup_{x\in\mathbb{R}^{d_1}} \sup_{y\in\mathbb{R}^{d_2}} \left(\log \int_{\mathbb{R}^{d_2}} e^{\langle \alpha, g(x,z)\rangle}\rho_x(y,dz)\right) < \infty.$$

$$\sup_{x\in\mathbb{R}^{d_1}} \sup_{y\in\mathbb{R}^{d_2}} \left(\log \int_{\mathbb{R}^{d_2}} e^{\langle \alpha, z\rangle}\rho_x(y,dz)\right) < \infty.$$

(A.8) *The sequence* $\{\varepsilon_k\}_{k\in\mathbb{N}}$ *satisfies* $\varepsilon_k > 0$ *for each* $k \geq 1$*,* $\lim_{k\to\infty}\varepsilon_k = 0$ *and* $\sum_k \varepsilon_k = \infty$*. Let* $\{\beta_n\} \doteq \{\mathbf{m}(t_n+T) - n\}$ *and suppose that the function* $h^n : [0,T] \to (0,\infty)$*, given by,*

$$h^n(t) = \beta_n \varepsilon_{n+i-1}, \qquad \textit{for } t \in [t_{n+i-1} - t_n, t_{n+i} - t_n], \quad i \in \{1,\dots,\beta_n\},$$

converges uniformly on $[0,T]$ *to some limit* h*.*

Assumption (A.1) is a standard assumption for the existence and uniqueness of a classical solution to an ordinary differential equation; Assumptions (A.2) and (A.5) guarantee the existence of an invariant probability measure for $\rho_x(y,dz)$; Assumption (A.6) ensures that the invariant probability measure is unique and the Markov chain with transition probability $\rho_x(y,dz)$ is ergodic. For each $x \in \mathbb{R}^{d_1}$, we let π_x denote this unique invariant measure for ρ_x. Assumption (A.7) is used to guarantee that the updates have finite exponential moments. Lastly, (A.8) is needed to prove convergence of the stochastic approximation algorithm and the limit function h may be interpreted as an asymptotic time-scale of the process $\{X^n\}$. For example, with $\varepsilon_k = 1/k$, a straightforward calculation shows that the limit function is given by $h(t) = e^{-t}(e^T - 1)$.

2.4.1 Form of the rate function

Before stating the Laplace principle for the sequence $\{X^n\}$, a heuristic calculation that suggests the correct form of the rate function is provided. The heuristic calculation contains several non-rigorous approximations and is only intended to give the reader a first suggestion of the form of the rate function. For simplicity we only consider a specific step-size sequence given by $\varepsilon_k = 1/k$, $k \geq 1$.

Recall first that the empirical measure of an ergodic Markov chain with transition probability $\rho(y, dz)$ satisfies a LDP with scaling sequence $\{n\}$ and rate function given by

$$J_0(\mu) = \inf_{\gamma \in \mathcal{A}(\mu)} R(\gamma || \mu \otimes \rho), \tag{2.2}$$

where $\mathcal{A}(\mu)$ is defined in Section 2.1, see, e.g., [9, Ch. 8]. Taking a bounded continuous function g on $\mathbb{R}^{d_2}$ the contraction principle, applied to the map $\mu \mapsto \int g(y)\mu(dy)$, implies that the sample average $\frac{1}{n}\sum_{i=1}^{n} g(Y_i)$ satisfies a LDP with rate function,

$$L_0(\beta) = \inf\{J_0(\mu) : \int g(y)\mu(dy) = \beta\} = \inf_{\mu}\left\{ \inf_{\gamma \in \mathcal{A}(\mu)} R(\gamma || \mu \otimes \rho) : \int g(y)\mu(dy) = \beta \right\}.$$

By incorporating a time variable the continuous linear interpolation of $\frac{1}{n}\sum_{i=1}^{[nt]} g(Y_i)$ satisfies a LDP on $C_0([0,T] : \mathbb{R}^d)$ with rate function

$$J_1(\varphi) = \int_0^T L_0(\dot{\varphi}(t))dt.$$

Consider now the stochastic approximation with fixed step-sizes where $\varepsilon_k = 1/n$ for all $k = 1, \ldots, n$. Take φ in $C([0,T] : \mathbb{R}^d)$ and consider the probability that the trajectory of X^n resides in a ball of radius $\sigma > 0$ around φ. In this case X^n may be approximated over a small interval $[s, s+\delta]$ of length $\delta > 0$ by,

$$X^n(s+\delta) - X^n(s) \approx \frac{1}{n}\sum_{i=\lfloor ns \rfloor}^{\lfloor n(s+\delta) \rfloor} g(X^n(s), Y_i) \approx \frac{1}{n}\sum_{i=\lfloor ns \rfloor}^{\lfloor n(s+\delta) \rfloor} g(\varphi(s), Y_i),$$

where Y_i is a Markov chain with transition probability $\rho_{\varphi(s)}(y, dz)$. Applying the Laplace principle for the sample average the increment $X^n(s+\delta) - X^n(s)$ satisfies a Laplace principle with rate function

$$\delta L(\varphi(s), \beta).$$

where

$$L(x, \beta) \doteq \inf_{\mu}\left\{ \inf_{\gamma \in \mathcal{A}(\mu)} R(\gamma || \mu \otimes \rho_x) : \beta = \int g(x,z)\mu(dz) \right\}. \tag{2.3}$$

By pasting together the local approximations over small intervals X^n satisfies a Laplace principle on $C([0,T]; \mathbb{R}^d)$ with rate function

$$J_2(\varphi) = \int_0^T L\left(\varphi(s), \dot{\varphi}(s)\right) ds.$$

Indeed,

$$
\begin{aligned}
-\frac{1}{n}\log P\{X^n(\cdot)\approx\varphi(\cdot)|\} &\approx -\frac{1}{n}\log P\left\{X^n(j\delta)\approx\varphi(j\delta)\text{ for all }1\le j\le\lfloor\frac{T}{\delta}\rfloor\right\}\\
&\approx -\frac{1}{n}\log P\left\{X^n((j+1)\delta)-X^n(j\delta)\approx\delta\dot{\varphi}(j\delta)\text{ for all }0\le j\le\frac{T}{\delta}-1\right\}\\
&\approx -\frac{1}{n}\log\prod_{j=0}^{\lfloor\frac{T}{\delta}\rfloor-1}P\left\{X^n((j+1)\delta)-X^n(j\delta)\in\delta\dot{\varphi}(j\delta)\mid X^n(j\delta)\right\}\\
&\approx -\frac{1}{n}\log\prod_{j=0}^{\lfloor\frac{T}{\delta}\rfloor-1}\exp\left\{-n\delta L\left(\varphi(j\delta),\dot{\varphi}(j\delta)\right)\right\}\\
&\approx \sum_{j=0}^{\lfloor\frac{T}{\delta}\rfloor-1}\delta L(\varphi(j\delta),\dot{\varphi}(j\delta))\\
&\approx \int_0^T L\left(\varphi(s),\dot{\varphi}(s)\right)ds.
\end{aligned}
$$

Consider now $\{X^n\}$ with decreasing step-size $\{\varepsilon_n\}$ as defined in Section 2.2. Take as the scaling sequence in the LDP the sequence $\{\beta_n\}=\{\mathbf{m}(t_n+T)-n\}$ and define $h^n(t)$ as in ((A.8)).

In this case the decreasing step-sizes corresponds to a change of time scale and the rate of change of X^n over a small interval $[s,s+\delta]$ of length $\delta>0$ may be approximated by

$$
\begin{aligned}
\frac{X^n(s+\delta)-X^n(s)}{\delta} &\approx \frac{1}{\delta}\sum_{i=\mathbf{m}(t_n+s)-n+1}^{\mathbf{m}(t_n+s+\delta)-n}\varepsilon_{n+i-1}g(\varphi(s),Y_i)\\
&\approx \frac{1}{\delta}\left(\frac{\sum_{i=\mathbf{m}(t_n+s)-n+1}^{\mathbf{m}(t_n+s+\delta)-n}\varepsilon_{n+i-1}}{\mathbf{m}(t_n+s+\delta)-\mathbf{m}(t_n+s)}\right)\left(\sum_{i=\mathbf{m}(t_n+s)-n+1}^{\mathbf{m}(t_n+s+\delta)-n}g(\varphi(s),Y_i)\right)\\
&\approx \frac{1}{\mathbf{m}(t_n+s+\delta)-\mathbf{m}(t_n+s)}\left(\sum_{i=\mathbf{m}(t_n+s)-n+1}^{\mathbf{m}(t_n+s+\delta)-n}g(\varphi(s),Y_i)\right),
\end{aligned}
$$

for which a LDP holds as in (2.3). Using a similar argument as in the case of fixed step sizes it follows that

$$\begin{aligned}
-\frac{1}{\beta_n}\log P\{X^n(\cdot)\approx\varphi(\cdot)\} &\approx -\frac{1}{\beta_n}\log P\left\{X^n(j\delta)\approx\varphi(j\delta)\text{ for all }1\le j\le\lfloor\frac{T}{\delta}\rfloor\right\}\\
&\approx -\frac{1}{\beta_n}\log P\left\{X^n((j+1)\delta)-X^n(j\delta)\approx\delta\dot\varphi(j\delta)\text{ for all }0\le j\le\frac{T}{\delta}-1\right\}\\
&\approx -\frac{1}{\beta_n}\log\prod_{j=0}^{\lfloor\frac{T}{\delta}\rfloor-1}P\left\{\frac{X^n((j+1)\delta)-X^n(j\delta)}{\delta}\approx\dot\varphi(j\delta)\mid X^n(j\delta)\right\}\\
&\approx \frac{1}{\beta_n}\sum_{j=0}^{\lfloor\frac{T}{\delta}\rfloor-1}(\mathbf{m}(t_n+(j+1)\delta)-\mathbf{m}(t_n+j\delta))L\left(\varphi(j\delta),\dot\varphi(j\delta)\right)\\
&\approx \frac{1}{\beta_n}\sum_{j=0}^{\lfloor\frac{T}{\delta}\rfloor-1}\left(\sum_{i=\mathbf{m}(t_n+j\delta)-n+1}^{\mathbf{m}(t_n+(j+1)\delta)-n}L(\varphi(\tau_i^n),\dot\varphi(\tau_i^n))\right)\\
&\approx \frac{1}{\beta_n}\sum_{i=1}^{\beta_n}\frac{1}{\varepsilon_{n+i-1}}L(\varphi(\tau_i^n),\dot\varphi(\tau_i^n))\varepsilon_{n+i-1}\\
&\approx \sum_{i=1}^{\beta_n}\frac{1}{h^n(\tau_i^n)}L(\varphi(\tau_i^n),\dot\varphi(\tau_i^n))\varepsilon_{n+i-1}\\
&\approx \int_0^T\frac{1}{h(s)}L\left(\varphi(s),\dot\varphi(s)\right)ds.
\end{aligned}$$

The above calculation indicates that the appropriate rate function in the LDP for X^n, the piecewise linear interpolation of the stochastic approximation, is given by

$$I(\varphi)=\int_0^T\frac{1}{h(t)}L(\varphi(t),\dot\varphi(t))dt,$$

where $h(t)=(e^T-1)e^{-t}$ is the limit of $h^n(t)$.

3 Statement of Main Results

The goal of this paper is to establish the LDP for the sequence of $X^n=\{X^n(t):t\in[0,T]\}$, the linear interpolations of $\{X_k^n\}_{k\ge n}$ starting from $X_n^n=x_0\in\mathbb{R}^{d_1}$. To this end, we define the function $I:C_{x_0}([0,T]:\mathbb{R}^{d_x})$ as,

$$I(\varphi)=\begin{cases}\int_0^T\frac{1}{h(t)}L(\varphi(t),\dot\varphi(t))dt, & \varphi\in AC_{x_0}([0,T]:\mathbb{R}^{d_1}),\\ +\infty, & \text{otherwise,}\end{cases}\tag{3.1}$$

with the local rate function L as in (2.3). Note that we suppress the dependence on the choice of starting point x_0 in the notation. The following Laplace princple is the main result of the paper, where I plays the role of the large deviation rate function.

Theorem 3.1 (Laplace principle) *Let $X^n = \{X^n(t) : t \in [0,T]\}$ be the continuous interpolations of $\{X_k^n\}_{k\geq n}$ given by (2.1) and take L as in (2.3). Under Assumptions (A.1)-(A.8), I is a rate function, and $\{X^n\}_{n\in\mathbb{N}}$ satisfies a Laplace principle with scaling sequence $\beta_n = \mathbf{m}(t_n+T)-n$ and rate function I.*

The proof of Theorem 3.1 relies on the weak convergence approach to large deviations, presented in the monographs [9, 4]. In particular it is divided into proving the upper bound,

$$\liminf_{n\to\infty} -\frac{1}{\beta_n}\log E\left[e^{-\beta_n F(X^n)}\right] \geq \inf_{\varphi}\left\{F(\varphi)+I(\varphi)\right\},$$

and the lower bound,

$$\limsup_{n\to\infty} -\frac{1}{\beta_n}\log E\left[e^{-\beta_n F(X^n)}\right] \leq \inf_{\varphi}\left\{F(\varphi)+I(\varphi)\right\},$$

where the infima are over $\varphi \in \mathcal{AC}_{x_0}([0,T] : \mathbb{R}^{d_1})$ and F is an arbitrary bounded continuous function. The proofs of the two bounds are given in Sections 6 and 7, respectively, and rely on the following representation formula that is a straightforward modification of Theorem 4.5 in [4].

Proposition 3.2 *Fix $n \in \mathbb{N}$ and let $\{X^n(t) : t \in [0,T]\}$ be the continuous interpolations of $\{X_k^n\}_{k\geq n}$ given by (2.1), and $X_n^n = x$. For any bounded continuous function $F : C([0,T] : \mathbb{R}^{d_1}) \to \mathbb{R}$,*

$$-\frac{1}{\beta_n}\log Ee^{-\beta_n F(X^n)} = \inf_{\{\bar\mu_i^n\}} E\left[F(\bar X^n) + \frac{1}{\beta_n}\sum_{i=n+1}^{\beta_n+n} R(\bar\mu_i^n(\cdot)\|\rho_{\bar X_{i-1}^n}(\bar Y_{i-1}^n,\cdot))\right], \tag{3.2}$$

where $\{\bar\mu_i^n\}_{i\in\{n+1,\dots,\beta_n+n\}}$ is a collection of random probability measures satisfying the following two conditions:

1. *$\bar\mu_i^n$ is measurable with respect to the σ-algebra $\mathcal{F}_{i-1}^n$, where $\mathcal{F}_n^n = \{\emptyset, \Omega\}$ and for $i \in \{n+1,\dots,\beta_n+n\}$, $\mathcal{F}_i^n = \sigma\{\bar Y_n^n,\dots,\bar Y_i^n\}$;*
2. *the conditional distribution of $\bar Y_i^n$, given $\mathcal{F}_{i-1}^n$, is $\bar\mu_i^n$.*

Moreover, $\{\bar X_k^n\}_{k\geq n}$ are defined by (2.1) with $\{Y_k\}$ replaced by $\{\bar Y_k^n\}$, and $\{\bar X^n(t) : t \in [0,T]\}$ is the continuous interpolations of $\{\bar X_k^n\}_{k\geq n}$.

Proof. Observe that $\{X^n(t) : t \in [0,T]\}$ are determined by $\{x, X_{n+1}^n, \dots, X_{\mathbf{m}(t_n+T)}^n\}$, which depends only on the state-dependent noise $\{Y_n,\dots,Y_{\mathbf{m}(t_n+T)-1}\}$ via the recursive formula. Therefore, the variational formula in [4, Proposition 2.3] combined with the chain rule for relative entropy [4, Theorem 2.6], with $\beta_n = \mathbf{m}(t_n+T)-n$ and base measure

$$\rho_{x_0^{\beta_n}}(y_0,dy_1)\rho_{x_1^{\beta_n}}(y_1,dy_2)\times\cdots\times\rho_{x_{\beta_n-1}^{\beta_n}}(y_{\beta_n-1},dy_{\beta_n}),$$

gives the claimed result. ∎

Let us briefly outline the main ideas of the proof. For the upper bound, for any $\varepsilon > 0$, from the representation formula we can choose a sequence of ε-optimal controls $\bar\mu^n = \{\bar\mu_i^n\}_{n+1}^{\beta_n+n}$.

This sequence in turn defines a controlled process $\bar{X}^n = \{\bar{X}^n(t) : t \in [0,T]\}$. To prove the upper bound, in Lemma 6.2 we show tightness of both the controls and the controlled process, and identify the limit $(\bar{X}, \bar{\mu})$ along a convergent subsequence of $\{(\bar{X}^n, \bar{\mu}^n)\}$. In particular we identify the limit ODE for $\bar{X}$, the limit of the controlled processes. With these results, the proof of the upper bound follows from fairly standard arguments involving Fatou's lemma, lower semi-continuity of relative entropy and the chain rule; see Section 6 for the complete details.

The difficult part of proving Theorem 3.1 is in proving the lower bound. Whereas for the upper bound we can use the definition of the infimum in (3.2) to obtain a suitable sequence of controls, for the lower bound we must explicitly construct a sequence of nearly-optimal controls $\bar{\nu}^n = \{\bar{\nu}_i^n\}_{i=n+1}^{\beta_n+n}$. This is carried out in Section 7.1. The first step is to show that for any trajectory ξ such that $I(\xi) < \infty$, for any $\varepsilon > 0$, there is a piece-wise linear ξ^* such that $||\xi^* - \xi||_\infty < \varepsilon$ and $I(\xi^*) \le I(\xi) + \varepsilon$ (see Lemma 7.3). Such trajectories, along with transition kernels that are nearly-optimal for the local rate function L—see Lemma 7.2—are used to construct the sequence of controls $\bar{\nu}^n$ for each n. Moreover, in Lemma 7.4 we show tightness of the sequence $\{\bar{\nu}^n\}_n$.

With suitable controls $\bar{\nu}^n$ identified, we obtain an upper bound of the right-hand side of the representation formula (3.2). It remains to show, that asymptotically in n, this upper bound is in turn bounded from above by $\inf_\rho\{F(\rho) + I(\rho)\}$. This is achieved in Section 7 through a series of approximations. An essential ingredient in the proof of the lower bound is to divide $[0,T]$ into subintervals, each containing a given number of time points t_j^n associated with the controlled process arising from the $\bar{\nu}_j^n$s. We use (local) ergodicity to show that, as the number of such time points in each subinterval grows, the controlled process converges and identify the corresponding limit process (7.9). Next, we show that as the number of intervals grows, this limit process converges to the trajectory ξ of interest. In Section 7 these approximations are combined to obtain the lower bound.

3.1 Alternative representations of the local rate function

Note that, for each $x \in \mathbb{R}^d$, J_x defined as in (2.2) with $\rho(y,dz)$ replaced by $\rho_x(y,dz)$ is the rate function associated with the empirical measure of a Markov chain with transition probability $\rho_x(y,dz)$. An alternative representation of $J_x(\mu)$, due to Donsker and Varadhan [6], is given by

$$\sup_{u>0} \int \log\left(\frac{u(y)}{\rho_x u(y)}\right)\mu(dy), \tag{3.3}$$

where the supremum is taken over strictly positive continuous functions u and $\rho_x u(y) = \int u(z)\rho_x(y,dz)$.

Another representation of J_x is provided by Dinwood and Ney, see [5] Lemma 3.1. For bounded Lipschitz functions, f, let T_f^x be the operator, on the space of bounded measurable functions with the uniform metric, given by

$$T_f^x(u)(y) = e^{f(y)}\rho_x u(y).$$

With $r_f(x)$ the spectral radius of T_f^x they prove that $J_x(\mu)$ can be represented as

$$\sup_f\left\{\int f(y)\mu(dy) - r_f(x)\right\}, \tag{3.4}$$

where the supremum is taken over bounded Lipschitz functions. Consequently, the local rate function in (2.3) can be written as,

$$L(x,\beta) = \inf_{\mu} \left\{ J_x(\mu) : \beta = \int g(x,z)\mu(dz) \right\},$$

where $J_x(\mu)$ is given by any of the expressions (2.2), (3.3) or (3.4).

3.2 The limiting Hamiltonian

Consider the Hamiltonian H given as the Fenchel-Legendre transform of the local rate function L in (2.3). That is,

$$H(x,\alpha) = \sup_{\beta} \{\alpha, \beta\rangle - L(x,\beta)\}.$$

Next, the Laplace principle for the empirical measure of a Markov chain and standard results from convex analysis will be used to show that $H(x,\alpha)$ can be interpreted as a limiting log-moment generating function associated with the transition probability $\rho_x(y,dz)$. We have the following result.

Proposition 3.3 *Suppose (A.5)-(A.7) of Assumption 2.2 holds. Take $x \in \mathbb{R}^d$, let $\{Y_i\}$ be a Markov chain with transition kernel ρ_x and J_x be defined as in* (2.2)*, with ρ replaced by ρ_x. Then,*

$$H(x,\alpha) = \lim_n \frac{1}{n} \log E\left[\exp\left\{\sum_{i=1}^n \langle \alpha, g(x,Y_i)\rangle\right\}\right], \qquad \alpha \in \mathbb{R}^d. \tag{3.5}$$

Proof. By (A.5) and (A.6) in Assumption 2.2 it follows that the empirical measure of $\{Y_i\}$ satisfies a Laplace principle on $\mathcal{P}(\mathbb{R}^{d_2})$ with rate function J_x, see [3, Theorem 6.6]. For every bounded and measurable function f the linear functional $\mu \mapsto \int f d\mu$, defined on $\mathcal{P}(\mathbb{R}^{d_2})$, is bounded and continuous. For each $x \in \mathbb{R}^{d_1}$, by the Laplace principle for the empirical measure of $\{Y_i\}$, the map $f \mapsto \hat{H}(x,f)$ given by the limit,

$$\begin{aligned}\hat{H}(x,f) &= \lim_n \frac{1}{n} \log E\left[\exp\left\{\sum_{i=1}^n f(Y_i)\right\}\right] \\ &= \lim_n \frac{1}{n} \log E\left[\exp\left\{-n\left\langle -f, \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}\right\rangle\right\}\right],\end{aligned}$$

is well defined on the set of bounded measurable functions. Moreover, $\hat{H}(x,f)$ may be identified as the Fenchel-Legendre transform of J_x,

$$\hat{H}(x,f) = \sup_{\mu} \{\langle f, \mu\rangle - J_x(\mu)\}.$$

By Assumption 2.2 (A.7), the function $\hat{H}(x,\cdot)$ may be extended to the, possibly unbounded, function $\langle \alpha, g(x,\cdot)\rangle$. Indeed, the function $\langle \alpha, g(x,\cdot)\rangle$ may be approximated from below by bounded measurable functions and the dominated convergence theorem can be applied since,

$$\begin{aligned}
&\sup_n \frac{1}{n}\log E\left[\exp\left\{\sum_{i=1}^{n}\langle \alpha, g(x,Y_i)\rangle\right\}\right]\\
&\quad= \sup_n \frac{1}{n}\log E\left[\exp\left\{\sum_{i=1}^{n-1}\langle \alpha, g(x,Y_i)\rangle\right\} E\left[\exp\left\{\langle \alpha, g(x,Y_n)\rangle\right\} \mid Y_{n-1},\dots,Y_1\right]\right]\\
&\quad= \sup_n \frac{1}{n}\log E\left[\exp\left\{\sum_{i=1}^{n-1}\langle \alpha, g(x,Y_i)\rangle\right\} \int \exp\left\{\langle \alpha, g(x,y_n)\rangle\right\}\rho_x(Y_{n-1},dy_n)\right]\\
&\quad\le \sup_n \frac{1}{n}\log \left(K\cdot E\left[\exp\left\{\sum_{i=1}^{n-1}\langle \alpha, g(x,Y_i)\rangle\right\}\right]\right)\\
&\quad\le K<\infty,
\end{aligned}$$

where $K=\sup_{x\in\mathbb{R}^{d_1}}\sup_{y\in K}\log\int_K e^{\langle\alpha,g(x,z)\rangle}\rho_x(y,dz)$. It remains to show that

$$H(x,\alpha)=\hat{H}(x,\langle\alpha,g(x,\cdot)\rangle),$$

is the Fenchel-Legendre transform of L, which is proved using a rather standard argument from convex analysis. Let us show that $L(x,\beta)=\sup_\alpha\{\langle\alpha,\beta\rangle-\hat{H}(x,\langle\alpha,g(x,\cdot)\rangle)\}$. Consider the set

$$\Gamma_x=\{(r,s)\subset\mathbb{R}\times\mathbb{R}^{d_1}: r\ge J_x(\mu),\int g(x,y)\mu(dy)=s,\ \text{some } \mu\in\mathcal{P}(\mathbb{R}^{d_2})\}.$$

Note that Γ_x is convex for each x. By taking a normal of the form $(1,\lambda_\beta)$ to the tangent plane of Γ_x at $(L(x,\beta),\beta)$ it follows that

$$\langle(1,\lambda_\beta),(r,s)-(L(x,\beta),\beta)\rangle\ge 0,\quad (r,s)\in\Gamma_x.$$

Moreover, the sup in

$$\sup_\alpha\inf_{(r,s)\in\Gamma_x}\{\langle(1,-\alpha),(r,s)-(L(x,\beta),\beta))\rangle\},$$

is attained at $-\alpha=\lambda_\beta$. Therefore, we have on one hand that

$$\sup_\alpha\inf_{(r,s)\in\Gamma_x}\{\langle(1,-\alpha),(r,s)-(L(x,\beta),\beta))\rangle\}=\inf_{(r,s)\in\Gamma}\{\langle(1,\lambda_\beta),(r,s)-(L(x,\beta),\beta))\rangle\}\ge 0.$$

On the other hand, for all α,

$$\inf_{(r,s)\in\Gamma_x}\{\langle(1,-\alpha),(r,s)-(L(x,\beta),\beta))\rangle\}\le 0.$$

Therefore,

$$\sup_\alpha\inf_{(r,s)\in\Gamma_x}\{\langle(1,-\alpha),(r,s)-(L(x,\beta),\beta))\rangle\}=\inf_{(r,s)\in\Gamma_x}\{\langle(1,\lambda_\beta),(r,s)-(L(x,\beta),\beta))\rangle\}=0,$$

which is equivalent to

$$\begin{aligned}
L(x,\beta) &= \sup_{\alpha} \inf_{(r,s)\in\Gamma_x} \{r - \langle \alpha, s-\beta\rangle\} \\
&= \sup_{\alpha} \{\langle \alpha, \beta\rangle + \inf_{(r,s)\in\Gamma_x} \{r - \langle \alpha, s\rangle\}\} \\
&= \sup_{\alpha} \{\langle \alpha, \beta\rangle + \inf_{\mu} \{J_x(\mu) - \langle \alpha, \int g(x,y)\mu(dy)\rangle\}\} \\
&= \sup_{\alpha} \{\langle \alpha, \beta\rangle - \sup_{\mu} \{\int \langle \alpha, g(x,y)\rangle \mu(dy) - J_x(\mu)\}\} \\
&= \sup_{\alpha} \{\langle \alpha, \beta\rangle - \hat{H}(x, \langle \alpha, g(x,\cdot)\rangle)\}.
\end{aligned}$$

This completes the proof. ■

Remark 3.4 *Using the representation* (3.5) *of the limiting Hamiltonian, it follows that the time-dependent limiting Hamiltonian, which is the Fenchel-Legendre transform of the time-dependent local rate function $L(t,x,\beta) = \frac{1}{h(t)}L(x,\beta)$ is given by,*

$$\begin{aligned}
H(t,x,\alpha) &= \sup_{\beta} \left\{ \{\alpha, \beta\rangle - \frac{1}{h(t)} L(x,\beta) \right\} \\
&= \frac{1}{h(t)} \sup_{\beta} \{\langle \alpha h(t), \beta\rangle - L(x,\beta)\} \\
&= \frac{1}{h(t)} H\left(x, \alpha h(t)\right).
\end{aligned}$$

3.3 Continuity of the local rate function

In this section we prove that, under Assumption 2.2, the local rate function L in (2.3) is continuous and every point where it is finite.

