Large deviations and parameter estimations for small noise diffusion processes

Hu, Y.

DOI
10.4233/uuid:4cdc69a7-68ea-4b6a-ab0d-fc82ad3be445

Publication date
2025

Document Version
Final published version

Citation (APA)
Hu, Y. (2025). *Large deviations and parameter estimations for small noise diffusion processes*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:4cdc69a7-68ea-4b6a-ab0d-fc82ad3be445

Large deviations and parameter estimations for small noise diffusion processes

Yanyan HU

Large deviations and parameter estimations for small noise diffusion processes

Dissertation

for the purpose of obtaining the degree of doctor
at the Delft University of Technology
by the authority of the Rector Magnificus prof.dr.ir.T.H.J.J. van der Hagen
Chair of the Board for Doctorates
to be defended publicly on
Monday 20 January 2025 at 12:30 o'clock

by

Yanyan HU

Master of Science in Mathematics, Central South University, China
born in Shandong, China

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	Chairperson
Prof. dr. F.H.J. Redig	Delft University of Technology, promotor
Dr. R.C. Kraaij	Delft University of Technology, copromotor
Prof. dr. F.B. Xi	Beijing Institute of Technology, external advisor

Independent members:

Prof. dr. M.R.H. Mandjes	University of Leiden
Prof. dr. ir. F.H. van der Meulen	VU University Amsterdam
Prof. dr. A. Papapantoleon	Delft University of Technology
Dr. ir. R. Versendaal	Delft University of Technology
Prof. dr. ir. G. Jongbloed	Delft University of Technology, reserve member

The doctoral research has been carried out in the context of an agreement on joint doctoral supervision between Delft University of Technology, the Netherlands, and Beijing Institute of Technology, China.

Printed by: Gideprint
Cover by: Yanyan Hu

ISBN 978-94-6496-315-1

An electronic version of this dissertation is available at:

`http://repository.tudelft.nl/`

Contents

Chapter 1

Introduction

This thesis focuses on large deviations and parameter estimations for diffusion processes with small noise. In the real world, the small noise environment is often considered ideal due to the inevitability of random noise. By reducing the impact of noise, the law of large numbers (LLN) provides a theoretical basis for understanding the behavior of complex systems as they evolve. The LLN suggests that, despite the existence of random noise, the average behavior of the system converges to the expected behavior over time.

Based on the above analysis, small noise for a system is needed in real applications. In this thesis, we mainly:

- prove large deviation principles for slow-fast processes, where the slow process is a class of diffusion processes with small noise, and the fast process is a switching process;
- prove the consistency and asymptotic normality of estimators for small noise diffusion processes.

In the following, we start with the law of large numbers, the central limit theorem, and the large deviations theory of random variables.

1.1 Classical probability limits

Probability theory is a branch of mathematics that investigates the probabilities associated with random phenomena. Random phenomena refer to objective events where the outcome cannot be predetermined by individuals; instead, these events can result in any one of multiple possible outcomes.

In both the natural world and human society, there are numerous random phenomena. For example, when tossing a coin, it may land on heads or tails; when measuring the length of an object, different results may be obtained due to variations in instruments and environmental factors; when manufacturing light bulbs under the same production conditions, the lifespan of the bulbs may vary. All these instances represent random phenomena.

In fact, people have gradually realized through long-term practical experience that although the occurrence of an event in an experiment is accidental, a large number of repeated experiments under the same conditions show obvious regularity. This is because the random errors of each trial will cancel each other out by averaging over multiple repetitions. For example, if an unbiased coin is tossed many times, the frequency of heads coming up gradually approaches the probability $1/2$ as the number of flips increases; when the length of the same object is measured many times, the average value of the measurement results gradually approaches the true length of the object with the increase of the number of measurements. The (strong) law of large numbers and the central limit theorem are results that describe and demonstrate these regular events (i.e., frequencies close to the mean).

In the simplest setup, we first consider mutually independent and identically distributed (i.i.d.) random variables X_1, $X_2, \ldots$ on the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$. Its mean and standard deviation are u and σ, respectively. For each n, let

$$S_n = \sum_{i=1}^{n} X_i \tag{1.1}$$

represent the sum of random variables $X_1, X_2, \ldots, X_n$. The average $\frac{1}{n}S_n$ represents the frequency of events occurring.

The Strong Law of Large Numbers (SLLN) states that the average of many independent samples is close to the mean of a single sample:

- **strong law of large numbers**

$$\frac{1}{n}S_n \to u \quad \mathbb{P} - a.s. \text{ as } n \to \infty.$$

Compared with the SLLN, the Central Limit Theorem (CLT) further considers the deviations between the average of multiple independent samples and the mean of a single sample:

- **central limit theorem**

$$\sqrt{n}\left(\frac{1}{n}S_n - u\right) \Rightarrow \sigma N(0,1) \quad \text{in} \quad \text{law w.r.t. } \mathbb{P} \text{ as } n \to \infty.$$

In contrast, the Large Deviation Principle (LDP) quantifies deviations at a large distance occurrence thus being rare events (i.e., frequencies far from the mean):

- **large deviation principle**

$$\mathbb{P}\left(\frac{1}{n}S_n \approx \gamma\right) \sim e^{-nI(\gamma)}, \quad \gamma \neq u,$$

where

$$I(\gamma) = \sup_{\lambda \in \mathbb{R}}\{\lambda\gamma - \log M(\lambda)\} \tag{1.2}$$

and $M(\lambda) = \mathbb{E}[e^{\lambda X_1}]$, $\lambda \in \mathbb{R}$, is the moment generating function of X_1. The LDP mentioned above is Cramér theorem; see [DZ98, Theorem 2.2.3] for the proof.