Lemma 3.5 *Suppose (A.1),(A.2),(A.5) and (A.7) hold. For any $(x_1, \beta_1) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_1}$ such that $L(x_1, \beta_1) < \infty$, L is continuous at (x_1, β_1).*

Proof. Let H be the limiting Hamiltonian given by (3.5). By Proposition 3.3, $L(x,\cdot)$ is equal to the Legendre-Fenchel transform of $H(x,\cdot)$. In [14] the authors show that $\alpha \mapsto H(x,\alpha)$ is convex and smooth; see also Section 4.3 in [8]. To prove the continuity of L at (x,β), by the arguments used in [4, Lemma 4.16 (f)], it suffices to show the continuity of $H(x,\alpha)$ in (x,α).

To prove that $(x,\alpha) \mapsto H(x,\alpha)$ is continuous it is sufficient to show that the family $\{H_n\}_{n\in\mathbb{N}}$ with

$$H_n(x,\alpha) = \frac{1}{n} \log E\left[\exp\left\{\sum_{i=1}^{n} \langle \alpha, g(x, Y_i)\rangle\right\}\right], \quad (x,\alpha) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_1},$$

is equicontinuous.

By Assumption (A.2), for each $\varepsilon_1 > 0$ there exists $\delta_1 > 0$ such that $|x_1 - x_2| < \delta_1$ and $|\alpha_1 - \alpha_2| < \delta_1$ implies that

$$-\varepsilon_1 \leq \Lambda(x_1, \alpha_1, y) - \Lambda(x_2, \alpha_2, y) \leq \varepsilon_1, \qquad y \in K.$$

By exponentiating each expression in the last display and selecting ε_1 sufficiently small there is, for each $\varepsilon > 0$, a $\delta > 0$ such that $|x_1 - x_2| < \delta$ and $|\alpha_1 - \alpha_2| < \delta$ implies that,

$$1 - \varepsilon \leq \frac{\int \exp\{\langle \alpha_1, g(x_1, z)\rangle\}\rho_{x_1}(y, dz)}{\int \exp\{\langle \alpha_2, g(x_2, z)\rangle\}\rho_{x_2}(y, dz)} \leq 1 + \varepsilon, \qquad y \in K.$$

Repeatedly applying the inequalities in the previous display yields

$$\begin{aligned}
(1-\varepsilon)^n &\int \cdots \int e^{\langle \alpha_2, g(x_2, y_1)\rangle + \cdots + \langle \alpha_2, g(x_2, y_n)\rangle} \rho_{x_2}(y_0, dy_1) \cdots \rho_{x_2}(y_{n-1}, dy_n) \\
&\leq \int \cdots \int e^{\langle \alpha_1, g(x_1, y_1)\rangle + \cdots + \langle \alpha_1, g(x_1, y_n)\rangle} \rho_{x_1}(y_0, dy_1) \cdots \rho_{x_1}(y_{n-1}, dy_n) \\
&\leq (1+\varepsilon)^n \int \cdots \int e^{\langle \alpha_2, g(x_2, y_1)\rangle + \cdots + \langle \alpha_2, g(x_2, y_n)\rangle} \rho_{x_2}(y_0, dy_1) \cdots \rho_{x_2}(y_{n-1}, dy_n).
\end{aligned}$$

Applying $\frac{1}{n}\log$ and rearranging the inequalities we obtain,

$$\log(1-\varepsilon) \leq H_n(x_1, \alpha_1) - H_n(x_2, \alpha_2) \leq \log(1+\varepsilon).$$

This proves that $\{H_n\}_{n \in \mathbb{N}}$ is equicontinuous and completes the proof. ■

4 Related work for constant and decreasing step size

The literature on large deviations for recursive algorithms of the form (2.1) studies, on the one hand, the setting with constant step size, where $\varepsilon_n = \varepsilon > 0$, does not depend on n, and consequently $t_n = \varepsilon n$ and $\mathbf{m}(t) = \lfloor n\varepsilon \rfloor$. In this setting, large deviations principles for the piecewise linearly interpolated process $X^\varepsilon(t)$ of $\{X_n^\varepsilon\}$ with interpolation time ε, are obtained as $\varepsilon \to 0$. The associated rate function takes the form of an action functional,

$$I(\varphi) = \int_0^T L(\varphi(t), \dot{\varphi}(t)) ds,$$

if φ is an absolutely continuous function, and $I(\varphi) = \infty$, otherwise, where L is a local rate function.

On the other hand, in the setting with decreasing step size, where $\varepsilon_n \to 0$ as $n \to \infty$, large deviations principles are obtained for the process $\{X^n\}$ defined in (2.1). In this case, with $\beta_n \doteq m(t_n + T) - n$ the limiting time scale is $h(t) = \lim_n h^n(t)$, where $h^n(t) = \beta_n \varepsilon_{n+k-1}$ for $t \in [t_{n+k-1}.t_n, t_{n+k} - t_n)$, $k \in \{1, \dots, \beta_n\}$, and the rate function takes the form

$$I(\varphi) = \int_0^T \frac{1}{h(t)} L(\varphi(t), \dot{\varphi}(t)) ds.$$

The main difference between the constant and decreasing step size settings is the inclusion of the limiting time scale $h(t)$ in the rate function. Note, however, that for some choices of ε_n, such as $\varepsilon_n = (n+1)^{-\alpha}$, for $\alpha \in (0,1)$, the limiting time scale $h(t)$ may be constant and equal to 1.

In the existing literature, the large deviations principle is obtained by identifying a Hamiltonian $H(x,\alpha)$, that sometimes can be interpreted as a limiting log-moment generating function and defining the local rate function $L(x,\beta)$ as the convex conjugate of $H(x,\alpha)$. A problem with this approach is that the Hamiltonian is defined as a limit and its relation to the underlying dynamics such as the transition kernel $\rho_x(y,dz)$ can only be established in some special cases.

4.1 Large deviations for constant step size

The large deviations theory for stochastic approximation with constant step size originates from the work of Freidlin [12, 13] who studies dynamical systems in continuous time, of the form,

$$\dot{x}^\varepsilon(t) = b(x^\varepsilon(t), \xi(t/\varepsilon)), \qquad x^\varepsilon(0) = x, \tag{4.1}$$

over a finite time interval, $[0,T]$, where the function b is bounded, with bounded derivatives, $\{\xi(t), t \geq 0\}$ is bounded and $\varepsilon \to 0$. Freidlin assumes that there is a limiting Hamiltonian $H(x,\alpha)$ such that for arbitrary step functions φ and α from $[0,T]$ to $\mathbb{R}^{d_1}$, the following limit exists

$$\lim_{\varepsilon\to 0} \varepsilon \log E\left[\exp\left\{\frac{1}{\varepsilon}\int_0^T \langle \alpha(t), b(\varphi(t), \xi(t/\varepsilon))\rangle dt\right\}\right] = \int_0^T H(\varphi(t), \alpha(t))dt. \tag{4.2}$$

With L as the convex conjugate of H Freidlin proves a large deviations principle on $C^x_{0,T} = \{\varphi \in C([0,T];\mathbb{R}^{d_1}), \varphi(0) = x\}$ for $\{x^\varepsilon\}$ as $\varepsilon \to 0$, with rate function given by

$$I(\varphi) = \int_0^T L(\varphi(t), \dot{\varphi}(t))ds,$$

if φ is absolutely continuous and $I(\varphi) = \infty$, otherwise. When $\{\xi(t)\, t \geq 0\}$ is a finite state Markov chain, Freidlin identifies the limiting Hamiltonian as the largest eigenvalue of a tilted intensity matrix.

Iscoe, Ney and Nummelin [14] generalize the results of Freidlin by considering large deviations principles for Markov-additive processes in both continuous time and discrete time. In the discrete time setting, which relates more closely to the results of this paper, they consider a process of the form (Y_n, X_n) where

$$P((Y_n, X_n - X_{n-1}) \in A \times \Gamma \,|\, (Y_{n-1}, X_{n-1}) = (y,x)) = P((Y_n, X_n - X_{n-1}) \in A \times \Gamma \,|\, Y_{n-1} = y).$$

That is, equations of the form (2.1) where $g(x,y) = g(y)$, does not depend on x. They assume that there exists a probability measure ν on $E \times \mathbb{R}^{d_1}$, an integer m_0, and real numbers $0 < a \leq b < \infty$ such that

$$a\nu(A \times \Gamma) \leq P^{m_0}(A \times \Gamma) \leq b\nu(A \times \Gamma),$$

for all $x \in E$, $A \in \mathscr{E}$, $\Gamma \in \mathbb{R}^{d_1}$. With $\hat{P}(\alpha) = \hat{P}(y, A; \alpha) = \int \exp\{\langle \alpha, x\rangle\}P(y, A \times dx)$ they derive a large deviations principle, and more detailed asymtotics, for $P^n(y, A \times nF) = P((Y_n, X_n - X_0) \in A \times nF \,|\, Y_0 = y)$. In particular, it follows from Lemma 3.1 (ii) in [14] that

$$\lim_{n\to\infty} \frac{1}{n} \log \hat{P}^n(y, A; \alpha) = \log \lambda(\alpha), \quad \alpha \in \mathcal{D}.$$

Dupuis [8] further develops the large deviations results for discrete systems of the form (2.1) with constant step size, using a milder conditions on the limiting Hamiltonian. More specifically, in Section 4.3 of his paper, Dupuis considers the model (2.1) with $\varepsilon_n = \varepsilon$ and $g(x, y)$ bounded and uniformly (in y) Lipschitz continuous in x, and measurable in y. He proves a large deviations principle with rate function I under the following assumptions. The process Y_n is sampled from a transition kernel $\rho_{X_n}(y, \cdot)$ with density $\eta_{X_n}(y, \cdot)$ with respect to a common reference measure $\lambda(dz)$ such that, for a given compact set F_1,

1. There are $0 < a \le A < \infty$ such that for all $x \in F_1$, and all y, z, $a \le \eta_x(y, z) \le A$, and
2. $\eta_x(y, z)$ is Lipschitz continuous in x, uniformly in y, z, for $x \in F_1$.

4.2 Large deviations for decreasing step size

Large deviations principles for the case of decreasing step size is not as well developed. The first results are obtained by Kushner [17] who considers (2.1), with $\varepsilon_n = (n+1)^{-\rho}$, $\rho \in (0, 1]$, and $g(x, y)) = b(x) + y$, with $b(\cdot)$ Lipschitz continuous and $\{Y_n\}$ a sequence of iid centered Gaussian variables. The discrete time and time-changed analogue of (4.2) is given by

$$\lim_{n\to\infty} \lambda_n \log E\left[\exp\left\{\sum_{i=0}^{N-1}\left\langle \alpha(i\Delta), \sum_{j=\mathbf{m}(t_n+i\Delta)}^{\mathbf{m}(t_n+(i+1)\Delta)-1} \varepsilon_j(b(x) + Y_j)/\lambda_n\right\rangle\right\}\right] = \int_0^T H(t, x, \alpha(t))dt,$$

where $T = N\Delta$, $\Delta > 0$, α is constant on intervals $[i\Delta, (i+1)\Delta)$. Kushner identifies the appropriate normalising sequence,

$$\lambda_n = \sum_{j=n}^{m(t_n+T)} \varepsilon_j^2,$$

which can be shown to be asymptotically proportional to $1/\beta_n$ with β_n as in Assumption 2.2 ((A.8)), and proves a large deviations principle with rate function

$$I(\varphi) = \int_0^T L(t, \varphi(t), \dot{\varphi}(t))ds,$$

where the local rate function $L(t, \varphi(t), \dot{\varphi}(t))$ is the convex conjugate of H.

Dupuis and Kushner [11] develop the theory further by considering recursions of the form (2.1) with $g(x, y) = \bar{b}(x) + b(x, y)$, $\varepsilon_n \ge 0$, $\sum_n \varepsilon_n = \infty$, $\varepsilon_n \to 0$ and $E[b(x, Y_n)] = 0$. They assume further that $Y_n = (\tilde{Y}_n, \hat{Y}_n)$, where $\{\tilde{Y}_n\}$ and $\{\hat{Y}_n\}$ are mutually independent, $\{\tilde{Y}_n\}$ is stationary and bounded and $\{\hat{Y}_n\}$ is stationary centered Gaussian process with summable correlation function. Moreover, $b(x, Y_n) = b_1(x, \tilde{Y}_n) + b_0(x)\hat{Y}_n$, where $b_1(\cdot, \tilde{y})$, b_0 and $\bar{b}$ are uniformly (in $\tilde{y}, x$)

Lipschitz and bounded. Note that, in contrast to our setting, in [11] it is not assumed that the distribution of the noise Y_{n+1} may depend on the state X_n. It is assumed that there exists a continuous function h_1 such that

$$\lim_{\delta\to 0}\lim_{n\to\infty}\frac{\varepsilon_{m_n(t+\delta)}}{\varepsilon_n}=h_1(t),$$

Further, the existence of a limiting Hamiltonian is assumed in [11]. That is, there is a continuous function $H(t,x,\alpha)$ with $\alpha\mapsto H(t,x,\alpha)$ continuously differentiable for each t,x such that the following limit exists,

$$\lim_{\delta\to 0}\lim_{n\to\infty}\lambda_n\log E\left[\exp\left\{\sum_{i=0}^{T/\delta-1}\left\langle\alpha(i\delta)\varepsilon_{m_n(i\delta)},\sum_{j=m(i\delta)}^{m((i+1)\delta)-1}b(\psi(i\delta),Y_j)\right\rangle\right\}\right]=\int_0^T H(t,x,\alpha(t))dt.$$

A particular example studied in [11] is when $\{Y_n,-\infty<n<\infty\}$ is bounded and stationary and there is a continuous $\hat{H}_0(\cdot,\cdot)$ with $\alpha\mapsto\hat{H}_0(\alpha,x)$ continuously differentiable for each x such that

$$\begin{aligned}&\lim_{N\to\infty}\frac{1}{N}\log E\left[\exp\left\{\left\langle\alpha,\sum_{j=0}^{N-1}b(\psi,Y_j)\right\rangle\right\}\right]\\&=\lim_{N\to\infty}\frac{1}{N}\log E_0\left[\exp\left\{\left\langle\alpha,\sum_{j=0}^{N-1}b(\psi,Y_j)\right\rangle\right\}\right]\\&=\hat{H}_0(\alpha,\psi),\end{aligned}$$

where the convergence is uniform in the conditioning data. Note that the limiting Hamiltonian established in Proposition 3.3 provides an analogous representation in the setting where the distribution of the noise may be state dependent.

5 Applications

In this section we present applications to learning algorithms in statistics, machine learning and statistical physics that can be stated as stochastic approximations satisfying Assumption 2.2.

5.1 Stochastic gradients

Consider minimizing a function $G(x)=\sum_{m=1}^{M}G_m(x)$, by stochastic gradient descent (SGD). Let us assume that $x\mapsto\nabla G_m(x)$ is bounded and Lipschitz continuous for all $m\in\{1,\dots,M\}$. Consider a standard SGD algorithm; in the kth iteration an index Y_{k+1} is selected uniformly at random on $\{1,\dots,M\}$ and updated according to

$$X_{n+1}=X_n-\varepsilon_{n+1}\nabla G_{Y_{n+1}}(X_n).$$

Consequently, $\{X_n\}$ satisfies the stochastic approximation (2.1) where $\{Y_k\}$ is an iid sequence, $\rho_x(y,\cdot)=\rho(\cdot)$ is the uniform distribution on the integers $\{1,\dots,M\}$, and $g(x,m)=-\nabla G_m(x)$.

Assumption 2.2 is automatically satisfied by the assumptions on ∇G_m, λ as the counting measure and since $\rho_x(y,m)$ does not depend on x, y.

By Theorem 3.1 the continuous interpolations of $\{X_k^n\}$ given by (2.1) satisfies a Laplace principle with rate function I given by (3.1) where the local rate function L is given by (2.3). Since, $\rho_x(y,m) = \rho(m) = 1/M$ does not depend on x, y the local rate function simplifies to

$$L(x,\beta) = \inf_{\mu} \left\{ R(\mu \| \rho) : \beta = -\sum_{m=1}^{M} \nabla G_m(x)\mu(m) \right\} = \sup_{\alpha} \{\langle \alpha, \beta \rangle - \bar{H}(x,\alpha)\},$$

where

$$\bar{H}(x,\alpha) = \log \left(\frac{1}{M} \sum_{m=1}^{M} \exp\left\{ -\langle \alpha, \nabla G_m(x) \rangle \right\} \right).$$

A concrete example arises in maximum likelihood estimation of a logistic regression with data $\{(\xi_m, \upsilon_m)\}_{m=1}^M$ where ξ_m are explanatory variables and υ_m labels in $\{-1, 1\}$ and ϕ represents a feature function. Then the negative log-likelihood to be minimized is given by

$$G(x) = \sum_{m=1}^{M} -\log \operatorname{sigm}\left(\upsilon_m x^T \phi(\xi_m) \right)$$

where $\operatorname{sigm}(t) = (1 + e^{-t})^{-1}$ is the sigmoid function and

$$\nabla G_m(x) = \upsilon_m \phi(\xi_m) \left(1 - \operatorname{sigm}(\upsilon_m x^T \phi(\xi_m) \right),$$

which is bounded and Lipschitz continuous in x, for all $m \in \{1, \dots, M\}$.

More general stochastic gradients appear in the minimization of functions of the form $\bar{G}(x) = \int G(x,y)\gamma(dy)$ for some distribution γ. With $\{Y_n\}$ iid with distribution γ the algorithm

$$X_{n+1} = X_n - \varepsilon_{n+1} \nabla_x G(X_n, Y_{n+1}),$$

can be used to minimize $\bar{G}$. If $\nabla_x G(x,y)$ is bounded and Lipschitz continuous in x, then Assumption 2.2 is satisfied and the Laplace principle holds with local rate function

$$L(x,\beta) = \inf_{\mu} \left\{ R(\mu \| \gamma) : \beta = -\int \nabla_x G(x,y)\mu(dy) \right\} = \sup_{\alpha} \{\langle \alpha, \beta \rangle - \bar{H}(x,\alpha)\},$$

where

$$\bar{H}(x,\alpha) = \log \left(\int \exp\left\{ -\langle \alpha, \nabla_x G(x,y) \rangle \right\} \gamma(dy) \right).$$

5.2 Persistent contrastive divergence

Consider parametrized a probability density of the form

$$p(v, h|x) = \exp\left\{ -E(v,h;x) + F(x) \right\},$$

where x denotes the parameters. We assume that v represents observed (visible) variables, h represents unobserved (hidden) variables, E is referred to as the energy and F as the free energy,

$$F(x) = -\log \int \exp\{-E(v,h;x)\}\lambda(dv,dh).$$

Given independent observations $v^{(1)},\dots,v^{(M)}$ from $p(v|x) = \int p(v,h|x)\lambda(dh)$ the parameters x may be estimated by minimizing the negative log-likelihood, which is proportional to:

$$-\log L(x) = -\frac{1}{m}\sum_{m=1}^{M} \log p(v^{(m)}|x).$$

A gradient descent algorithm would require knowledge of the gradient

$$\begin{aligned}
-\nabla_x \log L(x) &= -\frac{1}{m}\sum_{m=1}^{M} \frac{\nabla_x \int \exp\left\{-E(v^{(m)},h;x) + F(x)\right\}\lambda(dh)}{p(v^{(m)}|x)} \\
&= \frac{1}{m}\sum_{m=1}^{M} \frac{\int \left(\nabla_x E(v^{(m)},h;x) - \nabla_x F(x)\right) p(v^{(m)},h|x)\lambda(dh)}{p(v^{(m)}|x)} \\
&= \frac{1}{m}\sum_{m=1}^{M} \left[\int \nabla_x E(v^{(m)},h;x) p(h|v^{(m)},x)\lambda(dh) - \nabla_x F(x)\right] \\
&= \frac{1}{m}\sum_{m=1}^{M} \left[\int \nabla_x E(v^{(m)},h;x) p(h|v^{(m)},x)\lambda(dh) - \int \nabla_x E(v,h;x) p(v,h|x)\lambda(dv,dh)\right], \qquad (5.1)
\end{aligned}$$

which may be intractable. Simplifying model assumptions may assist in computing the first term explicitly as illustrated in some of the examples below. In the general case we may write

$$p(h|v,x) = \exp\left\{-E(v,h;x) + F_H(v,x)\right\},$$

where

$$F_H(v,x) = -\log \int \exp\left\{-E(v,h;x)\right\}\lambda(dh).$$

To approximate the gradient in (5.1) we may construct Markov kernels, $\rho_x^{(m,1)}(y^{(1)},dz^{(1)})$ and $\rho_x^{(2)}(y^{(2)},dz^{(2)})$, where $y^{(1)} = h$, $y^{(2)} = (v,h)$ and $\rho_x^{(m,1)}(y^{(1)},dz^{(1)})$ has invariant distribution $p(h|v^{(m)},x)$ and $\rho_x^{(2)}(y^{(2)},dz^{(2)})$ has invariant distribution $p(v,h|x)$. We sample $Y_{n+1} = (Y_{n+1}^{(1)}, Y_{n+1}^{(2)})$ by drawing an index m at random and drawing $Y_{n+1}^{(1)}$ from $\rho_{X_n}^{(m,1)}(Y_n^{(1)},dz^{(1)})$ and $Y_{n+1}^{(2)}$ from $\rho_{X_n}^{(2)}(Y_n^{(2)},dz^{(2)})$ independently of each other and updating

$$X_{n+1} = X_n - \varepsilon_{n+1}\left(\nabla_x E(v^{(m)}, Y_{n+1}^{(1)}; X_n) - \nabla_x E(Y_{n+1}^{(2)}; X_n)\right).$$

This can be identified as the stochastic recursion (2.1) with

$$\rho_x(y,dz) = \frac{1}{m}\sum_{m=1}^{M} \rho_x^{(m,1)}(y^{(1)},dz^{(1)})\rho_x^{(2)}(y^{(2)},dz^{(2)}),$$

and

$$g(x,y) = \nabla_x E(v^{(m)}, y^{(1)}; x) - \nabla_x E(y^{(2)}; x).$$

Example 5.1 *In Restricted Boltzmann Machines (RBMs) v and h are binary with $x = (W, b_V, b_H)$ where W is a matrix and b_V, b_H vectors and*

$$E(v, h; W, b_V, b_H) = -v^T W h - v^T b_V - h^T b_H,$$

which implies that the components of h are conditionally independent given v with success probability $p(h_j = 1|v, x) = \mathrm{sigm}(v^T W e_j + e_j^T b_H)$ and the first term in (5.1) reduces to

$$\begin{aligned}
&\frac{1}{m}\sum_{m=1}^{M}\sum_{h}\Big[\nabla_{W_{ij}} E(v^{(m)}, h; W, b_V, b_H)\Big]\, p(h|v^{(m)}, W, b_V, b_H)\\
&\quad = \frac{1}{m}\sum_{m=1}^{M}\sum_{h} -v_i^{(m)} h_j \mathrm{sigm}(v^T W e_j + e_j^T b_H)^{h_j} \mathrm{sigm}(-(v^T W e_j + e_j^T b_H))^{1-h_j}\\
&\quad = -\frac{1}{m}\sum_{m=1}^{M} v_i^{(m)} \mathrm{sigm}(v^T W e_j + e_j^T b_H).
\end{aligned}$$

The second term is given by the expectation $E[V_i H_j]$ under the joint distribution $p(v, h|x)$. Let $\rho_x((v_0, h_0), (v_1, h_1))$ denote a stochastic kernel with $p(v, h|x)$ as its invariant distribution and approximate $E[V_i H_j]$ by its expectation under $\rho_x((v_0, h_0), \cdot)$, where $\rho_x((v_0, h_0), \cdot)$ may be taken as the block-Gibbs sampler

$$\begin{aligned}
&\rho_x((v_0, h_0), (v_1, h_1)) = p(h_1|v_0, x) p(v_1|h_1, x)\\
&\quad = \prod_{j=1}^{d_H} \mathrm{sigm}(v_0^T W e_j + e_j^T b_H)^{h_{1j}} \mathrm{sigm}(-(v_0^T W e_j + e_j^T b_H))^{1-h_{1j}}\\
&\qquad \times \prod_{i=1}^{d_V} \mathrm{sigm}(e_i^T W h_1 + e_i^T b_V)^{v_{1i}} \mathrm{sigm}(-(e_i^T W h_1 + e_1^T b_V))^{1-v_{1i}}.
\end{aligned}$$

The persistent contrastive divergence algorithm for estimating the parameters is then given by (2.1) where $Y_{n+1} = (v_{n+1}, h_{n+1})$ is sampled from $\rho_{X_n}(Y_n, \cdot)$ and

$$g(x,y) = \frac{1}{m}\sum_{m=1}^{M}\left[\int \nabla_x E(v^{(m)}, h; x) p(h|v^{(m)}, x)\lambda(dh) - \nabla_x E(y; x)\right].$$

It is straightforward to verify Assumption 2.2, since sigm *is bounded and continuous and the state space $\{0,1\}^{d_V} \times \{0,1\}^{d_H}$ is finite.*

Example 5.2 *Consider an exponential family with $E(v, h; x) = E(v; x) = x^T \phi(v) - \log c(v)$, that does not depend on hidden variables and is linear in the parameters x. Then $\nabla_x E(v^{(m)}; x) = \phi(v^{(m)})$ whereas $\nabla_x F(x) = E[\phi(V)]$ where the expectation is taken under $p(v|x)$ and may be intractable. Thus, $g(x, y)$ becomes*

$$g(x,y) = \frac{1}{m}\sum_{m=1}^{M}\phi(v^{(m)}) - \int \phi(v) p(v|x)\lambda(dv).$$

5.3 The Wang-Landau Algorithm

The Wang-Landau algorithm for general state spaces includes many popular multicanonical Monte Carlo methods, such as simulated tempering. Let $\{(\mathcal{Y}_i, \mathcal{B}_i, \lambda_i)\}_{i=1}^d$ be measure spaces with λ_i being σ-finite for each i. Let $\mathcal{Y} = \cup_{i=1}^d \mathcal{Y}_i \times \{i\}$, be the union space equipped with the σ-field $\mathcal{B}$ generated by the sets $\{(A_i, i) : i \in \{1, \dots, d\}, A_i \in \mathcal{B}_i\}$ and define the measure λ on $\mathcal{B}$ by $\lambda(A, i) = \lambda_i(A) I\{A \in \mathcal{B}_i\}$. Given non-negative integrable functions f_i, $i = 1 \dots, d$, let $x(i) = \int_{\mathcal{Y}_i} f_i(y)\lambda_i(dy)/Z$, where $Z = \sum_{i=1}^d \int_{\mathcal{Y}_i} f_i(y)\lambda_i(dy)$. Assuming that $x(i) > 0$ for each $i = 1, \dots, d$, the aim is to sample from π on $\mathcal{B}$ given by

$$\pi(dy, i) \propto \frac{f_i(y)}{x(i)} I\{y \in \mathcal{Y}_i\}\lambda_i(dy),$$

and to estimate the normalizing constants $x(i)$. Let $\rho_x((y, i), (dz, j))$ be a Markov kernel with invariant density π. The original algorithm considers the case where π is uniform in i, whereas the general case considered here is due to [1]. The basic for of the Wang-Landau algorithm initiates $(Y_0, I_0) \in \mathcal{Y}$, $\phi_0 \in (0, \infty)^d$ and $x_0 = \phi_0 / \sum_{i=1}^d \phi_0(i)$. At each $k \geq 0$, given (Y_k, I_k), ϕ_k and x_k, sample (Y_{k+1}, I_{k+1}) from $\rho_{x_k}((Y_k, I_k), \cdot)$ and update

$$\begin{aligned} \phi_{k+1}(i) &= \phi_k(i)(1 + \varepsilon_k I\{I_{k+1} = i\}), \quad i = 1, \dots, d, \\ x_{k+1}(i) &= \frac{\phi_{k+1}(i)}{\sum_{j=1}^d \phi_{k+1}(j)}. \end{aligned}$$

The Wang-Landau algorithm is a stochastic approximation with update function $g(\phi, (z, j)) = \phi + \phi(j)e_j$, where e_j is the unit-vector in the jth coordinate.

Example 5.3 (Multicanonical Monte Carlo) *Let Σ be a finite state space, e.g. $\{-1, 1\}^N$, and $E : \Sigma \to \mathbb{R}$ an energy function and consider the Gibbs distribution with density $\bar{\pi}$ proportional to $\exp\{-E(\sigma)\}$. A collection of energy levels $-\infty \leq E_0 < \cdots < E_d \leq \infty$ induces a partition $\mathcal{Y}_i = \{\sigma \in \Sigma : E_{i-1} < E(\sigma) \leq E_i\}$. With $x(i) = \bar{\pi}(\mathcal{Y}_i)$, $f_i(y) = E(y)$ and $\lambda_i = \lambda$ samples from the measure $\pi(dy, i)$ may be obtained to estimate $x(i)$.*

Example 5.4 (Estimation of free energy differences) *Let Σ be a finite state space, e.g. $\{-1, 1\}^N$, and $E : \Sigma \times \Omega \to \mathbb{R}$ an energy function parametrized by a finite set Ω (for example temperatures) and consider the Gibbs distribution with density $p_{\Sigma,\Omega}$ proportional to $\exp\{-E(\sigma, \omega)\}$. The conditional density of the state given the parameter ω is given by $p_{\Sigma|\Omega}(\sigma|\omega) = \exp\{-E(\sigma, \omega) + F(\omega)\}$, where $F(\omega) = -\log \sum_\sigma \exp\{-E(\sigma, \omega)\}$ is the free energy. Consider the problem of estimating free energy differences. That is, fix $\omega_1 \in \Omega$ and consider estimating $F(\omega) - F(\omega_1)$ for $\omega \in \Omega$. By enumerating $\Omega = \{\omega_i\}_{i=1}^d$, letting $\mathcal{Y}_i = \Sigma \times \Omega$, λ_i be counting measure on $\Sigma \times \Omega$ and $f_i(\sigma, \omega) = \exp\{-E(\sigma, \omega_i)\} I\{\omega = \omega_i\}$ it follows that*

$$-\log(x(i)/x(1)) = \log \sum_\sigma \exp\{-E(\sigma, \omega_1)\} - \log \sum_\sigma \exp\{-E(\sigma, \omega_i)\} = F(\omega_i) - F(\omega_1).$$

Since, $-\log(x(i)/x(1))$ may be estimated by $-\log(\phi_k(i)/\phi_k(1))$ where ϕ_k is generated by the Wang-Landau algorithm, the free energy differences may be estimated accordingly.

6 Laplace upper bound

In this section we take the first step towards proving Theorem 3.1, by proving the Laplace principle upper bound.

Theorem 6.1 *Assume (A.1)-(A.8). With I defined as in* (3.1)*, for any bounded, continuous function $F : C([0,T] : \mathbb{R}^{d_1}) \to \mathbb{R}$,*

$$\liminf_{n\to\infty} -\frac{1}{\beta_n} \log E\left[e^{-\beta_n F(X^n)}\right] \geq \inf_{\varphi} \left(F(\varphi) + I(\varphi)\right), \tag{6.1}$$

where the infimum is over $\varphi \in AC_{x_0}([0,T] : \mathbb{R}^d)$.

From the representation formula (3.2), for fixed n and arbitrary (fixed) $\varepsilon > 0$, it is possible to choose a sequence of controls $\{\bar{\mu}^n\}$ such that

$$-\frac{1}{\beta_n} \log E\left[e^{-\beta_n F(X^n)}\right] + \varepsilon \geq E\left[F(\bar{X}^n) + \frac{1}{\beta_n}\sum_{i=n+1}^{\beta_n+n} R(\bar{\mu}_i^n(\cdot)||\rho_{\bar{X}_{i-1}^n}(\bar{Y}_{i-1}^n,\cdot))\right]. \tag{6.2}$$

We augment the controls to also keep track of the time dependence of the $\bar{\mu}_i^n$s: for a Borel set A and $t \in [t_n, t_n + T]$, define $\bar{\mu}^n(A|t)$ by

$$\bar{\mu}^n(A|t) = \bar{\mu}_i^n(A), \quad \text{for } i \text{ such that } t \in [\tau_i^n, \tau_{i+1}^n),$$

where $\tau_i^n = t_{n+i} - t_n$. The controlled measures $\bar{\mu}^n$ can now be defined as

$$\bar{\mu}^n(A \times C) = \int_C \frac{1}{h^n(t)} \bar{\mu}^n(A|t)dt,$$

where

$$h^n(t) = \beta_n \varepsilon_{n+i-1},$$

with $i \in \{n+1, \dots, \beta_n + n\}$ such that $t \in [\tau_i^n, \tau_{i+1}^n)$. Lastly, we define a collection of sequences of measures, involving the controlled process $\bar{X}^n$, the controlled noise $\bar{Y}^n$ and the noise distribution ρ, that will play a role in the convergence analysis of the controlled process $\bar{X}^n$ and the corresponding controls $\bar{\mu}^n$: for $A, B \subset \mathbb{R}^{d_2}, C \subset [0,T]$ Borel sets,

$$\begin{aligned}
\lambda^n(A \times B \times C) &= \int_C \frac{1}{h^n(t)} \lambda^n(A \times B|t)dt, \quad \lambda^n(A \times B|t) = \delta_{\bar{Y}_{i-1}^n}(A)\bar{\mu}_i^n(B),\\
\gamma^n(A \times B \times C) &= \int_C \frac{1}{h^n(t)} \gamma^n(A \times B|t)dt, \quad \gamma^n(A \times B|t) = \delta_{\bar{Y}_{i-1}^n}(A)\rho_{\bar{X}_{i-1}^n}(\bar{Y}_{i-1}^n, B).
\end{aligned}$$

In each definition, i is such that $t \in [\tau_i^n, \tau_{i+1}^n)$. From the definitions of $\bar{\mu}^n$ and λ^n, we have that $\bar{\mu}^n(A \times C) = \lambda^n(\mathbb{R}^{d_2} \times A \times C)$. The following lemma establishes the necessary tightness and characterises the limits of subsequences of the sequences of measures defined above.

Lemma 6.2 *Assume (A.1)-(A.8) hold. Then $\{\bar{X}^n\}$, $\{\bar{\mu}^n\}$, $\{\bar{\lambda}^n\}$ and $\{\gamma^n\}$ are tight sequences, and for every subsequence of $\{\bar{X}^n, \bar{\mu}^n\}$ there exists a further subsequence that converges to $(\bar{X}, \bar{\mu})$, which satisfies the following relations:*

$$\bar{\mu}(A \times C) = \int_C \frac{1}{h(t)} \bar{\mu}(A|t)dt, \tag{6.3}$$

$$\bar{X}(t) = x + \int_0^t \int_{\mathbb{R}^d} g(\bar{X}(s), y)\bar{\mu}(dy|s)ds. \tag{6.4}$$

Furthermore, any limit point λ and γ of a convergent subsequence of $\{\lambda^n\}$ and $\{\gamma^n\}$, respectively, will have the following properties,

$$\begin{aligned}\lambda(A \times B \times C) &= \int_C \frac{1}{h(t)} \lambda(A \times B|t)dt,\\ \gamma(A \times B \times C) &= \int_C \frac{1}{h(t)} \left(\int_A \rho_{\bar{X}(t)}(x, B)\bar{\mu}(dx|t) \right) dt,\end{aligned}$$

for some stochastic kernel $\lambda(dy \times dz|t)$, and

$$\lambda(A \times \mathbb{R}^{d_2} \times C) = \lambda(\mathbb{R}^{d_2} \times A \times C) = \bar{\mu}(A \times C) = \int_C \frac{1}{h(t)} \bar{\mu}(A|t)dt.$$

Before giving the proof of Lemma 6.2, we show how the result allows us to prove the upper bound (6.1).

Proof of Theorem 6.1. As a first step, we use the chain rule to decompose the relative entropy term on the right-hand side of (6.2),

$$\begin{aligned}R(\bar{\mu}_i^n(\cdot)||\rho_{\bar{X}_{i-1}^n}(\bar{Y}_{i-1}^n, \cdot)) &= R(\delta_{\bar{Y}_{i-1}^n}(\cdot)||\delta_{\bar{Y}_{i-1}^n}(\cdot)) + R(\bar{\mu}_i^n(\cdot)||\rho_{\bar{X}_{i-1}^n}(\bar{Y}_{i-1}^n, \cdot))\\ &= R(\delta_{\bar{Y}_{i-1}^n}(dy)\bar{\mu}_i^n(dz)||\delta_{\bar{Y}_{i-1}^n}(dy)\rho_{\bar{X}_{i-1}^n}(y, dz))\\ &= R(\lambda^n(dy \times dz|t)||\gamma^n(dy \times dz|t)). \end{aligned} \tag{6.5}$$

By tightness, we can pick a subsequence, also labelled by n for notational convenience, along which all the measures involved converge. Along this subsequence we also have the following lower bound:

$$\begin{aligned}\liminf_{n\to\infty} -\frac{1}{\beta_n} E\left[e^{-\beta_n F(X^n)}\right] + \varepsilon &\geq \liminf_{n\to\infty} E\left[F(\bar{X}^n) + \frac{1}{\beta_n} \sum_{i=n}^{\beta_n+n-1} R(\bar{\mu}_i^n(\cdot)||\rho_{\bar{X}_i^n}(\bar{Y}_i^n, \cdot))\right]\\ &= \liminf_{n\to\infty} E\left[F(\bar{X}^n) + R(\lambda^n(dy \times dz \times dt)||\gamma^n(dy \times dz \times dt))\right] && (6.6)\\ &\geq E\left[F(\bar{X}) + R(\lambda(dx \times dy \times dt)||\gamma(dx \times dy \times dt))\right]. && (6.7)\end{aligned}$$

In the first step in the last display, the equality (6.6), we use the decomposition (6.5) combined with the definition of h^n and the fact that the measures $\lambda^n(\cdot|t), \gamma^n(\cdot|t)$ are constant over the intervals $[\tau_i^n, \tau_{i+1}^n)$. In the second step, the inequality (6.7), we combine Lemma 6.2 with Fatou's lemma and the lower semi-continuity of relative entropy (see, e.g., [9, 4]). Next, we use the chain rule once more combined with the structure of the measures λ and γ,

$$\begin{aligned} &E\left[F(\bar{X}) + R(\lambda(dy \times dz \times dt)||\gamma(dy \times dz \times dt))\right] \\ &\quad = E\left[F(\bar{X}) + \int_0^T \frac{1}{h(t)} R(\lambda(dy \times dz|t)||\bar{\mu}(dy|t)\rho_{\bar{X}(t)}(y, dz|t))dt\right]. \end{aligned}$$

The relative entropy term on the right-hand side can be bounded from below by the local rate function L in (2.3):

$$\begin{aligned} &E\left[F(\bar{X}) + \int_0^T \frac{1}{h(t)} R(\lambda(dy \times dz|t)||\bar{\mu}(dy|t)\rho_{\bar{X}(t)}(y, dz|t))dt\right] \\ &\quad \geq E\left[F(\bar{X}) + \int_0^T \frac{1}{h(t)} L(\bar{X}(t), \dot{\bar{X}}(t))dt\right] \\ &\quad \geq \inf_{\varphi}\{F(\varphi) + \int_0^T \frac{1}{h(t)} L(\varphi(t), \dot{\varphi}(t))dt\}, \end{aligned}$$

where the infiumum is over $\varphi \in AC_{x_0}([0,T] : \mathbb{R}^{d_1})$. The integral on the right-hand side is precisely how the rate function I was defined in Theorem 3.1 and combining the inequalities leads to the

$$\liminf_{n\to\infty} -\frac{1}{\beta_n} E\left[e^{-\beta_n F(\bar{X}^n)}\right] + \varepsilon \geq \inf_{\varphi}\left\{F(\varphi) + I(\varphi)\right\}.$$

Since ε was chosen arbitrarily, this shows how the upper bound (6.1) for the subsequence used. A standard argument by contradiction extends the upper bound to hold for the full sequence, which shows how the Laplace principle upper bound follows from Lemma 6.2. ■

Proof of Lemma 6.2. Because we can always choose the controls $\{\bar{\mu}_i^n\}$ such that the expectation of the sum of the relative entropies, appearing in (3.2), tightness of $\{\bar{X}^n\}$ and $\{\bar{\mu}^n\}$ follows from Theorem 7.8, which also gives the characterisation of the limit points as in (6.4)-(6.3). From the definition of the controlled process, tightness of $\{\bar{\mu}^n\}_n$ implies tightness of $\{\delta_{\bar{Y}_i^n}\}_{i=n}^{\beta_n+n}$, as a sequence in n. This in turn gives tightness of $\{\lambda^n\}$. The tightness of $\{\gamma^n\}$ is obtained by the tightness of $\{\bar{X}^n\}$ and $\{\delta_{\bar{Y}_i^n}\}_{i=n}^{\beta_n+n}$ together with the uniform continuity of $\rho_x(y, dz)$.

To characterise limit points λ of subsequences of $\{\lambda^n\}$, by Lemma 3.3.1 in [9] and the uniform convergence of h^n we have the decomposition $\lambda(dy \times dz \times dt) = (h(t))^{-1}\lambda(dy \times dz|t)dt$, for some stochastic kernel $\lambda(dy \times dz|t)$. Moreover, note that $\lambda^n(\mathbb{R}^{d_2} \times A \times C) = \bar{\mu}^n(A \times C)$ implies that $\lambda(\mathbb{R}^{d_2} \times A \times C) = \bar{\mu}(A \times C)$. For the marginal obtained when integrating out the second variable, we use arguments similar to those used in proving Lemma 6.12 in [4]. Take $\{f_m\}$ as a countable collection of bounded continuous functions that is also a separating class on $\mathbb{R}^{d_2}$. We will prove that, for any $\varepsilon > 0$ and all $t \in [0,T]$, as $n \to \infty$,

$$P\left(\left|\left|\int_0^t \int \frac{1}{h^n(s)} f_m(y)\bar{\mu}^n(dy|s)ds - \int_0^t \int \frac{1}{h^n(s)} f_m(y)\lambda^n(dy \times \mathbb{R}^{d_2}|s)ds\right|\right| > \varepsilon\right) \to 0. \qquad (6.8)$$

Suppose this limit holds. Because the collection of sets of the form $[0,t]$, for $t \in [0,T]$, is a separating class of $[0,T]$, (6.8) combined with Fatou's lemma ensures that w.p. 1 the limit of λ^n will satisfy $\lambda(A \times \mathbb{R}^{d_2} \times C) = \bar{\mu}(A \times C)$.

To prove (6.8), define $K_m = \|f_m\|_\infty$. Suppose n is such that $\beta_n > 4K_m/\varepsilon$–since $\beta_n \to \infty$ as $n \to \infty$, this is possible. Using the definitions of $\bar{\mu}^n$ and λ^n, and an application of Markov's inequality we have

$$\begin{aligned}
&P\left(\left\|\int_0^t \int \frac{1}{h^n(s)} f_m(y)\bar{\mu}^n(dy|s)ds - \int_0^t \int \frac{1}{h^n(s)} f_m(y)\lambda^n(dy \times \mathbb{R}^{d_2}|s)ds\right\| > \varepsilon\right) \\
&= P\left(\left\|\frac{1}{\beta_n}\sum_{i=n+1}^{\mathbf{m}(t_n+t)} \int f_m(y)\bar{\mu}_i^n(dy) - \frac{1}{\beta_n}\sum_{i=n}^{\mathbf{m}(t_n+t)-1} f_m(\bar{Y}_i^n)\right\| > \varepsilon\right) \\
&\leq P\left(\left\|\frac{1}{\beta_n}\sum_{i=n+1}^{\mathbf{m}(t_n+t)} \int f_m(y)\bar{\mu}_i^n(dy) - \frac{1}{\beta_n}\sum_{i=n+1}^{\mathbf{m}(t_n+t)} f_m(\bar{Y}_i^n)\right\| > \frac{\varepsilon}{2}\right) \\
&\qquad + P\left(\left\|\frac{1}{\beta_n}\left(f_m(\bar{Y}^n_{\mathbf{m}(t_n+t)}) - f_m(\bar{Y}_n^n)\right)\right\| > \frac{\varepsilon}{2}\right) \\
&\leq P\left(\left\|\frac{1}{\beta_n}\sum_{i=n+1}^{\mathbf{m}(t_n+t)} \left(\int f_m(y)\bar{\mu}_i^n(dy) - f_m(\bar{Y}_i^n)\right)\right\| > \frac{\varepsilon}{2}\right) \\
&\leq \frac{4}{\varepsilon^2} E\left[\frac{1}{\beta_n^2}\sum_{i,j=n+1}^{\mathbf{m}(t_n+t)} \Delta_{m,i}^n \Delta_{m,j}^n\right],
\end{aligned}$$

where we have defined

$$\Delta_{m,i}^n = \int f_m(y)\bar{\mu}_i^n(dy) - f_m(\bar{Y}_i^n).$$

The term $P\left(\left\|\frac{1}{\beta_n}\left(f_m(\bar{Y}^n_{\mathbf{m}(t_n+t)}) - f_m(\bar{Y}_n^n)\right)\right\| > \frac{\varepsilon}{2}\right) = 0$ since $\frac{1}{\beta_n}\left(f_m(\bar{Y}^n_{\mathbf{m}(t_n+t)}) - f_m(\bar{Y}_n^n)\right) < \frac{\varepsilon}{4K_m} 2\|f_m\| = \frac{\varepsilon}{2}$. The sequence $\{\Delta_{m,i}^n\}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_i^n = \sigma\left((\bar{X}_j^n, \bar{Y}_j^n) : \ j < i\right)$. Therefore, the off-diagonal terms in the sum have expected value 0: for $i > j$,

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

Combined with the previous inequalities this leads to the upper bound

$$\begin{aligned}
&P\left(\left\|\int_0^t \int \frac{1}{h^n(s)} f_m(y)\bar{\mu}^n(dy|s)ds - \int_0^t \int \frac{1}{h^n(s)} f_m(y)\lambda^n(dy \times \mathbb{R}^{d_2}|s)ds\right\| > \varepsilon\right) \\
&\leq \frac{4}{\varepsilon^2} E\left[\frac{1}{\beta_n^2}\sum_{i=n+1}^{\mathbf{m}(t_n+t)} \left(\Delta_{m,i}^n\right)^2\right] \\
&\leq \frac{4}{\varepsilon^2} E\left[\frac{1}{\beta_n^2}\sum_{i=n+1}^{\beta_n+n} \left(2K_m\right)^2\right] \\
&\leq \frac{16K_m^2}{\varepsilon^2 \beta_n}.
\end{aligned}$$

We can make this arbitrarily small by choosing n large enough, which proves (6.8).

In order to show the claimed form for γ we use a strategy similar to the one used for λ. Take $\{f_m\}$ to now be a countable separating class on $\mathbb{R}^{d_2} \times \mathbb{R}^{d_2}$ of bounded continuous functions. We define a sequence of measures $\{\eta^n\}$ by

$$\eta^n(A \times B \times C) = \int_C \frac{1}{h(t)} \eta^n(A \times B|t)dt, \quad \eta^n(A \times B|t) = \int_A \rho_{\bar{X}^n_{i-1}}(y, B)\bar{\mu}^n_{i-1}(dy).$$

From the convergence of $\bar{\mu}^n$ and the continuity of ρ, η^n converges to γ. To finish the proof we therefore show that γ^n must have the same limit as η^n, by proving that, for arbitrary $\varepsilon > 0$,

$$P\left(\left\|\int_0^t \int\int \frac{1}{h^n(s)} f_m(y,z)\eta^n(dy \times dz|s)ds - \int_0^t \int\int \frac{1}{h^n(s)} f_m(y,z)\gamma^n(dy \times dz|s)ds\right\| > \varepsilon\right) \to 0.$$

Similar to before, take $K_m = \|f_m\|_\infty$. Then,

$$\begin{aligned}
&P\left(\left\|\int_0^t \int\int \frac{1}{h^n(s)} f_m(y,z)\eta^n(dy \times dz|s)ds - \int_0^t \int\int \frac{1}{h^n(s)} f_m(y,z)\gamma^n(dy \times dz|s)ds\right\| > \varepsilon\right) \\
&\quad = P\left(\left\|\frac{1}{\beta_n}\sum_{i=n+1}^{\mathbf{m}(t_n+t)} \int\int f_m(y,z)\rho_{\bar{X}^n_{i-1}}(y,dz)\bar{\mu}^n_{i-1}(dy) - \frac{1}{\beta_n}\sum_{i=n+1}^{\mathbf{m}(t_n+t)} \int f_m(\bar{Y}^n_{i-1}, z)\rho_{\bar{X}^n_{i-1}}(\bar{Y}^n_{i-1}, dz)\right\| > \varepsilon\right) \\
&\quad = P\left(\left\|\frac{1}{\beta_n}\sum_{i=n+1}^{\mathbf{m}(t_n+t)} \left(\int\int f_m(y,z)\rho_{\bar{X}^n_{i-1}}(y,dz)\bar{\mu}^n_{i-1}(dy) - \int f_m(\bar{Y}^n_{i-1}, z)\rho_{\bar{X}^n_{i-1}}(\bar{Y}^n_{i-1}, dz)\right)\right\| > \varepsilon\right) \\
&\quad \le \frac{1}{\varepsilon^2} E\left[\frac{1}{\beta_n^2}\sum_{i,j=n}^{\mathbf{m}(t_n+t)-1} \tilde{\Delta}^n_{m,i}\tilde{\Delta}^n_{m,j}\right],
\end{aligned}$$

where

$$\tilde{\Delta}^n_{m,i} = \int\int f_m(y,z)\rho_{\bar{X}^n_i}(y,dz)\bar{\mu}^n_i(dy) - \int f_m(\bar{Y}^n_i, z)\rho_{\bar{X}^n_i}(\bar{Y}^n_i, dz).$$

Similar to the convergence analysis for λ^n, $\{\tilde{\Delta}^n_{m,i}\}$ forms a martingale difference sequence with respect to the filtration $\mathcal{F}^n_j$. The off-diagonal terms thus disappear from the sum,

$$E\left[\frac{1}{\beta_n^2}\sum_{i,j=n}^{\mathbf{m}(t_n+t)-1} \tilde{\Delta}^n_{m,i}\tilde{\Delta}^n_{m,j}\right] = E\left[\frac{1}{\beta_n^2}\sum_{i=n}^{\mathbf{m}(t_n+t)-1} \left(\tilde{\Delta}^n_{m,i}\right)^2\right],$$

and we obtain the upper bound

$$\begin{aligned}
&P\left(\left\|\int_0^t\int\int\frac{1}{h^n(s)}f_m(y,z)\bar{\eta}^n(dy\times dz|s)ds-\int_0^t\int\int\frac{1}{h^n(s)}f_m(y,z)\gamma^n(dy\times dz|s)ds\right\|>\varepsilon\right)\\
&\quad\leq\frac{1}{\varepsilon^2}E\left[\frac{1}{\beta_n^2}\sum_{i=n}^{\mathbf{m}(t_n+t)-1}\left(\tilde{\Delta}_{m,i}^n\right)^2\right]\\
&\quad\leq\frac{1}{\varepsilon^2}E\left[\frac{1}{\beta_n^2}\sum_{i=n}^{\beta_n+n-1}(2K_m)^2\right]\\
&\quad=\frac{4K_m^2}{\varepsilon^2\beta_n}.
\end{aligned}$$

We can choose n large enough to make this as small as desired. Since ε was taking arbitrarily, this proves the claimed convergence. Having already established that $\eta^n\to\gamma$, this also shows that $\gamma^n\to\gamma$. ■

7 Laplace lower bound

In this section we prove the Laplace principle lower bound, which amounts to the following.

Theorem 7.1 *Assume (A.1)-(A.8). With I defined as in* (3.1)*, for any bounded, continuous function $F:C([0,T]:\mathbb{R}^{d_1})\to\mathbb{R}$,*

$$\limsup_{n\to\infty}-\frac{1}{\beta_n}\log E\left[e^{-\beta_nF(X^n)}\right]\leq\inf_{\varphi}\left(F(\varphi)+I(\varphi)\right), \tag{7.1}$$

where the infimum is over $\varphi\in AC_{x_0}([0,T]:\mathbb{R}^{d_1})$.

Together with the upper bound (6.1), this proves the limit in Theorem 3.1. The proof of the upper bound, given in Section 6, is aided by the fact that by definition of the infimum, we can choose a sequence of controls satisfying (6.2). Proving the lower bound (7.1) is considerably more involved as we must now explicitly construct a sequence of nearly optimal controls.

7.1 Construction and tightness of nearly-optimal controls

In this section we construct, for each n, a sequence of nearly-optimal controls to be used in proving the lower bound. As a first step, we show that the local rate function L defined in (2.3), is continuous (Lemma 3.5), and that for any (x,β) such that $L(x,\beta)<\infty$, there exists nearly-optimal transition kernels with respect to the infimum in the definition of $L(x,\beta)$ (Lemma 7.2). Next, in Lemma 7.3 we show that for any function $\zeta\in C([0,1]:\mathbb{R}^d)$ such that $I(\zeta)<\infty$, for any $\varepsilon>0$ we can find a piece-wise linear function, with a finite number of pieces, that is ε-close to ζ both in sup-norm and in evaluating I.

Recall that π_x is the unique invariant measure of ρ_x. The following result is a direct consequence of the definition of L and results in [4].

Lemma 7.2 *Suppose (A.2), (A.5) and (A.6) hold. For any $(x,\beta) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_1}$ such that $L(x,\beta) < \infty$ and $\varepsilon > 0$, there exists a probability measure $\nu^{x,\beta}(dy)$ such that*

$$\inf_{\gamma \in \mathcal{A}(\nu^{x,\beta})} R(\gamma \| \nu^{x,\beta} \otimes \rho_x(\cdot,\cdot)) \leq L(x,\beta) + \varepsilon \text{ and } \beta = \int g(x,y)\nu^{x,\beta}(dy).$$

Moreover, for any $\delta > 0$, define a probability measure $\mu^{x,\beta,\delta} \doteq (1-\delta/2)\nu^{x,\beta} + (\delta/2)\pi_x$. There exists a transition kernel $q^{x,\beta,\delta}(y,dz)$ such that $\mu^{x,\beta,\delta}$ is the unique invariant measure of $q^{x,\beta,\delta}(y,dz)$ and the associated Markov chain is ergodic. In addition,

$$R(\mu^{x,\beta,\delta} \otimes q^{x,\beta,\delta}(\cdot,\cdot) \| \mu^{x,\beta,\delta} \otimes \rho_x(\cdot,\cdot)) \leq \inf_{\gamma \in \mathcal{A}[\nu^{x,\beta}]} R(\gamma \| \nu^{x,\beta} \otimes \rho_x) \leq L(x,\beta) + \varepsilon.$$

Proof. Under (A.1)-(A.7), the existence of $\nu^{x,\beta}$ follows from the definition of $L(x,\beta)$ in terms of an infimum. The existence of $\mu^{x,\beta,\delta}$ and $q^{x,\beta,\delta}$ follow from Lemma 6.17 in [4]. ■

In proving the lower bound Theorem 7.1, we may assume $\inf_\varphi\{F(\varphi)+I(\varphi)\} < \infty$, as otherwise the bound is trivially true. By the definition of the infimum, for any $\varepsilon > 0$, there is $\zeta \in C([0,1];\mathbb{R}^{d_1})$ such that

$$F(\zeta) + I(\zeta) \leq \inf_\varphi\{F(\varphi) + I(\varphi)\} + \varepsilon.$$

Recall that F is bounded and I is of the form

$$I(\zeta) = \int_0^T \frac{1}{h(t)} L(\zeta(t), \dot{\zeta}(t))dt.$$

We can therefore assume that $L(\zeta(t), \dot{\zeta}(t)) < \infty$ for all $t \in [0,T]$. Moreover, the following lemma states that we can focus on ζ that are piece-wise linear with finitely many pieces.

Lemma 7.3 *Assume (A.2),(A.5),(A.6) and (A.7). For $\zeta \in C([0,1];\mathbb{R}^{d_1})$ satisfying $I(\zeta) < \infty$, for any $\varepsilon > 0$, there exists a $\zeta^* \in C([0,1] : \mathbb{R}^{d_1})$ that is piece-wise linear with finitely many pieces, such that $\|\zeta^* - \zeta\|_\infty < \varepsilon$ and*

$$I(\zeta^*) = \int_0^T \frac{1}{h(t)} L(\zeta^*(t), \dot{\zeta}^*(t))dt \leq \int_0^T \frac{1}{h(t)} L(\zeta(t), \dot{\zeta}(t))dt + \varepsilon = I(\zeta) + \varepsilon.$$

Proof. The proof relies on parts of several different results from [4]. First, since $(x,\beta) \mapsto L(x,\beta)$ is continuous by Lemma 3.5, it suffices—see the argument used for Part (e) of Lemma 4.21 in [4]—to show that, for the given $\varepsilon > 0$, there is a $\zeta_1^* \in C([0,T] : \mathbb{R}^{d_1})$ such that $\{\dot{\zeta}_1^*(t) : t \in [0,T]$ is bounded, $\|\zeta - \zeta_1^*\|_\infty < \varepsilon$, and

$$I(\zeta^*) \leq I(\zeta) + \varepsilon.$$

The existence of such an ζ is the topic of Lemma 4.17 in [4]. The same arguments as used in the proof of that result applies also in the setting considered here, if we can show that L is uniformly superlinear in β (see Section 2.1). Recall that H is the Lengendre-Fenchel transform of L. The uniform superlinearity of L then holds if,

$$\sup_{x \in \mathbb{R}^{d_1}} \sup_{\alpha \in \mathbb{R}^{d_1} : \|\alpha\| = M} H(x,\alpha) < \infty, \tag{7.2}$$

for every $M < \infty$; see [4, Lemma 4.14(c)] for why this bound ensures the superlinearity of L. Combining these arguments, to prove the existence of ζ_1^* with the properties described above, it is enough to prove (7.2).

To show (7.2), we recall the alternative representation from Proposition 3.3,

$$H(x,\alpha) \doteq \lim_{n\to\infty} \frac{1}{n} \log \left(\int \cdots \int e^{\langle \alpha, g(x,y_1)\rangle + \cdots + \langle \alpha, g(x,y_n)\rangle} \rho_x(y_0, dy_1) \cdots \rho_x(y_{n-1}, dy_n) \right).$$

Moreover, Assumption (A.7) ensures that, for every $\alpha \in \mathbb{R}^{d_1}$,

$$\hat{C}_\alpha = \sup_x \sup_y \left(\log \int_{\mathbb{R}^{d_2}} e^{\langle \alpha, g(x,z)\rangle} \rho_x(y, dz) \right) < \infty.$$

Combining the two, we have that, for any $\alpha \in \mathbb{R}^{d_1}$,

$$H(x,\alpha) \le \hat{C}_\alpha < \infty.$$

In addition, for any x, y, the function

$$\alpha \mapsto \log \int_{\mathbb{R}^{d_2}} e^{\langle \alpha, g(x,z)\rangle} \rho_x(y, dz)$$

is convex. Because the supremum of a collection of convex functions is also convex, it holds that $\alpha \mapsto \hat{C}_\alpha$ is a convex function, with finite values for all $\alpha \in \mathbb{R}^{d_1}$. Therefore, $\hat{C}_\alpha$ is continuous in α, due to it being convex and finite-valued for any α, and we have

$$\sup_{x\in\mathbb{R}^{d_1}} \sup_{\alpha\in\mathbb{R}^{d_1}:\|\alpha\|=M} H(x,\alpha) \le \sup_{\alpha\in\mathbb{R}^{d_1}:\|\alpha\|=M} \log(\hat{C}_\alpha) < \infty,$$

for every $M < \infty$.

This shows (7.2), which ensures the uniform superlinearity of L, and in turn the existence of an $\zeta_1^* \in C([0,T] : \mathbb{R}^{d_1})$ such that $\{\dot{\zeta}_1^*(t) : t \in [0,T]\}$ is bounded, $\|\zeta - \zeta_1^*\| < \varepsilon$, and $(\zeta_1^*) \le I(\zeta) + \varepsilon$. Using the continuity of L, a function ζ^* with the claimed properties can then be obtained as a piece-wise linear approximation of ζ_1^*. ■

With Lemmas 3.5, 7.2 and 7.3 established, we are now ready to construct the (nearly-optimal) controls that will play a central role in the proof of the lower bound Theorem 7.1. A crucial part of the construction of the controls is to split up the interval $\{n, n+1, \ldots, n+\beta_n\}$ into ℓ segments. Let $\ell, m \in \mathbb{N}$ where for any $\ell \le \beta_n$, m is the largest integer such that $\ell m \le \beta_n$. The idea is that, for fixed ℓ, we can freeze the state dependence of the noise sequence and therefore be able to use an ergodic argument in the convergence. Furthermore define the times $\tau_k^\ell, k = 0, 1, \ldots, \ell$ as

$$\tau_k^\ell = \lim_{n\to\infty} \sum_{i=n}^{\lfloor n+\frac{k}{l}\beta_n \rfloor} \varepsilon_i,$$

the corresponding times for the ℓ intervals. From the definition we have $\tau_0^\ell = 0$ and $\tau_\ell^\ell = T$.

The controls will be defined in terms of the transition probabilities obtained in Lemma 7.2. Set $\bar{X}_n^n = x_0$, $\bar{Y}_n^n = y_0$, and recall that $\zeta(0) = x_0$. Given $\delta > 0$, for $j = n, \dots, n+m-1$, we define $\hat{\nu}_j^n$ as

$$\hat{\nu}_j^n(dz) = \begin{cases} \rho_{\zeta^*(0)}(\bar{Y}_{j-1}^n, dz) & j < n + l_0, \\ q^{\zeta^*(0), \dot{\zeta}^*(0), \delta}(\bar{Y}_{j-1}^n, dz) & j \geq n + l_0, \end{cases}$$

where l_0 is the constant in the transitivity condition (A.6). The $\hat{\nu}_j^n$s define a controlled sequence $\{\bar{Y}_j^n\}_j$ in that the conditional distribution of $\bar{Y}_j^n$ given $\mathcal{F}_{j-1}^n$ is $\hat{\nu}_j^n$. These controlled measures are such that for the first l_0 variables $\bar{Y}_n^n, \dots, \bar{Y}_{n+l_0-1}^n$, the conditional distribution is the same as the noise distribution with fixed x-argument, and for the remaining variables, $\bar{Y}_{n+l_0}^n, \dots, \bar{Y}_{n+m-1}^n$, the conditional distribution is the transition kernel of Lemma 7.2 associated with the triplet $(\zeta(0), \dot{\zeta}(0), \delta)$.

Next, with m and l fixed according to the above, for each $k \in \mathbb{N}$, $1 \leq k \leq l$, for $j = n + km + 1, \dots, n + km + m$, we define

$$\hat{\nu}_j^n(dz) = \begin{cases} \rho_{\zeta^*(\tau_k^\ell)}(\bar{Y}_{j-1}^n, dz) & j < n + km + l_0 + 1, \\ q^{\zeta^*(\tau_k^\ell), \dot{\zeta}^*(\tau_k^\ell), \delta}(\bar{Y}_{j-1}^n, dz), & j \geq n + km + l_0 + 1, \end{cases}$$

where the conditional distribution of each $\bar{Y}_j^n$ given $\mathcal{F}_{j-1}^n$ is again $\hat{\nu}_j^n$, and we use the notation $t_j^n = t_j - t_n$. Lastly, for $j = n + \ell m + 1, \dots, n + \beta_n$, we set

$$\hat{\nu}_j^n(dz) = \rho_{\bar{X}_{j-1}^n}(\bar{Y}_{j-1}^n, dz),$$

where the controlled process $\bar{X}^n$ is defined as

$$\bar{X}_j^n = \bar{X}_{j-1}^n + \varepsilon_j g(\bar{X}_{j-1}^n, \bar{Y}_j^n), \quad j = n, \dots, n + \beta_n.$$

To make sure that the controlled process $\bar{X}_i^n$ is not too far away from path $\zeta^*(t)$ we define the stopping index $\hat{i}^n$ as

$$\hat{i}^n = \inf\left\{ i \leq n : \|\bar{X}_i^n - \zeta^*(t_i^n)\| > 1 \right\} \wedge (\beta_n + n),$$

and the stopping time $\hat{S}^n$ as

$$\hat{S}^n = \sum_{i=n}^{\hat{i}^n} \varepsilon_i.$$

Observe that since $\bar{X}_n^n = \zeta(0) = x$, we have that $\hat{i}^n > n$ and $\hat{S}^n > 0$. Now we define the controls $\bar{\nu}^n$ as

$$\bar{\nu}_j^n = \begin{cases} \hat{\nu}_j^n(dz) & j < \hat{i}^n, \\ \rho_{\bar{X}_{j-1}^n}(\bar{Y}_{j-1}^n, dz). \end{cases}$$

This defines the controls, that is the conditional distributions $\{\bar{\nu}_j^n\}_j$ for the noise, and the corresponding controlled process $\bar{X}^n = \{\bar{X}_j^n\}_j$. To have a control in continuous time instead, we define $\bar{\nu}^n(A|t) = \bar{\nu}_i^n(A)$ for $t \in [t_{i-1} - t_i, t_i - t_n]$ and the measure $\bar{\nu}^n \in \mathcal{P}(\mathbb{R}^{d_2} \times [0, T])$ by

$$\bar{\nu}^n(A \times B) = \int_B \frac{1}{h^n(t)} \bar{\nu}^n(A|t) dt.$$

Note that throughout the paper, unless where there is a need to emphasise it, we suppress the dependence on δ in the control sequence $\{\bar{\nu}_j^n\}$ in the notation.

As a step towards proving the lower bound, we prove tightness of the control sequence $\{\bar{\nu}^n\}_n$.

Lemma 7.4 *Under (A.1)-(A.7), the control sequence $\{\bar{\nu}^n\}_n$ is tight.*

Proof. The proof will be the same as Lemma 4.11 in [4] and proposition 5.3.2 in [10]. We require that

$$\sup_n E\left[\frac{1}{\beta_n}\sum_{i=n}^{\beta_n+n-1} R(\bar{\nu}_{i+1}^n(\cdot)||\rho_{\bar{X}_{i+1}^n}(\bar{Y}_i^n,\cdot))\right] < \infty,$$

which we prove in Lemma 7.7. It is sufficient to prove that $\bar{\nu}^n$ satisfies the uniform integrability property

$$\lim_{C\to\infty}\limsup_n E\left[\int_0^T\int_{\mathbb{R}^{d_2}}\int_{\|z\|>C}\|z\|\bar{\nu}^n(dy\times dt)\right] = 0.$$

The proof uses the inequality $ab \leq e^{\sigma a} + \frac{1}{\sigma}(b\log(b) - b + 1)$ with $a = \|\zeta\|$ and $b = \frac{d\bar{\nu}_i^n(\cdot)}{d\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,\cdot)}$. For $t \in [0,T]$, and fixed C and n, we have,

$$\begin{aligned}
&\int_{\|z\|>C}\|z\|d\bar{\nu}_i^n(dz)\\
&\quad = \int_{\|z\|>C}\|z\|\frac{d\bar{\nu}_i^n(z)}{d\rho_{\bar{X}_i^n}(\bar{Y}_i^n,z)}\rho_{\bar{X}_i^n}(\bar{Y}_i^n,dz)\\
&\quad \leq \int_{\|z\|>C}e^{\sigma\|z\|}\rho_{\bar{X}_i^n}(\bar{Y}_i^n,dz)\\
&\qquad + \frac{1}{\sigma}\int_{\|z\|>C}\left(\frac{d\nu_i^n(z)}{d\rho_{\bar{X}_i^n}(\bar{Y}_i^n,z)}\log\left(\frac{d\bar{\nu}_i^n(z)}{d\rho_{\bar{X}_i^n}(\bar{Y}_i^n,z)}\right) - \frac{d\bar{\nu}_i^n(z)}{d\rho_{\bar{X}_i^n}(\bar{Y}_i^n,z)} + 1\right)\rho_{\bar{X}_i^n}(\bar{Y}_i^n,dz)\\
&\quad \leq \int_{\|z\|>C}e^{\sigma\|z\|}\rho_{\bar{X}_i^n}(\bar{Y}_i^n,dz) + \frac{1}{\sigma}R(\bar{\nu}_i^n(\cdot)||\rho_{\bar{X}_i^n}(\bar{Y}_i^n,\cdot))\\
&\quad \leq e^{-\sigma C}\sup_x\sup_y\int e^{2\sigma\|z\|}\rho_x(y,dz) + \frac{1}{\sigma}R(\bar{\nu}_i^n(\cdot)||\rho_{\bar{X}_i^n}(\bar{Y}_i^n,\cdot)),
\end{aligned}$$

where in the last step we have used Assumption (A.7) to guarantee that the first term is finite. Moreover, (7.5) ensures that the second term is bounded in n. Using this bound yields

$$\begin{aligned}
&E\left[\int_0^T\int_{\mathbb{R}^{d_2}}\int_{\|z\|>C}\|z\|\nu^n(dy\times dt)\right]\\
&\leq E\left[\sum_{i=n}^{\beta_n+n-1}\int_{t_i,t_{i+1}}\frac{1}{h_n(t)}e^{-\sigma C}\sup_x\sup_y\int e^{2\sigma\|z\|}\rho_x(y,dz) + \frac{1}{h_n(t)}\frac{1}{\sigma}R(\tilde{\nu}_i^n(\cdot)||\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,\cdot))dt\right]\\
&=\leq e^{-\sigma C}\sup_x\sup_y\int e^{2\sigma\|z\|}\rho_x(y,dz) + \frac{1}{\sigma}E\left[\frac{1}{\beta_n}\sum_{i=n}^{n+\beta_n-1}R(\tilde{\nu}_i^n(\cdot)||\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,\cdot))\right].
\end{aligned}$$

The first term does not depend on n and the second is by Lemma 7.7 bounded. Now sending $C \to \infty$ and then $\sigma \to \infty$ yields the uniform integrability and also the tightness. ∎

7.2 Convergence of controls and controlled processes

A key step in the weak convergence approach is to show convergence of the controls and associated controlled processes, and to identify the limit objects and their properties. In this section we carry out such an analysis for the pairs $(\bar{\nu}^n, \bar{X}^n)$.

Take $\varepsilon > 0$ and for the function ζ^* from Lemma 7.3, consider the associated measures $\{\nu^{\zeta^*(t),\dot{\zeta}^*(t)} : t \in [0,T]\}$ from Lemma 7.2; throughout the section we suppress the dependence on ε. The following theorem is the main result of the section.

Theorem 7.5 *Fix $\varepsilon > 0$ and ζ^* according to the above. Under (A.1)-(A.8), for every subsequence of $\{(\bar{\nu}^n, \bar{X}^n)\}$, there exists a further subsequence that converges weakly to $(\bar{\nu}, \zeta^*)$, where $\bar{\nu}$ satisfies*

$$\bar{\nu}(A \times B) = \int_B \frac{1}{h(t)} \bar{\nu}(A|t)dt,$$

and $\bar{\nu}(A|t) = \nu^{\zeta^*(t),\dot{\zeta}^*(t)}$.

The proof relies on showing that the limit $\bar{X}$ of $\bar{X}^n$ satisfies

$$\bar{X}(t) = x + \int_0^t \int_{\mathbb{R}^{d_2}} g(\bar{X}(s), y)\bar{\nu}(dy|s)ds,$$

and that, by construction of the $\nu^{\zeta^*(t),\dot{\zeta}^*(t)}$-measures, ζ^* satisfies the ODE

$$\zeta^*(t) = \int_0^t \int g(\zeta^*(s), y)\nu^{\zeta^*(s),\dot{\zeta}^*(s)}(dy)ds.$$

That this ODE has a unique solution is shown in Lemma A.2 in the Appendix.

We begin with an ancillary result, which will be used to prove tightness of the controlled processes $\bar{X}^n$ for generic controlled measures with bounded relative entropy with respect to ρ along the controlled process (Theorem 7.8). The proof of the following result is the same as for Lemma 6.16(b) in [4]; we omit the details.

Lemma 7.6 *Let l_0 be the constant in the transitivity condition (A.6). If a Borel set A has the property that $\rho_x^{l_0}(y, A) > 0$ for some x, y, then $\pi_x(A) > 0$.*

Using Lemma 7.6, we now prove that the expected running cost associated with the controlled measures $\{\bar{\nu}^n\}_n$ is bounded.

Lemma 7.7 *Under (A.2), (A.4), (A.5), (A.6) and (A.7), with $\bar{\nu}^n = \{\bar{\nu}_j^n\}_{j=n+1}^{\beta_n+n}$ defined as in Section 7.1,*

$$\sup_n E\left[\frac{1}{\beta_n} \sum_{i=n}^{\beta_n+n-1} R(\bar{\nu}_{i+1}^n(\cdot)||\rho_{\bar{X}_{i+1}^n}(\bar{Y}_i^n, \cdot))\right] < \infty.$$

Proof. First observe that for $i \geq \hat{i}^n$ we have that

$$R(\bar{\nu}^n_{i+1}(\cdot)||\rho_{\bar{X}^n_i}(\bar{Y}^n_i,\cdot)) = 0.$$

The proof below is carried out for indices $j < \hat{i}^n$. For each n, with ℓ and m as in Section 7.1, from the definition of the $\bar{\nu}^n_j$s we have

$$\frac{1}{\beta_n}\sum_{i=n}^{\beta_n+n-1} R\left(\bar{\nu}^n_{i+1}(\cdot)\|\rho_{\bar{X}^n_{i+1}}(\bar{Y}^n_i,\cdot)\right) = \frac{1}{\beta_n}\sum_{k=0}^{\ell-1}\sum_{j=1}^{m} R\left(\bar{\nu}^n_{n+km+j}(\cdot)\|\rho_{\bar{X}^n_{n+km+j-1}}(\bar{Y}^n_{n+km+j-1},\cdot)\right).$$

Using the definition of relative entropy, for each k and j in the relevant ranges, we can re-write the relative entropy-term on the right-hand side of the last display as

$$\begin{aligned} &R\left(\bar{\nu}^n_{n+km+j}(\cdot)\|\rho_{\bar{X}^n_{n+km+j}}(\bar{Y}^n_{n+km+j-1},\cdot)\right) \\ &\quad= R\left(\bar{\nu}^n_{n+km+j}(\cdot)\|\rho_{\zeta(\tau^\ell_k)}(\bar{Y}^n_{km+j-1},\cdot)\right) + \int_{\mathbb{R}^{d_2}}\left(\log\frac{d\rho_{\zeta(\tau^\ell_k)}(\bar{Y}^n_{n+km+j-1},y)}{d\rho_{\bar{X}^n_{km+j}}(\bar{Y}^n_{n+km+j-1},y)}\right)\bar{\nu}^n_{n+km+j}(dy). \end{aligned} \tag{7.3}$$

Note also that $R\left(\bar{\nu}^n_{n+km+j-1}(\cdot)\|\rho_{\zeta(\tau^\ell_k)}(\bar{Y}^n_{km+j-1},\cdot)\right) = 0$ for $j \leq l_0$.

Take $\delta > 0$. For any $k \in \{1,\dots,\ell\}$, consider the integral

$$\int_{\mathbb{R}^{d_2}} R\left(q^{\zeta(\tau^\ell_k),\dot{\zeta}(\tau^\ell_k),\delta}(y,\cdot)\|\rho_{\zeta(\tau^\ell_k}(y,\cdot)\right)\mu^{\zeta(\tau^\ell_k),\dot{\zeta}(\tau^\ell_k),\delta}(dy).$$

We will show that, as $m \to \infty$, which corresponds to the limit $n \to \infty$, this integral approximates

$$\frac{1}{m}\sum_{j=l_0+1}^{m} R\left(q^{\zeta(\tau^\ell_k),\dot{\zeta}(\tau^\ell_k),\delta}(\bar{Y}^n_{km+j-1},\cdot)\|\rho_{\zeta(\tau^\ell_k)}(\bar{Y}^n_{km+j-1},\cdot)\right),$$

that is the first part in the alternative representation of the running cost (7.3). To show this, we use arguments similar to those used in the proof of Proposition 6.15 in [4]. First, from Lemma 7.2,

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{m}\sum_{j=l_0+1}^{m} R\left(q^{\zeta(\tau^\ell_k),\dot{\zeta}(\tau^\ell_k),\delta}(\bar{Y}^n_{km+j-1},\cdot)\|\rho_{\zeta(\tau^\ell_k)}(\bar{Y}^n_{km+j-1},\cdot)\right)\right] \\ &\quad= \int_{\mathbb{R}^{d_2}} R\left(q^{\zeta(\tau^\ell_k),\dot{\zeta}(\tau^\ell_k),\delta}(y,\cdot)\|\rho_{\zeta(\tau^\ell_k}(y,\cdot)\right)\mu^{\zeta(\tau^\ell_k),\dot{\zeta}(\tau^\ell_k),\delta}(dy) \\ &\quad\leq L(\zeta^*(t^n_{n+km}),\dot{\zeta}^*(t^n_{n+km})) + \varepsilon, \end{aligned}$$

and from the properties of ζ^* this is finite. From this, the non-negativity of the relative entropy, and the properties of the $\mu^{\zeta^*(\tau^\ell_k),\dot{\zeta}^*(\tau^\ell_k),\delta}$-measures, and associated Markov chains, the L^1-ergodic theorem implies the convergence

$$\lim_{m\to\infty} E\left[\left|\frac{1}{m}\sum_{j=l_0+1}^{m} R\left(q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(\bar{Y}^n_{km+j-1},\cdot)\|\rho_{\zeta^*(\tau_k^\ell)}(\bar{Y}^n_{km+j-1},\cdot)\right)\right.\right.$$
$$\left.\left.-\int R\left(q^{\zeta^*(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(y,\cdot)\|\rho_{\zeta^*(\tau_k^\ell)}(y,\cdot)\right)\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy)\right|\right]=0.$$

From this convergence, it follows that, for any y_k,

$$E_{y_k}\left[\left|\frac{1}{m}\sum_{j=l_0+1}^{m} R\left(q^{\zeta^*(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(\bar{Y}^n_{km+j-1},\cdot)\|\rho_{\zeta^*(\tau_k^\ell)}(\bar{Y}^n_{km+j-1},\cdot)\right)\right.\right.$$
$$\left.\left.-\int R\left(q^{\zeta^*(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(y,\cdot)\|\rho_{\zeta^*(\tau_k^\ell)}(y,\cdot)\right)\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy)\right|\right],$$

converges in probability to 0, as $m\to\infty$. This is turn ensures that, for any $k\in\{1,\dots,\ell\}$, there is a further subsequence of $\{m\}$—we abuse notation and denote this subsequence by $\{m\}$ as well—and a Borel set Φ_k such that $\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(\Phi_k)=1$, and for any $\bar{Y}^n_{n+km+l_0}=y_k\in\Phi_k$,

$$\lim_{m\to\infty} E_{y_k}\left[\left|\frac{1}{m}\sum_{j=l_0+1}^{m} R\left(q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(\bar{Y}^n_{km+j-1},\cdot)\|\rho_{\zeta^*(\tau_k^\ell)}(\bar{Y}^n_{km+j-1},\cdot)\right)\right.\right.$$
$$\left.\left.-\int R\left(q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(y,\cdot)\|\rho_{\zeta^*(\tau_k^\ell)}(y,\cdot)\right)\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy)\right|\right]=0.$$

We now show that $\bar{Y}^n_{n+km+l_0}\in\Phi_k$ w.p. 1. Because $\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(\Phi_k)=1$ and $\pi_{\zeta^*(\tau_k^\ell)} << \mu^{\zeta^*(t^n_{km}),\dot{\zeta}^*(t^n_{km}),\delta}$, it holds that $\pi_{\zeta^*(\tau_k^\ell)}(\Phi_k^c)=0$. Lemma 7.6 then implies that $\rho^{l_0}_{\zeta^*(\tau_k^\ell)}(y,\Phi_k^c)=0$. This, combined with the fact that we only consider a finite number ℓ terms, gives the convergence

$$\lim_{m\to\infty}\max_{k\in\{1,\dots,\ell\}} E\left[\left|\frac{1}{m}\sum_{j=l_0}^{m-1} R\left(q^{\zeta^*(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(\bar{Y}^n_{km+j},\cdot)\|\rho_{\zeta^*(\tau_k^\ell}(\bar{Y}^n_{km+j},\cdot)\right)\right.\right.$$
$$\left.\left.-\int R\left(q^{\zeta^*(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(y,\cdot)\|\rho_{\zeta^*(\tau_k^\ell}(y,\cdot)\right)\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)\right|\right]=0.$$

It follows that

$$\sup_m E\left[\frac{1}{m}\sum_{j=0}^{m-1} R\left(q^{\zeta^*(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(\bar{Y}^n_{km+j},\cdot)\|\rho_{\zeta^*(\tau_k^\ell)}(\bar{Y}^n_{km+j},\cdot)\right)\right]<\infty.$$

Next, we consider the second term in (7.3),

$$\int_{\mathbb{R}^{d_2}}\left(\log\frac{d\rho_{\zeta^*(\tau_k^\ell)}(\bar{Y}^n_{n+km+j-1},y)}{d\rho_{\bar{X}^n_{km+j}}(\bar{Y}^n_{n+km+j-1},y)}\right)\bar{\nu}^n_{n+km+j}(dy). \tag{7.4}$$

Since $km + j < \hat{i}^n$ and by the continuity of ζ^*, there exists a compact set K such that $\zeta^*(\tau^l_k), \bar{X}^n_{km+j} \in K$. By Assumption (A.4), there exists a C such that $\log \frac{d\rho_{\zeta^*(\tau^\ell_k)}(y,z)}{d\rho_{\bar{X}^n_{km+j}}(y,z)} \leq C$, for all n. This ensures that (7.4) is bounded.

■

In the process of proving Theorem 7.5, we will work with a generic sequence of control measures $\{\tilde{\nu}^n\}$, and associated controlled processes $\tilde{X}^n$. That is, we have a sequence of measures $\tilde{\nu}^n_i \in \mathcal{P}(\mathbb{R}^{d_2})$ and define the corresponding controlled process $\{\tilde{X}^n_k\}_{k \geq n}$ as before: $\tilde{X}^n_n = x$ and

$$\tilde{X}^n_{k+1} = \tilde{X}^n_k + \varepsilon_k g(\tilde{X}^n_k, \tilde{Y}^n_k),$$

where $\tilde{\nu}^n_k$ is the conditional distribution for $\tilde{Y}^n_k$ given $\sigma\left(\tilde{Y}^n_n, \dots, \tilde{Y}^n_{i-1}\right)$. Similar to before, we take $\tilde{X}^n \in C([0,T] : \mathbb{R}^{d_x})$ as the linear interpolation with breakpoints $\tilde{X}^n(t_{n+k} - t_n) = \tilde{X}^n_k$. We also abuse notation a bit and define $\tilde{\nu} \in \mathcal{P}(\mathbb{R}^{d_2} \times [0,T])$ as

$$\tilde{\nu}^n(A \times B) = \int_B \frac{1}{h^n(t)} \tilde{\nu}^n(A|t)dt,$$

where $\tilde{\nu}^n(A|t) = \tilde{\nu}^n_i(A)$ when $t \in [t_{n+i-1} - t_n, t_{n+i} - t_n)$.

The assumption we will make on the $\tilde{\nu}^n_i$s is that they satisfy the condition of bounded expected running cost,

$$\sup_n E\left[\frac{1}{\beta_n} \sum_{i=n}^{\beta_n+n-1} R(\tilde{\nu}^n_i(\cdot)||\rho_{\tilde{X}^n_i}(\tilde{Y}^n_i, \cdot))\right] < \infty. \tag{7.5}$$

Because of Lemma 7.7, we know that results that hold under this condition will also apply to our specific choice of controlled measures $\bar{\nu}^n$, defined in Section 7.1.

The main step towards proving Theorem 7.5 is to prove a version of it with such a generic sequence of controls.

Theorem 7.8 *Under (A.1)-(A.8), for every subsequence of $\{(\tilde{\nu}^n, \tilde{X}^n)\}$ where $\{\tilde{\nu}^n\}$ satisfies* (7.5)*, there exists a further subsequence that converges weakly to $(\tilde{\nu}, \tilde{X})$. Furthermore, there exists a stochastic kernel $\tilde{\nu}(dy|t)$ such that*

$$\tilde{\nu}(A \times B) = \int_B \tilde{\nu}(A|t) \frac{1}{h(t)} dt,$$

and $\tilde{X}$ satisfies

$$\tilde{X}(t) = x + \int_0^t \int_{\mathbb{R}^{d_2}} g(\tilde{X}(s), y) \tilde{\nu}(dy|s) ds. \tag{7.6}$$

Note that the form of the limit measure $\tilde{\nu}$ is a direct consequence of Lemma 3.3.1 in [9] and the uniform convergence $h^n \to h$, ensured by (A.8). We also have that since $\hat{S}^n$ takes values in the compact set $[0,T]$, there is a subsequence that converges to $\hat{S} \in [0,T]$. The proof of Theorem 7.8 is presented in Section 8. Theorem 7.5 follows from this result if we can show that the limit point for the appropriate subsequences of the specific choice of control measures in Section 7.1 have the claimed form. We will prove this in two steps, the strategy being to first send m to infinity, and find the corresponding limit point $\bar{x}^\ell$ of $\bar{X}^n$. Recall that from how we chose m and ℓ, for fix ℓ, taking n to infinity also means taking m to infinity, and vice versa. Moreover, note that we here suppress the dependence on δ and ε. Next, we send ℓ to infinity and δ to 0, and show that the corresponding limit for the $\bar{x}^\ell$ is ζ, as claimed. That is, we show the following convergence results:

$$\bar{X}^n \xrightarrow[\textit{Lemma 7.9}]{m\to\infty} \bar{x}^\ell \xrightarrow[\textit{Lemma 7.10}]{\delta\to 0,\ \ell\to\infty} \zeta.$$

We start with the first part: using Theorem 7.8 we characterise the limit point $\bar{x}^\ell$ and prove that $\bar{X}^n \to \bar{x}^\ell$ in probability as $m \to \infty$.

Lemma 7.9 *Under (A.1)-(A.8), for any $\delta > 0$, $\ell \in \mathbb{N}$, $\{\bar{X}^n\}_n$ is tight. Moreover, the convergent subsequences of $\{\bar{X}^n\}$ converge to $\bar{x}^\ell$ in probability, as $m \to \infty$, where $\bar{x}^\ell$ satisfies*

$$\begin{aligned}\bar{x}^\ell(t) = x_0 &+ \sum_{i=0}^{k-1} \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int_{\mathbb{R}^{d_2}} g(\bar{x}^\ell(s), y)\mu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell),\delta}(dy)ds \\ &+ \int_{\tau_k^\ell}^{t} \int_{\mathbb{R}^{d_2}} g(\bar{x}^\ell(s), y)\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy)ds,\end{aligned}$$

for $t \in [\tau_k^\ell, \tau_{k+1}^\ell)$ and $t \le \hat{S^n}$.

Proof.

For $j \in \{n+km+1, \dots, n+(k+1)m\}$ and $k \in \{0, \dots, \ell-1\}$, consider $t \in [t_{j-1}^n, t_j^n)$. Because we will consider the limit as $m \to \infty$, to emphasise the dependence on m in the $\bar{\nu}_j^n$s, we define

$$\gamma^m(dz|t) = \bar{\nu}_j^n(dz) = \begin{cases} \rho_{\zeta^*(\tau_k^\ell)}(\bar{Y}_{j-1}^n, dz), & n+km+1 \le j \le n+km+l_0, \\ q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(\bar{Y}_{j-1}^n, dz), & n+km+l_0+1 \le j \le n+(k+1)m-1. \end{cases}$$

Moreover, for $j \in \{n+\ell m+1, \dots, n+\beta_n\}$, i.e., when $t \in [t_{\ell m}^n, T)$ or $t \ge \hat{S}^n$, we set

$$\gamma^m(dz|t) = \rho_{\bar{X}_{j-1}^n}(\bar{Y}_{j-1}^n, dz).$$

For notational brevity and clarity, we also define

$$\gamma^m(A \times B) \doteq \int_B \gamma^m(A|t)dt,$$

and

$$\gamma(A \times B) \doteq \int_B \gamma(A|t)dt, \quad \gamma(A|t) = \mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(A), t \le \hat{S}.$$

where t and k are as above. For $t > \hat{S}$ we set $\gamma(A|t) = \lim_{m\to\infty} \gamma^m(A|t)$. Note that γ^m and $\gamma^m(\cdot|t)$ are playing the roles of $\bar{\nu}^n$ and $\bar{\nu}^n(\cdot|t)$. Combining Lemma 7.7 and Theorem 7.8, with these definitions of γ^m and γ, it is enough to show that γ^m converges weakly to γ w. p. 1.

To prove the convergence of γ^m, consider any bounded and uniformly continuous function $f : \mathbb{R}^{d_2} \times [0,T] \to \mathbb{R}$. By the Portmanteau theorem, it is enough to prove that

$$\int_{\mathbb{R}^{d_2}\times[0,T]} f(y,t)\gamma^m(dydt),$$

converges, as $m \to \infty$, to

$$\int_{\mathbb{R}^{d_2}\times[0,T]} f(y,t)\gamma(dydt).$$

Since $\gamma^m(dy|t) \to \gamma(dy|t)$ by definition for $t > \hat{S}$ the only interesting case is for the interval $[0, \hat{S}]$. Below the proof is constructed with $\hat{S} = T$. The case with $\hat{S} < T$ would be the same but over a shorter time interval. From the definition of γ^m we have

$$\begin{aligned}\int_{\mathbb{R}^{d_2}\times[0,T]} f(y,t)\gamma^m(dydt) &= \sum_{k=0}^{\ell-1}\sum_{j=n+km+1}^{n+(k+1)m}\int_{t^n_{j-1}}^{t^n_j}\int_{\mathbb{R}^{d_2}} f(y,t)\gamma^m(dy|t)dt \\ &+ \int_{t^n_{n+\ell m}}^{T}\int_{\mathbb{R}^{d_2}} f(y,t)\gamma^m(dy|t)dt.\end{aligned}$$

As a first step, for each $k \in \{0,1,\ldots,\ell-1\}$, we now consider the difference

$$\sum_{j=n+km+1}^{n+(k+1)m}\int_{t^n_{j-1}}^{t^n_j}\int_{\mathbb{R}^{d_2}} f(y,t)\gamma^m(dy|t)dt - \int_{\mathbb{R}^{d_2}\times[\tau^\ell_k,\tau^\ell_{k+1})} f(y,t)\gamma(dydt). \tag{7.7}$$

In preparation for studying (7.7) in the limit $m \to \infty$, we make the following definitions.

$$\begin{aligned}
C^k_1(m) &= \sum_{j=n+km+1}^{n+km+l_0}\int_{t^n_{j-1}}^{t^n_j}\int_{\mathbb{R}^{d_2}} f(x,t)\rho_{\bar{X}^n_{j-1}}(\bar{Y}^n_{j-1},dx)dt,\\
C^k_2(m) &= \sum_{j=n+km+l_0+1}^{n+(k+1)m}\int_{t^n_{j-1}}^{t^n_j}\int_{\mathbb{R}^{d_2}} f(x,t)q^{\zeta^*(\tau^\ell_k),\dot{\zeta}^*(\tau^\ell_k),\delta}(\bar{Y}^n_{j-1},dx)dt\\
&\quad - \sum_{j=n+km+l_0+1}^{n+(k+1)m}\int_{t^n_{j-1}}^{t^n_j}\int_{\mathbb{R}^{d_2}} f\left(x,t^n_{j-1}\right)q^{\zeta^*(\tau^\ell_k),\dot{\zeta}^*(\tau^\ell_k),\delta}(\bar{Y}^n_{j-1},dx)dt,\\
C^k_3(m) &= \sum_{j=n+km+l_0+1}^{n+(k+1)m}\varepsilon_j\int_{\mathbb{R}^{d_2}} f\left(x,t^n_{j-1}\right)q^{\zeta^*(\tau^\ell_k),\dot{\zeta}^*(\tau^\ell_k),\delta}(\bar{Y}^n_{j-1},dx)\\
&\quad - \sum_{j=n+km+l_0+1}^{n+(k+1)m}\varepsilon_j\int_{\mathbb{R}^{d_2}}\int_{\mathbb{R}^{d_2}} f(x,t^n_{j-1})q^{\zeta^*(\tau^\ell_k),\dot{\zeta}^*(\tau^\ell_k),\delta}(y,dx)\mu^{\zeta^*(\tau^\ell_k),\dot{\zeta}^*(\tau^\ell_k),\delta}(dy),\\
C^k_4(m) &= \sum_{j=n+km+l_0+1}^{n+(k+1)m}\varepsilon_j\int_{\mathbb{R}^{d_2}} f(y,t^n_{j-1})\mu^{\zeta^*(\tau^\ell_k),\dot{\zeta}^*(\tau^\ell_k),\delta}(dy)\\
&\quad - \int_{[\tau^\ell_k,\tau^\ell_{k+1}]}\int_{\mathbb{R}^{d_2}} f(y,t)\mu^{\zeta^*(\tau^\ell_k),\dot{\zeta}^*(\tau^\ell_k),\delta}(dy)dt.
\end{aligned}$$

With these definitions, we now rewrite $\sum_{j=n+km+1}^{n+(k+1)m}\int_{t^n_{j-1}}^{t^n_j}\int_{\mathbb{R}^{d_2}} f(y,t)\gamma^m(dy|t)dt$ in terms of $C^k_i(m)$, $i = 1,\ldots,4$, and $\int_{\mathbb{R}^{d_2}\times[t^n_{n+km},t^n_{n+(k+1)m})} f(y,t)\gamma(dydt)$. First, we split the sum over j into two terms according to l_0:

$$\begin{aligned}
&\sum_{j=n+km+1}^{n+(k+1)m} \int_{t_{j-1}^n}^{t_j^n} \int_{\mathbb{R}^{d_2}} f(y,t)\gamma^m(dy|t)dt \\
&\quad = \sum_{j=n+km+l_0+1}^{n+(k+1)m} \int_{t_{j-1}^n}^{t_j^n} \int_{\mathbb{R}^{d_2}} f(y,t) q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(\bar{Y}_{j-1}^n, dy)dt + C_1^k(m).
\end{aligned}$$

Next, for each interval $[t_{j-1}^n, t_j^n)$, we freeze the time-variable t inside $f(y,t)$ at t_{j-1}^n:

$$\begin{aligned}
&\sum_{j=n+km+l_0+1}^{n+(k+1)m} \int_{t_{j-1}^n}^{t_j^n} \int_{\mathbb{R}^{d_2}} f(y,t) q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(\bar{Y}_{j-1}^n, dy)dt + C_1^k(m) \\
&\quad = \sum_{j=n+km+l_0+1}^{n+(k+1)m} \int_{t_{j-1}^n}^{t_j^n} \int_{\mathbb{R}^{d_2}} f\left(y, t_{j-1}^n\right) q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(\bar{Y}_{j-1}^n, dy)dt + C_1^k(m) + C_2^k(m) \\
&\quad = \sum_{j=n+km+l_0+1}^{n+(k+1)m} \varepsilon_j \int_{\mathbb{R}^{d_2}} f\left(y, t_{j-1}^n\right) q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(\bar{Y}_{j-1}^n, dy) + C_1^k(m) + C_2^k(m),
\end{aligned}$$

where in the second step we have used that there is integral over the time variable now amounts to $t_j^n - t_{j-1}^n = \varepsilon_j$. As a next step, by averaging over the controlled variable $\bar{Y}_{j-1}^n$, we can write the last display as

$$\begin{aligned}
&\sum_{j=n+km+l_0+1}^{n+(k+1)m} \varepsilon_j \int_{\mathbb{R}^{d_2}} f\left(y, t_{j-1}^n\right) q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(\bar{Y}_{j-1}^n, dy) + C_1^k(m) + C_2^k(m) \\
&\quad = \sum_{j=n+km+l_0+1}^{n+(k+1)m} \varepsilon_j \int_{\mathbb{R}^{d_2}} \int_{\mathbb{R}^{d_2}} f(y, t_{j-1}^n) q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(y, dz) \mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy) \\
&\qquad + C_1^k(m) + C_2^k(m) + C_3^k(m).
\end{aligned}$$

Because $\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}$ is invariant for $q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}$, we have

$$\int_{\mathbb{R}^{d_2}} \int_{\mathbb{R}^{d_2}} f(y, t_{j-1}^n) q^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(y, dz) \mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy) = \int_{\mathbb{R}^{d_2}} f(y, t_{j-1}^n) \mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy).$$

Moreover, from the definition of $C_4^k(m)$, we have

$$\sum_{j=n+km+l_0+1}^{n+(k+1)m} \varepsilon_j \int_{\mathbb{R}^{d_2}} f(y, t_{j-1}^n) \mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy) = \int_{\tau_k^\ell}^{\tau_{k+1}^\ell} \int_{\mathbb{R}^{d_2}} f(y, t_{j-1}^n) \mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy)dt + C_4^k(m)$$

Combining the steps above, and the definition of γ, we can express the difference (7.7) as

$$\sum_{j=n+km+1}^{n+(k+1)m} \int_{t_{j-1}^n}^{t_j^n} \int_{\mathbb{R}^{d_2}} f(y,t)\gamma^m(dy|t)dt - \int_{\mathbb{R}^{d_2}} \int_{\tau_k^\ell}^{\tau_{k+1}^\ell} f(y,t)\gamma(dydt) = \sum_{i=1}^{4} C_i^k(m).$$

We now consider the $C_i^k(m)$-terms, for a fixed $k \in \{0, 1, \dots, \ell - 1\}$, as we let m go to infinity.

For $C_1^k(m)$, because f is bounded, the sum only contains a finite number of terms, and

$$t^n_{n+km+l_0} - t^n_{n+km+1} \to 0, \ m \to \infty,$$

we have that $C_1^k(m) \to 0$.

For $C_2^k(m)$, we can write this term as

$$C_2^k(m) = \sum_{j=n+km+l_0+1}^{n+(k+1)m} \int_{t^n_{j-1}}^{t^n_j} \int_{\mathbb{R}^{d_2}} \left(f(y,t) - f(y, t^n_{j-1})\right) q^{\zeta^*(\tau^\ell_k), \dot{\zeta}^*(\tau^\ell_k), \delta}(\bar{Y}^n_{j-1}, dx) dt.$$

Using the uniform continuity of f, these terms can be made arbitrarily small.

Next, for $C_3^k(m)$, we use an argument similar to what is used in the proof of Lemma 7.7.

For $C_4^k(m)$, we utilise Riemann integrability of the function $\hat{f} : [0, T] \to \mathbb{R}$ defined by

$$t \mapsto \hat{f}(t) = \int_{\mathbb{R}^{d_2}} f(y,t) \mu^{\zeta^*(\tau^\ell_k), \dot{\zeta}^*(\tau^\ell_k), \delta}(dy).$$

Noting that $t^n_{n+(k+1)m} \to \tau^\ell_{k+1}$ and $t^n_{n+km+l_0+1} \to \tau^\ell_k$,

$$\sum_{j=n+km+l_0+1}^{n+(k+1)m} \varepsilon_j \int_{\mathbb{R}^{d_2}} f(y, t^n_{j-1}) \mu^{\zeta^*(\tau^\ell_k), \dot{\zeta}^*(\tau^\ell_k), \delta}(dy),$$

is a Riemann sum and converges to $\int_{\tau^\ell_k}^{\tau^\ell_{k+1}} \hat{f}(t) dt$ as $m \to \infty$. Thus $C_4^k(m) \to 0$ as $m \to \infty$.

We have established that, for each $k \in \{0, \dots, \ell - 1\}$, $\sum_{i=1}^4 C_i^k(m) \to 0$, as $m \to \infty$. It follows that

$$\sum_{k=0}^{\ell-1} \sum_{i=1}^4 C_i^k(m) \to 0, \quad m \to \infty.$$

By extension, as $m \to \infty$,

$$\begin{aligned} \sum_{k=0}^{\ell-1} \sum_{j=n+km+1}^{n+(k+1)m} \int_{t^n_{j-1}}^{t^n_j} \int_{\mathbb{R}^{d_2}} f(y,t) \gamma^m(dy|t) dt &\to \sum_{k=0}^{\ell-1} \int_{\tau^\ell_k}^{\tau^\ell_{k+1}} \int_{\mathbb{R}^{d_2}} f(y,t) \gamma(dy|t) dt \\ &= \int_0^T \int_{\mathbb{R}^{d_2}} f(y,t) \gamma(dydt). \end{aligned}$$

It remains to consider the term

$$\int_{t^n_{n+\ell m}}^T \int_{\mathbb{R}^{d_2}} f(y,t) \gamma^m(dy|t) dt, \tag{7.8}$$

in the limit as $m \to \infty$. Since f is bounded, $\gamma^m(\cdot|t)$ is a probability measure for each $t \in [0, T]$, and $t^n_{n+\ell m} \to \tau^\ell_\ell = T$ as $m \to \infty$, we have that (7.8) vanishes in this limit. Thus, we have shown that, w. p. 1, for arbitrary bounded and uniformly continuous $f : \mathbb{R}^{d_2} \times [0, T] \to \mathbb{R}$,

$$\int_{\mathbb{R}^{d_2} \times [0,T]} f(y,t) \gamma^m(dydt) \to \int_{\mathbb{R}^{d_2} \times [0,T]} f(y,t) \gamma(dydt), \quad m \to \infty.$$

That is, w. p. 1 we have the weak convergence $\gamma^m \to \gamma$, as $m \to \infty$. This completes the proof. ■

The next step is to prove the convergence of $\bar{x}^\ell$ when taking $\delta \to 0$ and $\ell \to \infty$, in that order. We have the following result.

Lemma 7.10 *Assume (A.1)-(A.8) hold and let $\{\bar{x}^\ell\}$ be the process defined in Lemma 7.9. Then, $\{\bar{x}^\ell\}$ converges to ζ^*, on the time interval $[0, \hat{S}]$, in the limit as $\delta \to 0$ and $\ell \to \infty$.*

Before proving Lemma 7.10, we show the integrability of $g(x, \cdot)$ with respect to π_x, for each $x \in \mathbb{R}^{d_1}$, which is used in the proof.

Lemma 7.11 *Under (A.1)-(A.8), for any $x \in \mathbb{R}^{d_1}$, the function $y \mapsto g(x, y)$ is integrable with respect to π_x.*

Proof. Since $1 + x \leq e^x$ for all $x \in \mathbb{R}$, from (A.7) we have that for all $x, \alpha \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$,

$$1 + \int_{\mathbb{R}^{d_2}} \langle \alpha, g(x, z) \rangle \rho_x(y, dz) \leq \sup_y \int_{\mathbb{R}^{d_2}} e^{\langle \alpha, g(x,z) \rangle} \rho_x(y, dz) < \infty.$$

By taking α as the unit vectors e_i, for every $i = 1, \ldots, d_1$, in the last display, the upper bound implies the finiteness of every component of

$$\sup_y \int_{\mathbb{R}^{d_2}} g(x, z) \rho_x(y, dz).$$

Moreover, since $\pi_x \rho_x = \pi_x$, we have,

$$\int_{\mathbb{R}^{d_2}} g(x, z) \pi_x(dz) = \int_{\mathbb{R}^{d_2}} \left(\int_{\mathbb{R}^{d_2}} g(x, z) \rho_x(y, dz) \right) \pi_x(dy).$$

Therefore, every component of the left integral is finite, which proves the claim. ■

Before the proof of Lemma 7.10 we prove that if the process $\bar{x}^l$ converges to ζ^* on $[0, \hat{S}]$, then it also converges to ζ^* on $[0, T]$

Corollary 7.12 *Assume that $\bar{x}^l$ converges to ζ^* on $[0, \hat{S}]$, then $\hat{S} = T$.*

Proof. Assume that $\hat{S} < T$. By the convergence of $\bar{x}^l$ we have that $\|\bar{x}^\ell(\hat{S}) - \zeta^*(\hat{S})\|$ can be made arbitrarily small for small enough δ and large enough ℓ . ζ^* is continuous by definition and by Theroem 7.8 $\bar{x}^\ell$ is continuous on the whole interval $[0, T]$. From the definition of $\hat{S}$ we have that

$$\lim_{t \to \hat{S}^+} |\xi^\ell(t) - \zeta^*(t)\| \geq 1.$$

But this contradicts the continuity of ζ^* and x^ℓ and we conclude that $\hat{S} = T$. ■

We now move to the proof of Lemma 7.10. The proof uses arguments similar to those used in Section 8 to prove Theorem 7.8. Specifically, we employ arguments similar to those used in the proof of Lemma 8.4. For simplicity the proof is done with $\hat{S} = T$, the proof in the case $\hat{S} < T$ is the same but on on a smaller interval.

Proof of Lemma 7.10. As already noted, by construction of the $\nu^{\zeta^*(t),\dot{\zeta}^*(t)}$-measures, for all $t \in [0,T]$, ζ^* satisfies

$$\zeta^*(t) = x_0 + \int_0^t \int_{\mathbb{R}^{d_2}} g(\zeta^*(s), y)\nu^{\zeta^*(t),\dot{\zeta}^*(t)}(dy)ds,$$

and Lemma A.2 ensures that the solution is unique. To show the claimed convergence, we consider the difference between $\bar{x}^\ell$ and ζ^*:

$$\begin{aligned}
&\|\bar{x}^\ell - \zeta^*\|_\infty = \sup_{t\in[0,T]} \|\bar{x}^\ell(t) - \zeta^*(t)\| \\
&= \sup_{t\in[0,T]} \Bigg\| \sum_{i=0}^{k-1} \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(s), y)\mu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell),\delta}(dy)ds + \int_{\tau_k^\ell}^t \int g(\zeta^*(s), y)\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy)ds \\
&\qquad - \int_0^t \int g(\zeta^*(s), y)\nu^{\zeta^*(s),\dot{\zeta}^*(s)}(dy)ds \Bigg\| \\
&\le \sup_{t\in[0,T]} \Bigg\| \sum_{i=0}^{k-1} \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(s), y)\mu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell),\delta}(dy)ds + \int_{\tau_k^\ell}^t \int g(\zeta^*(s), y)\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy)ds \\
&\qquad - \sum_{i=0}^{k-1} \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(s), y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds - \int_{\tau_k^\ell}^t \int g(\zeta^*(s), y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds \Bigg\| \\
&\quad + \sup_{t\in[0,T]} \Bigg\| \sum_{i=0}^{k-1} \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(s), y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds + \int_{\tau_k^\ell}^t \int g(\zeta^*(s), y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds \\
&\qquad - \int_0^t \int g(\zeta^*(s), y)\nu^{\zeta^*(s),\dot{\zeta}^*(s)}(dy)ds \Bigg\|
\end{aligned}$$

We now treat the two suprema in the upper bound separately, and start by considering a fixed but arbitrary $t \in [0,T]$. For the terms inside the first supremum, for any $i \in \{0, 1, \dots, k-1\}$, we have the upper bound

$$\begin{aligned}
&\left\| \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(s), y)\mu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell),\delta}(dy)ds - \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(s), y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds \right\| \\
&\quad \le \left\| \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(\tau_i^\ell), y)\mu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell),\delta}(dy)ds - \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(\tau_i^\ell), y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds \right\| \\
&\qquad + \left\| \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int \Big(g(\zeta^*(\tau_i^\ell), y) - g(\zeta^*(s), y) \Big) \mu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell),\delta}(dy)ds \right\| \\
&\qquad + \left\| \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int \Big(g(\zeta^*(\tau_i^\ell), y) - g(\zeta^*(s), y) \Big) \nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds \right\|.
\end{aligned}$$

Similarly, for the term involving integrals from τ_k^ℓ to t,

$$\begin{aligned}
&\left\| \int_{\tau_k^\ell}^t \int g(\zeta^*(s), y) \mu^{\zeta^*(\tau_k^\ell), \dot{\zeta}^*(\tau_k^\ell), \delta}(dy) ds - \int_{\tau_k^\ell}^t \int g(\zeta^*(s), y) \nu^{\zeta^*(\tau_k^\ell), \dot{\zeta}^*(\tau_k^\ell)}(dy) ds \right\| \\
&\quad \leq \left\| \int_{\tau_k^\ell}^t \int g(\zeta^*(\tau_k^\ell), y) \mu^{\zeta^*(\tau_k^\ell), \dot{\zeta}^*(\tau_k^\ell), \delta}(dy) ds - \int_{\tau_k^\ell}^t \int g(\zeta^*(\tau_k^\ell), y) \nu^{\zeta^*(\tau_k^\ell), \dot{\zeta}^*(\tau_k^\ell)}(dy) ds \right\| \\
&\qquad + \left\| \int_{\tau_k^\ell}^t \int \left(g(\zeta^*(\tau_k^\ell), y) - \int_{\tau_k^\ell}^t \int g(\zeta^*(s), y) \right) \mu^{\zeta^*(\tau_k^\ell), \dot{\zeta}^*(\tau_k^\ell), \delta}(dy) ds \right\| \\
&\qquad + \left\| \int_{\tau_k^\ell}^t \int \left(g(\zeta^*(\tau_k^\ell), y) - \int_{\tau_k^\ell}^t \int g(\zeta^*(s), y) \right) \nu^{\zeta^*(\tau_k^\ell), \dot{\zeta}^*(\tau_k^\ell), \delta}(dy) ds \right\|
\end{aligned}$$

From the definitions of $\mu^{\zeta^*(\tau_i^\ell), \dot{\zeta}^*(\tau_i^\ell), \delta}$ and $\nu^{\zeta^*(\tau_k^\ell), \dot{\zeta}^*(\tau_k^\ell), \delta}$, we have that

$$\begin{aligned}
&\left\| \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(\tau_i^\ell), y) \mu^{\zeta^*(\tau_i^\ell), \dot{\zeta}^*(\tau_i^\ell), \delta}(dy) ds - \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(\tau_i^\ell), y) \nu^{\zeta^*(\tau_i^\ell), \dot{\zeta}^*(\tau_i^\ell)}(dy) ds \right\| \\
&\quad = |\tau_{i+1}^\ell - \tau_i^\ell| \frac{\delta}{2} \left\| \int g(\zeta^*(\tau_i^\ell), y) \pi_{\zeta^*(\tau_i^\ell)}(dy) - \int g(\zeta^*(\tau_i^\ell), y) \nu^{\zeta^*(\tau_i^\ell), \dot{\zeta}^*(\tau_i^\ell)}(dy) \right\| \\
&\quad = |\tau_{i+1}^\ell - \tau_i^\ell| \frac{\delta}{2} \left\| \int g(\zeta^*(\tau_i^\ell), y) \pi_{\zeta^*(\tau_i^\ell)}(dy) - \dot{\zeta}^*(\tau_i^\ell) \right\|
\end{aligned}$$

The integral inside the norm is finite by Lemma 7.11.

Next, by the uniform Lipschitz property for g, we have

$$\begin{aligned}
&\left\| \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int \left(g(\zeta^*(\tau_i^\ell), y) - g(\zeta^*(s), y) \right) \mu^{\zeta^*(\tau_i^\ell), \dot{\zeta}^*(\tau_i^\ell), \delta}(dy) ds \right\| \\
&\quad \leq L_g \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \|\zeta^*(s) - \zeta^*(\tau_i^\ell)\| ds.
\end{aligned}$$

In precisely the same way we have

$$\begin{aligned}
&\left\| \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int \left(g(\zeta^*(\tau_i^\ell), y) - g(\zeta^*(s), y) \right) \nu^{\zeta^*(\tau_i^\ell), \dot{\zeta}^*(\tau_i^\ell)}(dy) ds \right\| \\
&\quad \leq L_g \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \|\zeta^*(s) - \zeta^*(\tau_i^\ell)\| ds.
\end{aligned}$$

Combining these inequalities yields the upper bound,

$$\begin{aligned}
&\left\| \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(s), y) \mu^{\zeta^*(\tau_i^\ell), \dot{\zeta}^*(\tau_i^\ell), \delta}(dy) ds - \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \int g(\zeta^*(s), y) \nu^{\zeta^*(\tau_i^\ell), \dot{\zeta}^*(\tau_i^\ell)}(dy) ds \right\| \\
&\quad \leq (\tau_{i+1}^\ell - \tau_i^\ell) \frac{\delta}{2} \left\| \int g(\zeta^*(\tau_i^\ell), y) \pi_{\zeta^*(\tau_i^\ell)}(dy) - \dot{\zeta}^*(\tau_i^\ell) \right\| + 2L_g \int_{\tau_i^\ell}^{\tau_{i+1}^\ell} \|\zeta^*(s) - \zeta^*(\tau_i^\ell)\| ds.
\end{aligned}$$

We can use the same arguments as above once more to obtain an upper bound for the term involving integrals from τ_k^ℓ to t:

$$\begin{aligned}&\left\|\int_{\tau_k^\ell}^t \int g(\zeta^*(s),y)\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy)ds - \int_{\tau_k^\ell}^t \int g(\zeta^*(s),y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds\right\|\\ &\quad\le (t-\tau_k^\ell)\frac{\delta}{2}\left\|\int g(\zeta^*(\tau_k^\ell),y)\pi_{\zeta^*(\tau_k^\ell)}(dy) - \dot{\zeta}^*(\tau_k^\ell)\right\| + 2L_g\int_{\tau_k^\ell}^t \|\zeta^*(s)-\zeta^*(\tau_k^\ell)\|ds.\end{aligned}$$

Combining the upper bounds yields

$$\begin{aligned}&\sup_{t\in[0,T]}\left\|\sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int g(\zeta^*(s),y)\mu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell),\delta}(dy)ds + \int_{\tau_k^\ell}^t\int g(\zeta^*(s),y)\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy)ds\right.\\ &\qquad\left.-\sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int g(\zeta^*(s),y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds - \int_{\tau_k^\ell}^t\int g(\zeta^*(s),y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds\right\|\\ &\quad\le \sup_{t\in[0,T]}\left\{\sum_{i=1}^{k-1}\left((\tau_{i+1}^\ell-\tau_i^\ell)\frac{\delta}{2}\left\|\int g(\zeta^*(\tau_i^\ell),y)\pi_{\zeta^*(\tau_i^\ell)}(dy)-\dot{\zeta}^*(\tau_i^\ell)\right\| + 2L_g\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\|\zeta^*(s)-\zeta^*(\tau_i^\ell)\|ds\right)\right.\\ &\qquad\left. + \ (t-\tau_k^\ell)\frac{\delta}{2}\left\|\int g(\zeta^*(\tau_k^\ell),y)\pi_{\zeta^*(\tau_k^\ell)}(dy)-\dot{\zeta}^*(\tau_k^\ell)\right\| + 2L_g\int_{\tau_k^\ell}^t\|\zeta^*(s)-\zeta^*(\tau_k^\ell)\|ds\right\},\end{aligned}$$

where the value for k depends on t.

To deal with the supremum over t, we note that increasing t will only add more non-negative terms, and the terms corresponding to time-differences will be maximal for $t=T$. This results in $k=l$ and we have

$$\begin{aligned}&\sup_{t\in[0,T]}\left\|\sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int g(\zeta^*(s),y)\mu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell),\delta}(dy)ds + \int_{\tau_k^\ell}^t\int g(\zeta^*(s),y)\mu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell),\delta}(dy)ds\right.\\ &\qquad\left.-\sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int g(\zeta^*(s),y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds - \int_{\tau_k^\ell}^t\int g(\zeta^*(s),y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds\right\|\\ &\quad\le \sum_{i=1}^{\ell-1}\left((\tau_{i+1}^\ell-\tau_i^\ell)\frac{\delta}{2}\left\|\int g(\zeta^*(\tau_i^\ell),y)\pi_{\zeta^*(\tau_i^\ell)}(dy)-\dot{\zeta}^*(\tau_i^\ell)\right\| + 2L_g\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\|\zeta^*(s)-\zeta^*(\tau_i^\ell)\|ds\right),\end{aligned}$$

where we have also used that $\tau_\ell^\ell = T$.

For the second supremum, we split it according to

$$\begin{aligned}
&\sup_{t\in[0,T]}\left\|\sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int g(\zeta^*(s),y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds+\int_{\tau_k^\ell}^{t}\int g(\zeta^*(s),y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds\right.\\
&\qquad\qquad\left.-\int_0^t\int g(\zeta^*(s),y)\nu^{\zeta^*(s),\dot{\zeta}^*(s)}(dy)ds\right\|\\
&\le\sup_{t\in[0,T]}\left\{\left\|\sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int g(\zeta^*(s),y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds-\sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int g(\zeta^*(\tau_i^\ell),y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds\right.\right.\\
&\qquad\left.+\int_{\tau_k^\ell}^{t}\int g(\zeta^*(s),y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds-\int_{\tau_k^\ell}^{t}\int g(\zeta^*(\tau_k^\ell),y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds\right\|\\
&\qquad+\left\|\sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int g(\zeta^*(\tau_i^\ell),y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds+\int_{\tau_k^\ell}^{t}\int g(\zeta^*(\tau_k^\ell),y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds\right.\\
&\qquad\qquad\left.\left.-\int_0^t\int g(\zeta^*(s),y)\nu^{\zeta^*(s),\dot{\zeta}^*(s)}(dy)ds\right\|\right\}
\end{aligned}$$

Similar to before, we start by treating the terms inside the supremum to obtain suitable upper bounds. In this direction the second norm-term is the easiest to treat. From the definitions of the $\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}$-measures and the properties of ζ^*,

$$\begin{aligned}
&\left\|\sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int g(\zeta^*(\tau_i^\ell),y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds+\int_{\tau_k^\ell}^{t}\int g(\zeta^*(\tau_k^\ell),y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds\right.\\
&\qquad\left.-\int_0^t\int g(\zeta^*(s),y)\nu^{\zeta^*(s),\dot{\zeta}^*(s)}(dy)ds\right\|\\
&\quad=\left\|\sum_{i=0}^{k-1}(\tau_{i+1}^\ell-\tau_i^\ell)\dot{\zeta}^*(\tau_i^\ell)+(t-\tau_k^\ell)\dot{\zeta}^*(\tau_k^\ell)-\zeta^*(t)\right\|.
\end{aligned}$$

This term will converge to 0 uniformly in t as l grows, due to the properties of ζ^* (see Lemma 7.3).

For the other term inside the supremum, we have the upper bound

$$\begin{aligned}
&\Bigg\|\sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int g(\zeta^*(s),y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds - \sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int g(\zeta^*(\tau_i^\ell),y)\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds \\
&\qquad + \int_{\tau_k^\ell}^{t}\int g(\zeta^*(s),y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds - \int_{\tau_k^\ell}^{t}\int g(\zeta^*(\tau_k^\ell),y)\nu^{\zeta^*(\tau_k^\ell),\dot{\zeta}^*(\tau_k^\ell)}(dy)ds\Bigg\| \\
&\quad\leq \sum_{i=0}^{k-1}\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\int\Big\|g(\zeta^*(s),y)-g(\zeta^*(\tau_i^\ell),y)\Big\|\,\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds \\
&\qquad + \int_{\tau_k^\ell}^{t}\int\Big\|g(\zeta^*(s),y)-g(\zeta^*(\tau_k^\ell),y)\Big\|\,\nu^{\zeta^*(\tau_i^\ell),\dot{\zeta}^*(\tau_i^\ell)}(dy)ds \\
&\quad\leq \sum_{i=0}^{k-1}L_g\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\|\zeta^*(s)-\zeta^*(\tau_i^\ell)\|ds + L_g\int_{\tau_k^\ell}^{t}\|g(\zeta^*(s))-g(\zeta^*(\tau_k^\ell))\|ds.
\end{aligned}$$

Similar to before, we see that the supremum is achieved at $t=T$, and thus $k=\ell$. Together with the preceding calculations this yields the upper bound

$$\begin{aligned}
\|\bar{x}^\ell-\zeta^*\|_\infty &= \sup_{t\in[0,T]}\|\bar{x}^\ell(t)-\zeta^*(t)\| \\
&\leq \sum_{i=1}^{\ell-1}\Bigg((\tau_{i+1}^\ell-\tau_i^\ell)\frac{\delta}{2}\left\|\int g(\zeta^*(\tau_i^\ell),y)\pi_{\zeta^*(\tau_i^\ell)}(dy)-\dot{\zeta}^*(\tau_i^\ell)\right\| \\
&\qquad\qquad +3L_g\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\|\zeta^*(s)-\zeta^*(\tau_i^\ell)\|ds\Bigg).
\end{aligned}$$

Note that for any ℓ, by sending δ to 0, we have

$$\sum_{i=1}^{\ell-1}(\tau_{i+1}^\ell-\tau_i^\ell)\frac{\delta}{2}\left\|\int g(\zeta^*(\tau_i^\ell),y)\pi_{\zeta^*(\tau_i^\ell)}(dy)-\dot{\zeta}^*(\tau_i^\ell)\right\|\to 0.$$

Next, by the uniform continuity of ζ^*, for ℓ large enough, we have that for any $\delta>0$,

$$\max_{i\in\{0,\ell-1\}}\sup_{s\in[\tau_{i+1}^\ell,\tau_i^\ell]}\|\zeta^*(s)-\zeta^*(\tau_i^\ell)\|<\delta.$$

Using this we get

$$\sum_{i=1}^{\ell-1}3L_g\int_{\tau_i^\ell}^{\tau_{i+1}^\ell}\|\zeta^*(s)-\zeta^*(\tau_i^\ell)\|ds<\sum_{i=1}^{\ell-1}3\delta L_g(\tau_{i+1}^\ell-\tau_i^\ell)=3\delta L_gT,$$

which can be made arbitrarily small. ■

We can now show that the second term in Equation 7.3 is negligible.

Lemma 7.13 *Under (A.1)-(A.7), we have the following convergence*

$$\limsup_{\ell\to\infty}\limsup_{\delta\to 0}\limsup_{m\to\infty} E\left[\frac{1}{n}\sum_{k=0}^{\ell-1}\sum_{j=0}^{m-1}\int\left(\log\frac{d\rho_{\zeta(\tau_k^\ell)}(\bar{Y}^n_{km+j},\cdot)}{d\rho_{\bar{X}^n_{km+j}}(\bar{Y}^n_{km+j},\cdot)}\right) q^{\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(\bar{Y}^n_{km+j},dy)\right]=0.$$

Proof. Recall that $\rho_x(y,dz)=\eta_x(y,z)\lambda(dz)$ where $\eta_x(y,z)$ is continuous in x, uniformly in y,z, this implies that

$$\log\frac{d\rho_x(y,\cdot)}{d\rho_w(y,\cdot)}=\log\frac{\eta_x(y,\cdot)}{\eta_w(y,\cdot)}\to 0 \text{ as } x\to w.$$

It suffice to prove that $\{\bar{X}^n_{km}\}$ and converges to $\zeta(\tau_k^\ell)$. This is true by Lemmas 7.9 and 7.10. ■

With the continuity of F, the above lemma and Lemma 7.9, we find that for any $\varepsilon,\delta>0$.

$$\begin{aligned}
&\limsup_{m\to\infty}-\frac{1}{\beta_n}\log Ee^{-\beta_n F(X^n)}\\
&\quad\le\limsup_{m\to\infty}E\left[F(\bar{X}^n)+\frac{1}{m\ell}\sum_{k=0}^{\ell-1}\sum_{j=0}^{m-1}R\left(\nu^n_{km+j+1}(\cdot)\|\rho_{\bar{X}^n_{km+j}}(\bar{Y}^n_{km+j},\cdot)\right)\right]\\
&\quad\le\limsup_{m\to\infty}E\left[F(\bar{X}^n)+\frac{1}{m\ell}\sum_{k=0}^{\ell-1}\sum_{j=l_0}^{m-1}R\left(q^{\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(\bar{Y}^n_{n+km+j},\cdot)\|\rho_{\zeta(\tau_k^\ell}(\bar{Y}^n_{km+j},\cdot)\right)\right]\\
&\quad\le E\left[F(\hat{X}^\ell)+\frac{1}{\ell}\sum_{k=0}^{\ell-1}\int R\left(q^{\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(y,\cdot)\|\rho_{\zeta(\tau_k^\ell)}(y,\cdot)\right)\mu^{\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(dy)\right].
\end{aligned}$$

Notice that from Lemma 7.2,

$$\begin{aligned}
&E\left[\int R\left(q^{\zeta(\tau_k^\ell),\zeta\dot{,}\delta(\tau_k^\ell)}(y,\cdot)\|\rho_{\zeta(\tau_k^\ell)}(y,\cdot)\right)\mu^{\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(dy)\right]\\
&\quad=E\left[R\left(\mu^{\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}\otimes q^{\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}(\cdot,\cdot)\|\mu^{\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell),\delta}\otimes\rho_{\zeta(\tau_k^\ell)}(\cdot,\cdot)\right)\right]\\
&\quad\le E\left[L\left(\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell)\right)\right]+\varepsilon.
\end{aligned}$$

Thus,

$$\limsup_{m\to\infty}-\frac{1}{\beta_n}\log Ee^{-\beta_n F(X^n)}\le E\left[F(\hat{X}^\ell)+\frac{1}{\ell}\sum_{k=0}^{\ell-1}L\left(\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell)\right)\right]+\varepsilon.$$

This we will now rewrite as a Riemann sum

$$\frac{1}{\ell}\sum_{k=0}^{\ell-1}L\left(\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell)\right)=\sum_{k=0}^{\ell-1}\frac{1}{\ell(\tau_{k+1}^\ell-\tau_k^\ell)}L\left(\zeta(\tau_k^\ell),\dot{\zeta}(\tau_k^\ell)\right)(\tau_{k+1}^\ell-\tau_k^\ell).$$

The term $\frac{1}{\ell(\tau^\ell_{k+1}-\tau^\ell_k)}$ is bounded from above by $\frac{1}{h(\tau^\ell_{k+1})}$

$$\begin{aligned}\ell(\tau^\ell_{k+1}-\tau^\ell_k) &= \lim_{n\to\infty}\sum_{i=\lfloor n+\frac{k}{\ell}\beta_n\rfloor}^{\lfloor n+\frac{k+1}{\ell}\beta_n\rfloor}\varepsilon_i\ell = \lim_{n\to\infty}\sum_{i=\lfloor n+\frac{k}{\ell}\beta_n\rfloor}^{\lfloor n+\frac{k+1}{\ell}\beta_n\rfloor}\varepsilon_i\beta_n\frac{\ell}{\beta_n}\\ &\geq \lim_{n\to\infty}\varepsilon_{\lfloor n+\frac{k+1}{\ell}\beta_n\rfloor}\beta_n = h(\tau^\ell_{k+1}).\end{aligned}$$

Lastly, since δ is arbitrary, F is continuous, ζ is piecewise linear with finitely many pieces (Lemma 7.3) and the use of Lemma 7.10, we find

$$\begin{aligned}&\limsup_{\ell\to\infty}\limsup_{m\to\infty}-\frac{1}{\beta_n}\log Ee^{-\beta_n F(X^n)}\\ &\leq \limsup_{\ell\to\infty}E\left[\limsup_{\delta\to0}F(\hat{X}^\ell)+\sum_{k=0}^{\ell-1}\frac{1}{h(\tau^\ell_{k+1})}L\left(\zeta(\tau^\ell_k),\dot{\zeta}(\tau^\ell_k)\right)(\tau^l_{k+1}-\tau^\ell_k)\right]+\varepsilon\\ &\leq F(\zeta)+\int_0^T\frac{1}{h(t)}L(\zeta(t),\dot{\zeta}(t))dt+\varepsilon\\ &= F(\zeta)+I(\zeta)\leq\inf_\varphi(F(\varphi)+I(\varphi))+2\varepsilon.\end{aligned}$$

Sending $\varepsilon\to0$, we get

$$\limsup_{\ell\to\infty}\limsup_{m\to\infty}-\frac{1}{n}\log Ee^{-nF(X^n)}\leq\inf_\varphi(F(\varphi)+I(\varphi)).$$

8 Proof of Theorem 7.8

In this section we carry out the proof of Theorem 7.8, the convergence result for $(\tilde{\nu}^n,\tilde{X}^n)$ when $\{\tilde{\nu}^n\}$ is a generic sequence of control measures satisfying bounded expected running cost.

The first step towards proving Theorem 7.8 is the following uniform integrability property.

Lemma 8.1 *Under (A.1)-(A.8), if $\{\tilde{\nu}^n\}$ has bounded running cost, i.e.*

$$\sup_n E\left[\frac{1}{\beta_n}\sum_{i=n}^{\beta_n+n-1}R(\tilde{\nu}^n_i(\cdot)||\rho_{\tilde{X}^n_i}(\bar{Y}^n_i,\cdot))\right]<\infty, \tag{8.1}$$

then it satisfies the uniform integrability property,

$$\lim_{C\to\infty}\sup_n E\left[\int_0^T\int_{\|g(\tilde{X}^n(t),z)\|>C}\|g(\tilde{X}^n(t),z)\|\tilde{\nu}^n(dz\times dt)\right]=0. \tag{8.2}$$

Proof. The proof uses the inequality $ab\leq e^{\sigma a}+\frac{1}{\sigma}(b\log(b)-b+1)$ with $a=\|g(\tilde{X}^n(t),z)\|$ and

$b = \frac{d\tilde{\nu}_i^n(\cdot)}{d\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,\cdot)}$. For $t \in [0,T]$, and fixed C and n, we have,

$$
\begin{aligned}
&\int_{\|g(\tilde{X}^n(t),z)\|>C} \|g(\tilde{X}^n(t),z)\| d\tilde{\nu}_i^n(dz) \\
&= \int_{\|g(\tilde{X}^n(t),z)\|>C} \|g(\tilde{X}^n(t),z)\| \frac{d\tilde{\nu}_i^n(z)}{d\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,z)} \rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,dz) \\
&\leq \int_{\|g(\tilde{X}^n(t),z)\|>C} e^{\sigma\|g(\tilde{X}^n(t),z)\|} \rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,dz) \\
&\quad + \frac{1}{\sigma}\int_{\|g(\tilde{X}^n(t),z)\|>C} \left(\frac{d\tilde{\nu}_i^n(z)}{d\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,z)} \log\left(\frac{d\tilde{\nu}_i^n(z)}{d\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,z)} \right) - \frac{d\tilde{\nu}_i^n(z)}{d\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,z)} + 1 \right) \rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,dz) \\
&\leq \int_{\|g(\tilde{X}^n(t),z)\|>C} e^{\sigma\|g(\tilde{X}^n(t),z)\|} \rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,dz) + \frac{1}{\sigma} R(\tilde{\nu}_i^n(\cdot)||\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,\cdot)) \\
&\leq e^{-\sigma C} \sup_x \sup_y \int e^{2\sigma\|g(x,z)\|} \rho_x(y,dz) + \frac{1}{\sigma} R(\tilde{\nu}_i^n(\cdot)||\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,\cdot)),
\end{aligned}
$$

Assumption (A.7) to guarantee that the first term is finite. Using this bound yields

$$
\begin{aligned}
&E\left[\int_0^T \int_{\|g(\tilde{X}^n(t),z)\|>C} \|g(\tilde{X}^n(t),z)\| \tilde{\nu}^n(dz \times dt) \right] \\
&\leq E\left[\sum_{i=n}^{\beta_n+n-1} \int_{t_i,t_{i+1}} \frac{1}{h_n(t)} e^{-\sigma C} \sup_x \sup_y \int e^{2\sigma\|z\|} \rho_x(y,dz) + \frac{1}{h_n(t)} \frac{1}{\sigma} R(\tilde{\nu}_i^n(\cdot)||\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,\cdot)) dt \right] \\
&=\leq e^{-\sigma C} \sup_x \sup_y \int e^{2\sigma\|z\|} \rho_x(y,dz) + \frac{1}{\sigma} E\left[\frac{1}{\beta_n} \sum_{i=n}^{\beta_n+n-1} R(\tilde{\nu}_i^n(\cdot)||\rho_{\tilde{X}_i^n}(\tilde{Y}_i^n,\cdot)) \right].
\end{aligned}
$$

Equation (7.5) ensures that the second term is bounded in n. Therefore, since the first term no longer depends on n, taking $C \to \infty$, followed by $\sigma \to \infty$ yields the claimed convergence. ■

The proof Theorem 7.8 contains many steps and the idea follows closely from [9, Section 5.3]. We first prove the tightness of the controls $\{\tilde{\nu}^n\}$ which assures the existence of a convergent subsequence. We also show an important property of $\{\tilde{\nu}^n\}$ called "uniform integrability" (see Lemma 8.1). After that we consider a stochastic process S^n which is defined as

$$S^n(t) = x + \int_0^t \int_{\mathbb{R}^{d_2}} g(\hat{X}^n(s), y) \tilde{\nu}^n(dy|s) ds,$$

where $\hat{X}^n(t)$ is the picewise constant function that takes the values $\hat{X}^n(t_{n+k} - t_n) = \tilde{X}_k^n$. With this intermediate process S^n, we can then bridge the gap between $\tilde{X}_n$ and $\tilde{X}$. To be more precise, we will show that there exists a subsequence of $\{S^n\}$ (also denote it by the same notation) such that $S^n \xrightarrow{p} \tilde{X}$ and

$$P\left(\sup_{t\in[0,T]} \|\tilde{X}^n(t) - S^n(t)\| > \varepsilon \right) = 0 \text{ for any } \varepsilon > 0$$

to complete the proof. We begin with proving uniform integrability and tightness of the control sequences $\{\tilde{\nu}^n\}$.

Before we show the convergence of $\{S^n\}$, we prove that $\{S^n\}$ is tight in the following lemma.

Lemma 8.2 *Define the modulus of continuity of S^n as*

$$w^n(\delta) = \sup_{|s-t|<\delta} \|S^n(t) - S^n(s)\|.$$

for any $\delta > 0$. Then

(a) for all $\varepsilon > 0$ and $\eta > 0$ there exists a $\delta > 0$ such that $P(w^n(\delta) > \varepsilon) < \eta$ for all n.

(b) S^n is tight.

Proof. For part (a), given any $\varepsilon > 0$ and $\eta > 0$, from (8.2) we can choose $C > 0$ such that

$$\sup_n E\left[\int_0^T \int_{\|g(S^n(t),y)\|>C} \|g(\hat{X}^n(t), y)\|\tilde{\nu}^n(dy \times dt)\right] \leq \frac{\eta\varepsilon}{2e^T}.$$

Then setting $\delta = \varepsilon/(2C)$, using Markov's inequality yields

$$\begin{aligned} P(w^n(\delta) > \varepsilon) &= P\left(\sup_{|s-t|<\delta} \|S^n(t) - S^n(s)\| > \varepsilon\right) \\ &\leq P\left(\sup_{|s-t|<\delta} \int_s^t \int_{\mathbb{R}^{d_2}} \|g(\hat{X}^n(r), y)\|\tilde{\nu}^n(dy|r)dr > \varepsilon\right) \\ &\leq P\left(\sup_{|s-t|<\delta} \int_s^t \int_{\|g(\hat{X}^n(r),y)\|>C} \|g(\hat{X}^n(r), y)\|\tilde{\nu}^n(dy|r)dr\right. \\ &\qquad\qquad \left. + \sup_{|s-t|<\delta} \int_s^t \int_{\|g(\hat{X}^n(r),y)\|\leq C} \|g(\hat{X}^n(r), y)\|\tilde{\nu}^n(dy|r)dr > \varepsilon\right) \\ &\leq P\left(\sup_{|s-t|<\delta} \int_s^t \int_{\|g(\hat{X}^n(r),y)\|>C} \|g(\hat{X}^n(r), y)\|\frac{h^n(r)}{h^n(r)}\tilde{\nu}^n(dy|r)dr + C\delta > \varepsilon\right) \\ &= P\left(\sup_{|s-t|<\delta} \int_s^t \int_{\|g(\hat{X}^n(r),y)\|>C} \|g(\hat{X}^n(r), y)\|h^n(r)\tilde{\nu}^n(dy \times dr) > \frac{\varepsilon}{2T}\right) \\ &\leq P\left(\int_0^T \int_{\|g(\hat{X}^n(r),y)\|>C} \|g(\hat{X}^n(r), y)\|\tilde{\nu}^n(dy \times dr) > \frac{\varepsilon}{2e^T}\right) \\ &\leq \frac{2e^T}{\varepsilon} E\left[\int_0^T \int_{\|g(\hat{X}^n(r),y)\|>C} \|g(\hat{X}^n(r), y)\|\tilde{\nu}^n(dy \times dr)\right] < \eta. \end{aligned}$$

As for part (b), since we have part (a) and we also know that $S^n(0) = x$ for all n, the tightness of $\{S^n\}$ follows from Theorem A.3.22 in [9]. ■ We next show that $\tilde{X}^n$ and S^n are close. To be more precise, we prove the following lemma.

Lemma 8.3 *Let S^n and $\tilde{X}^n$ be defined as before, then for any $\varepsilon > 0$*

$$P\left(\sup_{t\in[0,T]} \|S^n(t) - \tilde{X}^n(t)\| > \varepsilon\right) \to 0,$$

as $n \to \infty$.

Proof. We will use the notation $t_j^n = t_j - t_n$. We first use the fact that $\tilde{X}^n$ is the piecewise linear interpolation of the random vector $\{\tilde{X}_j^n\} = \{\tilde{X}^n(t_j^n)\}$ to find

$$\begin{aligned}\sup_{t\in[0,T]} \|S^n(t) - \tilde{X}^n(t)\| &\leq \max_{k\in J} \sup_{t\in[t_k^n, t_{k+1}^n]} \|S^n(t) - \tilde{X}^n(t)\| \leq \max_{k\in J} w^n(\varepsilon_k^n) + \max_{k\in J} \|S^n(t_k^n) - \tilde{X}^n(t_k^n)\| \\ &\leq w^n(\varepsilon_n) + \max_{k\in J} \|S^n(t_k^n) - \tilde{X}^n(t_k^n)\|,\end{aligned}$$

where $J \doteq \{n, \dots, m(T + t_n)\}$. Lemma 8.2 implies that $w^n(\varepsilon_n) \xrightarrow{p} 0$. Hence it suffices to show that $\max_{k\in J} \|S^n(t_k^n) - \tilde{X}^n(t_k^n)\| \xrightarrow{p} 0$. For any $\varepsilon > 0$, we use Markov's inequality to get

$$P\left(\max_{k\in J} \|S^n(t_k^n) - \tilde{X}^n(t_k^n)\| > \varepsilon\right) \leq \frac{1}{\varepsilon} E\left[\max_{k\in J} \|S^n(t_k^n) - \tilde{X}^n(t_k^n)\|\right].$$

It remains to show the expectation converges to 0. Given $\theta > 0$, we introduce a variable Λ_j^n defined by

$$\Lambda_j^n = \begin{cases} \tilde{X}_{j+1}^n - \tilde{X}_j^n & \text{if } \|\tilde{X}_{j+1}^n - \tilde{X}_j^n\| < \theta \\ 0 & \text{if } \|\tilde{X}_{j+1}^n - \tilde{X}_j^n\| \geq \theta \end{cases}.$$

Observe that Λ_j^n is a truncation of $\varepsilon_j g(X_j^n, \bar{Y}_j^n)$, since $\tilde{X}_{j+1}^n = \tilde{X}_j^n + \varepsilon_j g(\tilde{X}_j^n, \bar{Y}_j^n)$. With this notation and by the definitions of S^n and $\tilde{X}^n$, and $\tilde{\nu}^n(dy|s) = \tilde{\nu}_j^n(dy)$ for $s \in [t_j^n, t_{j+1}^n)$, we find that

$$\begin{aligned} &E\left[\max_{k\in J} \|S^n(t_k^n) - \tilde{X}^n(t_k^n)\|\right] \\ &\leq E\left[\max_{k\in J} \left\| x + \sum_{j=n}^{k-1} \varepsilon_j \int_{\mathbb{R}^{d_2}} g(\hat{X}^n(t_j^n), y)\tilde{\nu}_j^n(dy) - \tilde{X}^n(t_k^n)\right\|\right] \\ &\leq E\left[\max_{k\in J} \left\| x + \sum_{j=n}^{k-1} \Lambda_j^n - \tilde{X}^n(t_k^n)\right\|\right] \qquad (8.3) \\ &\quad + E\left[\max_{k\in J} \left\| \sum_{j=n}^{k-1} \left(\Lambda_j^n - \varepsilon_j \int_{\mathbb{R}^{d_2}} g(\tilde{X}_j^n, y)1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| < \theta\}} \tilde{\nu}_j^n(dy)\right)\right\|\right] \\ &\quad + E\left[\max_{k\in J} \left\| \sum_{j=n}^{k-1} \varepsilon_j \left(\int_{\mathbb{R}^{d_2}} g(\tilde{X}_j^n, y)1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| < \theta\}} \tilde{\nu}_j^n(dy) - \int_{\mathbb{R}^{d_2}} g(\bar{X}_j^n, y)\tilde{\nu}_j^n(dy)\right)\right\|\right] \\ &\qquad (8.4) \end{aligned}$$

Therefore, it suffices to show that the last three terms on the right side of the inequality converge to 0 as $n \to \infty$.

For the second term, we observe that with respect to the sigma algebra $\bar{\mathcal{F}}_j^n \doteq \sigma(\tilde{X}_n^n, \ldots, \tilde{X}_{n+j}^n)$ the sequence $\{M_j^n\}$ defined by

$$M_{j+1}^n \doteq \Lambda_j^n - \varepsilon_j \int_{\mathbb{R}^{d_2}} g(\tilde{X}_j^n, y) 1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| < \theta\}} \tilde{\nu}_j^n(dy)$$

is a martingale difference sequence. Indeed, by the definition of $\tilde{\nu}_j^n$

$$\begin{aligned} E[\Lambda_j^n | \bar{\mathcal{F}}_j^n] &= E\left[(\tilde{X}_{j+1}^n - \tilde{X}_j^n) 1_{\|\tilde{X}_{j+1}^n - \tilde{X}_j^n\| < \theta} | \bar{\mathcal{F}}_j^n\right] \\ &= \varepsilon_j E\left[g(X_j^n, \bar{Y}_j^n) 1_{\{\|\varepsilon_j g(X_j^n, \bar{Y}_j^n)\| < \theta\}} \bar{\mathcal{F}}_j^n\right] \\ &= \varepsilon_j \int_{\mathbb{R}^{d_2}} g(\tilde{X}_j^n, y) 1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| < \theta\}} \tilde{\nu}_j^n(dy). \end{aligned}$$

Therefore, $\{(\sum_{j=n}^k M_j^n, \bar{\mathcal{F}}_{k-1}^n)\}$ is a martingale and for any $i \neq j$, $E\left[\left\langle M_i^n, M_j^n \right\rangle\right] = 0$. In addition, the second term becomes

$$E\left[\max_{k \in J} \left\| \sum_{j=n}^{k-1} \left(\Lambda_j^n - \varepsilon_j \int_{\mathbb{R}^{d_2}} g(\tilde{X}_j^n, y) 1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| < \theta\}} \tilde{\nu}_j^n(dy) \right) \right\| \right] = E\left[\max_{k \in J} \left\| \sum_{j=n}^{k-1} M_{j+1}^n \right\| \right].$$

We can then bound the second term by

$$E\left[\max_{k \in J} \left\| \sum_{j=n}^{k-1} M_{j+1}^n \right\| \right] \leq \left(E\left[\max_{k \in J} \left\| \sum_{j=n}^{k-1} M_{j+1}^n \right\|^2 \right] \right)^{1/2} \leq 2 \left(E\left[\left\| \sum_{j=n}^{\beta_n} M_{j+1}^n \right\|^2 \right] \right)^{1/2},$$

where the first inequality comes from Cauchy-Schwarz inequality and the second one holds since Doob's submartingale inequality. Furthermore, because

$$\begin{aligned} E\left[\left\| \sum_{j=n}^{\beta_n + n} M_{j+1}^n \right\|^2 \right] &= \sum_{j=n}^{\beta_n + n} E\left[\left\| \Lambda_j^n - \varepsilon_j \int_{\mathbb{R}^{d_2}} g(\tilde{X}_j^n, y) 1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| < \theta\}} \tilde{\nu}_j^n(dy) \right\|^2 \right] \\ &\leq \sum_{j=n}^{\beta_n + n} E\left[\|\Lambda_j^n\|^2\right] \leq \theta \sum_{j=n}^{\beta_n + n} E\left[\|\Lambda_j^n\|\right] \\ &\leq \theta \sum_{j=n}^{\beta_n + n} E\left[\int_{t_j^n}^{t_{j+1}^n} \int_{\mathbb{R}^{d_2}} \|g(\tilde{X}_j^n, y)\| \tilde{\nu}^n(dy|t) dt \right] \\ &\leq \theta E\left[\int_0^T \int_{\mathbb{R}^{d_2}} \|g(\tilde{X}_j^n, y)\| h^n(t) \tilde{\nu}^n(dy \times dt) \right], \end{aligned}$$

where the first inequality comes from the fact that

$$E[\Lambda_j^n | \bar{\mathcal{F}}_j^n] = \varepsilon_j \int_{\mathbb{R}^{d_2}} g(\tilde{X}_j^n, y) 1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| < \theta\}} \tilde{\nu}_j^n(dy),$$

and the second inequality holds due to $\|\Lambda_j^n\| \leq \theta$.

The expectation in the last display is bounded by a constant (which is independent of n) due to the uniform integrability property, so by sending $\theta \to 0$ this term disappears.

Now for the first and the third term, we need to consider $c^n(\theta)$ which is defined as

$$c^n(\theta) \doteq E\left[\int_0^T \int_{\mathbb{R}^{d_2}} \|g(\tilde{X}_j^n, y)\| 1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| > \theta\}} \tilde{\nu}^n(dy|t)dt\right].$$

Notice that as $n \to \infty, c^n(\theta) \to 0$ because of the uniform integrability property and $\varepsilon_j \to 0$. Thus, we will complete the proof by bounding the first and the third term with $c^n(\theta)$, and then we send $n \to \infty$ followed by sending $\theta \to 0$.

For the first term, we write $\tilde{X}^n(t_k^n) - x$ as a telescoping sum to find

$$\begin{aligned} E\left[\max_{k\in J}\left\|x + \sum_{j=n}^{k-1}\Lambda_j^n - \tilde{X}^n(t_k^n)\right\|\right] &\leq E\left[\sum_{j=n}^{m(T+t_n)} \|\tilde{X}^n(t_{j+1}^n) - \tilde{X}^n(t_j^n) - \Lambda_j^n\|\right] \\ &\leq E\left[\sum_{j=n}^{\beta_n+n} \varepsilon_j \|g(\tilde{X}_j^n, \bar{Y}_j^n)\| 1_{\{\|\varepsilon_j g(\tilde{X}_j^n, \bar{Y}_j^n)\| \geq \theta\}}\right] \\ &= E\left[\sum_{j=n}^{\beta_n+n} \int_{t_j^n}^{t_{j+1}^n} \int_{\mathbb{R}^{d_2}} \|g(\tilde{X}_j^n, y)\| 1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| \geq \theta\}} \tilde{\nu}_j^n(dy)dt\right] \\ &\leq c^n(\theta). \end{aligned}$$

As for the third term

$$\begin{aligned} &E\left[\max_{k\in J}\left\|\sum_{j=n}^{k-1}\varepsilon_j\left(\int_{\mathbb{R}^{d_2}} g(\tilde{X}_j^n, y) 1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| < \theta\}} \tilde{\nu}_j^n(dy) - \int_{\mathbb{R}^{d_2}} g(\tilde{X}_j^n, y)\tilde{\nu}_j^n(dy)\right)\right\|\right] \\ &\leq E\left[\sum_{j=n}^{m(T+t_n)} \varepsilon_j \int_{\mathbb{R}^{d_2}} \|g(\tilde{X}_j^n, y)\| 1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| \geq \theta\}} \tilde{\nu}_j^n(dy)\right] \\ &= E\left[\sum_{j=n}^{\beta_n+n} \int_{t_j^n}^{t_{j+1}^n} \int_{\mathbb{R}^{d_2}} \|g(\tilde{X}_j^n, y)\| 1_{\{\|\varepsilon_j g(\tilde{X}_j^n, y)\| \geq \theta\}} \tilde{\nu}^n(dy|t)dt\right] \\ &\leq c^n(\theta). \end{aligned}$$

The proof is complete. ■

Now we know that there is a convergent subsequence of $\{S^n\}$ according to Prohorov's theorem, and if we denote its limit as S, then it remains to show that with probability one (w.p.1) S equals $\tilde{X}$, where $\tilde{X}$ is defined in (7.6).

Lemma 8.4 *Let S^n and $\tilde{X}$ be defined as before, then for any $\varepsilon > 0$*

$$P\left(\sup_{t\in[0,T]} \|S^n(t) - \tilde{X}(t)\| > \varepsilon\right) \to 0,$$

as $n \to \infty$.

Proof. For any $\varepsilon > 0$, we first use Markov's inequality to find

$$P\left(\sup_{t\in[0,T]} \|S^n(t) - \tilde{X}(t)\| > \varepsilon\right) \leq \frac{1}{\varepsilon} E\left[\sup_{t\in[0,T]} \|S^n(t) - \tilde{X}(t)\|\right].$$

It remains to show the expectation converges to 0. To show this, we notice that

$$\begin{aligned}
&E\left[\sup_{t\in[0,T]} \|S^n(t) - \tilde{X}(t)\|\right] \\
&= E\left[\sup_{t\in[0,T]} \left\|\int_0^t \int_{\mathbb{R}^{d_2}} g(\hat{X}^n(s), y) d\tilde{\nu}^n(y|s)ds - \int_0^t \int_{\mathbb{R}^{d_2}} g(\tilde{X}(s), y) d\tilde{\nu}(y|s)ds\right\|\right] \\
&\leq E\left[\sup_{t\in[0,T]} \left\|\int_0^t \int_{\mathbb{R}^{d_2}} g(\hat{X}^n(s), y) d\tilde{\nu}^n(y|s)ds - \int_0^t \int_{\mathbb{R}^{d_2}} g(\hat{X}^n, y) d\tilde{\nu}(y|s)ds\right\|\right] \\
&\quad + E\left[\sup_{t\in[0,T]} \int_0^t \int_{\mathbb{R}^{d_2}} \left\|g(\hat{X}^n, y) - g(\tilde{X}(s), y)\right\| d\tilde{\nu}(y|s)ds\right] \\
&\leq E\left[\sup_{t\in[0,T]} \int_{\mathbb{R}^{d_2}} \|g(\hat{X}^n, y)\|\|h^n(s) - h(s)\|\tilde{\nu}^n(d \times ds)\right] \\
&\quad + E\left[\sup_{t\in[0,T]} \left\|\int_{\mathbb{R}^{d_2}\times[0,t]} g(\hat{X}^n, y)h(s)\tilde{\nu}^n(dy \times ds) - \int_{\mathbb{R}^{d_2}\times[0,t]} g(\hat{X}^n, y)h(s)\tilde{\nu}(dy \times ds)\right\|\right] \\
&\quad + E\left[\sup_{t\in[0,T]} \int_0^t \int_{\mathbb{R}^{d_2}} K\|\hat{X}^n(s) - \tilde{X}(s)\| d\tilde{\nu}(y|s)ds\right] \\
&\leq E\left[\sup_{t\in[0,T]} \int_{\mathbb{R}^{d_2}} \|g(\hat{X}^n, y)\|\|h^n(s) - h(s)\|\tilde{\nu}^n(dy \times ds)\right] && (8.5) \\
&\quad + E\left[\sup_{t\in[0,T]} \left\|\int_{\mathbb{R}^{d_2}\times[0,t]} g(\hat{X}^n, y)h(s)\tilde{\nu}^n(dy \times ds) - \int_{\mathbb{R}^{d_2}\times[0,t]} g(\hat{X}^n, y)h(s)\tilde{\nu}(dy \times ds)\right\|\right] && (8.6) \\
&\quad + E\left[\sup_{t\in[0,T]} \int_0^t \int_{\mathbb{R}^{d_2}} K\|\hat{X}^n(s) - S^n(s)\| d\tilde{\nu}(y|s)ds\right] && (8.7) \\
&\quad + E\left[\sup_{t\in[0,T]} \int_0^t \int_{\mathbb{R}^{d_2}} K\|S^n(s) - \bar{X}(s)\| d\tilde{\nu}(y|s)ds\right].
\end{aligned}$$

By Grönwald's inequality we now get that

$$E\left[\sup_{t\in[0,T]} \|S^n(t) - \tilde{X}(t)\|\right] \leq E\left[((8.5) + (8.6) + (8.7))e^{KT}\right].$$

All that is left is to prove that (8.5), (8.6) and (8.7) converges to zero. For (8.5), due to (8.2), we know that there exists some $C > 0$ such that

$$\sup_n E\left[\int_0^T \int_{\|g(\hat{X}^n(t),y)\|>C} \|g(x, y)\|\tilde{\nu}^n(dy \times dt)\right] \leq 1.$$

With this $C > 0$, we can find that

$$\begin{aligned}
&E\left[\sup_{t\in[0,T]}\int_{\mathbb{R}^{d_2}\times[0,t]}\|g(\hat{X}^n(s),y)\|\|h^n(s)-h(s)\|\tilde{\nu}^n(dy\times ds)\right]\\
&\quad\le E\left[\int_0^T\int_{\|g(\hat{X}^n(s),y)\|\le C}\|g(\hat{X}^n(s),y)\|\|h^n(s)-h(s)\|\tilde{\nu}^n(dy\times ds)\right]\\
&\qquad+E\left[\int_0^T\int_{\|g(\hat{X}^n(s),y)\|> C}\|g(\hat{X}^n(s),y)\|\|h^n(s)-h(s)\|\tilde{\nu}^n(dy\times ds)\right]\\
&\quad\le (CT+1)\sup_{t\in[0,T]}\|h^n(t)-h(t)\|.
\end{aligned}$$

Thus, (8.5) converges to 0 due to the uniform converge of $\{h^n\}$ to h as $n \to \infty$.

As for (8.6), it requires more analysis. Notice that for any $C > 0$ and $n_0 > 0$,

$$\begin{aligned}
&E\left[\sup_{t\in[0,T]}\left\|\int_{\mathbb{R}^{d_2}\times[0,t]}g(\hat{X}^n(s),y)h(s)\tilde{\nu}^n(dy\times ds)-\int_{\mathbb{R}^{d_2}\times[0,t]}g(\hat{X}^n(s),y)h(s)\tilde{\nu}(dy\times ds)\right\|\right]\\
&\le E\left[\sup_{t\in[0,T]}\left\|\int_{\mathbb{R}^{d_2}\times[0,t]}g(\tilde{X}^{n_0}(s),y)h(s)\tilde{\nu}^n(dy\times ds)-\int_{\mathbb{R}^{d_2}\times[0,t]}g(\tilde{X}^{n_0}(s),y)h(s)\tilde{\nu}(dy\times ds)\right\|\right]\\
&+E\left[\sup_{t\in[0,T]}\left\|\int_{\mathbb{R}^{d_2}\times[0,t]}g(\hat{X}^n(s),y)h(s)\tilde{\nu}^n(dy\times ds)-\int_{\mathbb{R}^{d_2}\times[0,t]}g(\tilde{X}^{n_0}(s),y)h(s)\tilde{\nu}^n(dy\times ds)\right\|\right]\\
&+E\left[\sup_{t\in[0,T]}\left\|\int_{\mathbb{R}^{d_2}\times[0,t]}g(\tilde{X}^{n_0}(s),y)h(s)\tilde{\nu}(dy\times ds)-\int_{\mathbb{R}^{d_2}\times[0,t]}g(\hat{X}^n(s),y)h(s)\tilde{\nu}(dy\times ds)\right\|\right]\\
&\le E\left[\sup_{t\in[0,T]}\left\|\int_{\mathbb{R}^{d_2}\times[0,t]}g(\tilde{X}^{n_0}(s),y)h(s)\tilde{\nu}^n(dy\times ds)-\int_{\mathbb{R}^{d_2}\times[0,t]}g(\tilde{X}^{n_0}(s),y)h(s)\tilde{\nu}(dy\times ds)\right\|\right]\\
&+E\left[\sup_{t\in[0,T]}\int_{\mathbb{R}^{d_2}\times[0,t]}\|\hat{X}^n(s)-\bar{X}^{n_0}(s)\|h(s)\tilde{\nu}^n(dy\times ds)\right]\\
&+E\left[\sup_{t\in[0,T]}\int_{\mathbb{R}^{d_2}\times[0,t]}\|\bar{X}^{n_0}(s)-\hat{X}^{n_0}(s)\|h(s)\tilde{\nu}(dy\times ds)\right]
\end{aligned}$$

The last two terms can be made arbitrarily small by choosing large enough n and n_0 due to $\tilde{X}^n$ and $\hat{X}^n$ converging to the same process. For the first term we split it up into two parts, one that is bounded where we can use the weak convergence of $\tilde{\nu}^n$ and one part that can be made arbitrarily small due to uniform integrability property.

$$
\begin{aligned}
&E\left[\sup_{t\in[0,T]}\left\|\int_{\mathbb{R}^{d_2}\times[0,t]} g(\tilde{X}^{n_0}(s),y)h(s)\tilde{\nu}^n(dy\times ds)-\int_{\mathbb{R}^{d_2}\times[0,t]} g(\tilde{X}^{n_0}(s),y)h(s)\tilde{\nu}(dy\times ds)\right\|\right]\\
&\le E\left[\sup_{t\in[0,T]}\left\|\int_{\mathbb{R}^{d_2}\times[0,t]}\Big(g(\tilde{X}^{n_0}(s),y)1_{\{\|g(\tilde{X}^{n_0}(s),y)\|\le C\}}\right.\right.\\
&\qquad\qquad\qquad +C\frac{g(\tilde{X}^{n_0}(s),y)}{\|g(\tilde{X}^{n_0}(s),y)\|}1_{\{\|g(\tilde{X}^{n_0}(s),y)\|>C\}}\Big)h(s)\tilde{\nu}^n(dy\times ds)\\
&\qquad\quad -\int_{\mathbb{R}^{d_2}\times[0,t]}\Big(g(\tilde{X}^{n_0}(s),y)1_{\{\|g(\tilde{X}^{n_0}(s),y)\|\le C\}}\\
&\qquad\qquad\qquad \left.\left.+C\frac{g(\tilde{X}^{n_0},y)}{\|g(\tilde{X}^{n_0}(s),y)\|}1_{\{\|g(\tilde{X}^{n_0}(s),y)\|>C\}}\Big)h(s)\tilde{\nu}(dy\times ds)\right\|\right]\\
&\quad +E\left[\sup_{t\in[0,T]}\left\|\int_{\mathbb{R}^{d_2}\times[0,t]}\left(g(\tilde{X}^{n_0}(s),y)-C\frac{g(\tilde{X}^{n_0}(s),y)}{\|g(\tilde{X}^{n_0}(s),y)\|}\right)1_{\{\|g(\tilde{X}^{n_0}(s),y)\|>C\}}h(s)\tilde{\nu}^n(dy\times ds)\right.\right.\\
&\quad \left.\left.-\int_{\mathbb{R}^{d_2}\times[0,t]}\left(g(\tilde{X}^{n_0}(s),y)-C\frac{g(\tilde{X}^{n_0}(s),y)}{\|g(\tilde{X}^{n_0}(s),y)\|}\right)1_{\{\|g(\tilde{X}^{n_0}(s),y)\|>C\}}h(s)\tilde{\nu}(dy\times ds)\right\|\right].
\end{aligned}
$$

For the new first term, we define notations

$$\varphi_C(x)\doteq x1_{\{\|x\|\le C\}}+C\frac{x}{\|x\|}1_{\{\|x\|>C\}}$$

and

$$\ell^n(t)\doteq\int_{\mathbb{R}^{d_2}\times[0,t]}\varphi_C(g(\tilde{X}^{n_0}(s),y))h(s)\tilde{\nu}^n(dy\times ds)-\int_{\mathbb{R}^{d_2}\times[0,t]}\varphi_C(g(\tilde{X}^{n_0}(s),y))h(s)\tilde{\nu}(dy\times ds).$$

Then the new first term can then be expressed as $E[\sup_{t\in[0,T]}\|\ell^n(t)\|]$. Now given any $t\in[0,T]$, since $\tilde{\nu}(\mathbb{R}^{d_2}\times\{t\})=0$, $\tilde{\nu}^n$ converges weakly to $\tilde{\nu}$ w.p.1, and notice that φ_C is bounded and continuous, this implies that $\ell^n(t)\to 0$ w.p.1 as $n\to\infty$. Without loss of generality, we assume $\ell^n(t,\omega)\to 0$ for all $\omega\in\Omega$ and $t\in[0,T]$. Now for any fixed $\omega\in\Omega$, it is not hard to see that $\{\ell^n(t,\omega):t\in[0,T]\}_{n\in\mathbb{N}}$ is equicontinuous and uniformly bounded by $2C$. Thus, by Arzelà-Ascoli theorem and since $\ell^n(t,\omega)\to 0$, also the fact that if every subsequence has a further subsequence which converges uniformly to the same limit, then the whole sequence converges uniformly to the same limit, we can conclude that $\sup_{t\in[0,T]}\|\ell^n(t,\omega)\|\to 0$ for that given ω. Since ω is arbitrary, this means that $\sup_{t\in[0,T]}\|\ell^n(t)\|\to 0$ w.p.1. Moreover, due to the fact that $\sup_{t\in[0,T]}\|\ell^n(t)\|\le 2C<\infty$, we use the Lebesgue dominated convergence theorem to find that the new first term, i.e. $E[\sup_{t\in[0,T]}\|\ell^n(t)\|]$, converges to 0.

As for the new second term, we can use that $h(s)<e^T$ and

$$\left\|x-C\frac{x}{\|x\|}\right\|1_{\{\|x\|>C\}}\le(\|x\|+C)1_{\{\|x\|>C\}}\le 2\|x\|1_{\{\|x\|>C\}}$$

for all x and (8.2) to bound it from above by

$$4e^T \sup_n E\left[\int_0^T \int_{\|g(\bar{X}^{n_0}(t),y)\|>C} \|g(x,y)\|\tilde{\nu}^n(dy \times dt)\right],$$

which converges to 0 by sending $C \to \infty$.

Finally we have the term (8.7) that converges to zero due to Lemma 8.3. ■

Now the proof of Theorem 7.8 is complete. The tightness of $\{\tilde{\nu}^n\}$ follows from Lemma 8.1, the tightness of $\{\tilde{X}^n\}$ follows from Lemma 8.2 and Lemma 8.3 and finally the limit (7.6) follows from Lemma 8.4 and 8.3.

A Appendix

A.1 Proof I is a rate function

Lemma A.1 *Under Conditions 2.2, The function $I : C[0,T] \to [0,\infty]$ defined by*

$$I(\varphi) = \int_0^T \frac{1}{h(t)} L(\varphi(t), \dot{\varphi}(t))dt,$$

where

$$L(x,\beta) = \inf_{\mu \in \mathcal{P}(\mathbb{R}^{d_2})} \left\{ \inf_{\eta \in \mathcal{A}(\mu)} \{R(\eta \| \mu \otimes \rho_x(\cdot,\cdot))\} : \beta = \int_{\mathbb{R}^{d_2}} g(x,y)\mu(dy) \right\},$$

and

$$h(t) = (e^T - 1)e^{-t},$$

is a rate function, i.e., I has compact level sets.

Proof. We need to show that for any sequence of continuous functions $\{\phi^j\}$ such that $I(\phi^j) \leq K$ have a convergent subsequence where the corresponding limit ϕ fulfills $I(\phi) \leq K$. For any ε, there exist a probability measure $\mu^j(dy \times dt) = \mu^j(dy|t)dt$ and a transition kernel $q^j(y, dz|t)$ such that

$$\begin{aligned}
&\int_0^T \frac{1}{h(t)} \int_{\mathbb{R}^{d_2}} R(q^j(y,\cdot|t)||\rho_{\phi^j(t)}(y,\cdot))\mu^j(dy|t)dt \leq I(\phi^j) + \varepsilon \\
&\int_{\mathbb{R}^{d_2}} g(\phi^j(t), y)\mu^j(dy|t) = \dot{\phi}^j(t) \\
&\mu^j q^j = \mu^j
\end{aligned}$$

Now we prove that $\{\mu^j\}_j$ is uniformly integrable. It is sufficient to prove that

$$\lim_m \limsup_n \int_0^T \int_{\mathbb{R}^{d_2}} \int_{\|z\|>M} \|z\| q^j(y, dz|t)\mu^j(dy \times dt) = 0$$

We use the inequality $ab \leq e^{\sigma a} + \frac{1}{\sigma}(b\log(b) - b + 1)$

$$\begin{aligned}
&\int_0^T \int_{\mathbb{R}^{d_2}} \int_{\|z\|>M} \|z\| q^j(y,dz)\mu^j(dy \times dt) \\
&\leq \int_0^T \int_{\mathbb{R}^{d_2}} \int_{\|z\|>M} e^{\sigma\|z\|}\rho_{\phi^j(t)}(y,dz)\mu^j(dy|t)dt + \frac{1}{\sigma}\int_0^T R(q^j(y,dz|t)\mu^j(dy|t)||\rho_{\phi^j(t)}(y,dz)\mu^j(dy|t))dt \\
&\leq \int_0^T \sup_y \int_{\|z\|>M} e^{\sigma\|z\|}\rho_{\phi^j(t)}(y,dz)dt + \frac{1}{\sigma}(K+\varepsilon) \\
&\leq \int_0^T e^{-\sigma M} \sup_x \sup_y \int_{\|z\|>M} e^{2\sigma\|z\|}\rho_x(y,dz)dt + \frac{1}{\sigma}(K+\varepsilon).
\end{aligned}$$

Now as in Lemma7.4 and Lemma 8.1, sending $n \to \infty$, $M \to \infty$ and then $\sigma \to \infty$ yields the desired limit. This proves that μ^j is tight. Now for a convergent subsequence of μ^j with limit μ. We define ϕ as the solution to the following ODE

$$\phi(t) = x_0 + \int_0^t g(\phi(s,y))\mu(dy|s)ds.$$

We need to prove that this ODE has a unique solution.

Lemma A.2 *The ODE*

$$\phi(t) = x_0 + \int_0^t \int g(\phi(s),y)\nu^{\zeta(s),\dot{\zeta}(s)}(dy)ds,$$

has a unique solution.

Proof. Let ϕ^1 and ϕ^2 be solutions to the ODE

$$\phi^1(t) = x_0 + \int_0^t \int g(\phi^1(s),y)\nu^{\zeta(t),\dot{\zeta}(t)}(dy)ds,$$

$$\phi^2(t) = x_0 + \int_0^t \int g(\phi^2(s),y)\nu^{\zeta(s),\dot{\zeta}(s)}(dy)ds,$$

and let Δ be such that $\Delta K < 1$, where K is the Lipschitz constant to g. Then we will prove that for $t \in [0,\Delta]$, $\phi^1(t) = \phi^2(t)$.

$$\begin{aligned}
&\sup_{t\in[0,\Delta]} ||\phi^1(t) - \phi^2(t)|| = \sup_{t\in[0,\Delta]} || \int_0^t \int g(\phi^1(s),y)\nu^{\zeta(s),\dot{\zeta}(s)}(dy)ds - \int_0^t \int g(\phi^2(s),y)\nu^{\zeta(s),\dot{\zeta}(s)}(dy)ds|| \\
&\leq \int_0^\Delta \int \sup_{t\in[0,\Delta]} ||g(\phi^1(t),y) - g(\phi^2(t),y)||\nu^{\zeta(s),\dot{\zeta}(s)}(dy)ds \\
&\leq \int_0^\Delta \int K \sup_{t\in[0,\Delta]} ||\phi^1(t) - \phi^2(t)||\nu^{\zeta(s),\dot{\zeta}(s)}(dy)ds = K\Delta \sup_{t\in[0,\Delta]} ||\phi^1(t) - \phi^2(t)||.
\end{aligned}$$

This is a contraction and the same procedure can be iterated arbitrary number of times leading to

$$\sup_{t\in[0,\Delta]} ||\phi^1(t)-\phi^2(t)|| \leq (K\Delta)^N \sup_{t\in[0,\Delta]} ||\phi^1(t)-\phi^2(t)|| \to 0, \quad N\to\infty.$$

Now we need to extend this to $t\in[0,T]$. For $t\in[0,2\Delta]$ we can use the above argument to get

$$\begin{aligned}\sup_{t\in[0,2\Delta]} ||\phi^1(t)-\phi^2(t)|| &= \sup_{t\in[0,2\Delta]} ||\phi^1(t)-\phi^1(\Delta)-(\phi^2(t)-\phi^2(\Delta))|| \\ &= \sup_{t\in[\Delta,2\Delta]} ||\phi^1(t)-\phi^1(\Delta)-(\phi^2(t)-\phi^2(\Delta))|| \\ &\leq \Delta K \sup_{t\in[\Delta,2\Delta]} ||\phi^1(t)-\phi^2(t)|| \leq \Delta K \sup_{t\in[0,2\Delta]} ||\phi^1(t)-\phi^2(t)||\end{aligned}$$

The same argument as above can now be applied to show that $\phi^1(t)=\phi^2(t)$ for $t\in[0,2\Delta]$. Repeating this procedure yields the result for $t\in[0,T]$. ■

A contradiction argument with different ε proves that ϕ is independent of ε. This proves that there exists a subsequence of ϕ^j that converges and therefore is precompact. The proof is finished if we can show $I(\phi)\leq K$.

$$\begin{aligned}K &\geq \liminf I(\phi^j) \geq \liminf \int_0^T R(q^j(y,dz|t)\mu^j(dy|t)||\rho_{\phi^j(t)}(y,dz)\mu^j(dy|t))dt - \varepsilon \\ &\geq \int_0^T R(q(y,dz|t)\mu(dy|t)||\rho_{\phi(t)}(y,dz)\mu(dy|t))dts \geq I(\phi)-\varepsilon,\end{aligned}$$

where the second inequality uses Fatou's lemma, the lower semi-continuity of the relative entropy and the feller property of ρ. Since ε is arbitrary we have $I(\phi)\leq K$ and therefore I has compact level sets and is a rate function. ■

References

[1] Y.F. AtchadÃ© and J.S. Liu. The wang-landau algorithm in general state spaces: Applications and convergence analysis. *Statistica Sinica*, 20(1):209–233, 2010.

[2] V.S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint.* Cambridge University Press, 2008.

[3] A. Budhiraja and P. Dupuis. Simple necessary and sufficient conditions for the stability of constrained processes. *SIAM J. on Applied Math.*, 59:1686–1700, 1999.

[4] A. Budhiraja and P. Dupuis. *Analysis and Approximation of Rare Events: Representations and Weak Convergence Methods.* Number 94 in Probability Theory and Stochastic Modelling. Springer-Verlag, New York, 2019.

[5] I.H. Dinwoodie and P. Ney. Occupation measures for markov chains. *Journal of Theoretical Probability*, 8:679–691, 1995.

[6] M.D. Donsker and S.R.S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Comm. Pure Appl. Math.*, 28:1–47, 1975.

[7] M. Duflo and S. Wilson. *Random Iterative Models.* Springer-Verlag, Berlin, Heidelberg, 1st edition, 1997.

[8] P. Dupuis. Large deviations analysis of some recursive algorithms with state dependent noise. *The Annals of Probab.*, 16:1509–1536, 1988.

[9] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations.* John Wiley & Sons, New York, 1997.

[10] P. Dupuis and R.S. Ellis. Large deviations for Markov processes with discontinuous statistics, II: Random walks. *Probab. Th. Rel. Fields*, 91:153–194, 1992.

[11] P. Dupuis and H. J. Kushner. Stochastic approximation via large deviations: Asymptotic properties. *SIAM J. Control Optimization*, 23:675–696, 1985.

[12] M. I. Freidlin. Fluctuations in dynamical systems with averaging. *Dokl. Akad. Nauk SSSR*, 226:273–276, 1976.

[13] M. I. Freidlin. The averaging principle and theorems on large deviations. *Russ. Math. Surv.*, 33(July-Dec), 1978.

[14] I. Iscoe, P. Ney, and E. Nummelin. Large deviations of uniformly recurrent Markov additive processes. *Adv. Appl. Math.*, 6:373–412, 1985.

[15] J. Kiefer and J. Wolfowitz. Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics*, 23(3):462 – 466, 1952.

[16] H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications.* Stochastic Modelling and Applied Probability. Springer New York, 2003.

[17] H. J. Kushner. Asymptotic behavior of stochastic approximation and large deviations. *IEEE Trans. Automat. Contr.*, AC-29(11):984–990, 1984.

[18] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.