The LDP describes events where the average of many independent samples diverges from the mean of a single sample at an exponential rate, which is the probability of rare events. Precisely calculating the probability of these rare events is crucial in the fields of statistics, information theory, statistical physics, financial mathematics, and so on.

1.2 Random walks and Brownian motions

In the previous section, we provided a framework for the emergence and description of classical probability limits: SLLN, CLT, and LDP for random variables. However, many random phenomena evolve, such as fluctuations in stocks and exchange rates, sound signals, video signals, body temperature changes, and so on. This evolution can be described by stochastic processes. Continuous time stochastic processes can be divided into continuous time stochastic processes with discrete states, random walks, and continuous time stochastic processes with continuous states, Brownian motion. Random walks and Brownian motion are essential concepts in probability theory. They have numerous applications across various fields and provide a basis for understanding the behavior of stochastic processes.

A simple random walk describes the movement of a particle that moves randomly in any direction with equal probability. It is also known as a drunkard's walk. Simple random walks have many applications in computer science, where they are used to model the behavior of random algorithms and the spread of computer viruses.

Brownian motion is the most famous stochastic process coming from physical phenomena. Let us begin with an overview of the development of Brownian motion.

In 1827, Robert Brown, a British botanist, observed under a microscope that pollen particles suspended in water were constantly moving in irregular motion. Although the phenomenon had been observed before, Brown was the first to conduct a systematic scientific study of it. Therefore, this movement came to be known as Brownian motion.

Some 50 years later, in 1877, Joseph Delsaulx correctly pointed out that Brownian motion is caused by particles suspended in a liquid being unbalanced by the collisions of surrounding molecules. But this is just a description, not a theory, and not experimentally proven. It was not until 1905 that Einstein published the paper "On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat" [Ein05], which was the first theoretical and quantitative study of Brownian motion. Informally, he described Brownian motion as

$$\mathrm{d}X(t) = \text{“}\Gamma(t)\mathrm{d}t\text{”}, \tag{1.3}$$

where $\Gamma(t)$ represents the "swelling and falling force" of a unit mass of pollen particles when colliding with liquid molecules.

On the one hand, French physicist Langevin and others found that using Einstein's Brownian motion model (1.3) to describe the random movement of tiny pollen particles observed by Brown in liquid was not entirely satisfactory. In 1908, three years after Einstein's landmark paper, Langevin published another groundbreaking paper. In this paper, he summarized Einstein's theory and developed a new equation to describe Brownian motion:

$$\mathrm{d}X(t) = -\theta X(t)\mathrm{d}t + \text{“}\Gamma(t)\mathrm{d}t\text{”}, \tag{1.4}$$

where θ represents the damping coefficient per unit mass, and $X(t)$ is the speed which a particle moves in the liquid.

On the other hand, although Einstein did not develop a general theory of Brownian motion, his work influenced American mathematician Norbert Wiener. In 1923, Wiener constructed a stochastic process, $W(t)$, based on Einstein's equation to describe Brownian motion, also known as the Wiener process. It is uniquely determined by the following characteristics:

(1) $W(t) - W(s)$ and $\{W(r)\}_{0\le r\le s}$ are independent;

(2) The law of $W(t) - W(s)$ is $N(0, t-s)$ for $t > s$.

Wiener thus determined the distribution of Brownian motion, giving rise to the Wiener measure. It is a probability measure on the path space $C([0,\infty);\mathbb{R})$, supported on the set of trajectories that are everywhere continuous but nowhere differentiable.

In mathematics, we can replace "$\Gamma(t)\mathrm{d}t$" in (1.3) and (1.4) to be $\mathrm{d}W(t)$. Hence, we get

$$\mathrm{d}X(t) = \mathrm{d}W(t)$$

and

$$\mathrm{d}X(t) = -\theta X(t)\mathrm{d}t + \mathrm{d}W(t), \tag{1.5}$$

respectively. The second equation is the prototype of the stochastic differential equation, named Ornstein-Uhlenbeck (OU) process.

1.3 Limit theorems of random walks

In this section, we aim to extend the limit theorems for random variables discussed in Section 1.1 to the context of random walks, in parallel.

Recall the definition of S_n defined in (1.1), similarly define the stochastic process

$$S_{\lfloor nt \rfloor} = \sum_{i=1}^{\lfloor nt \rfloor} X_i, \quad t \geq 0$$

where $\lfloor nt \rfloor$ is the integer part of nt. The process $\{S_{\lfloor nt \rfloor} : t \geq 0\}$ records the discrete-time random walk $\{S_n : n = 0, 1, 2 \ldots\}$ on a continuous time scale such that in one unit ($t = 1$) of continuous time there will be the contribution from n discrete time units.

In this setting, for $S_{\lfloor nt \rfloor}$, we have the Functional Strong Law of Large Number (FSLLN):

- **functional strong law of large number**
  For $t \geq 0$,
$$\frac{1}{n} S_{\lfloor nt \rfloor} \to ut, \quad \mathbb{P} - a.s. \ \text{ as } \ n \to \infty.$$

We further know the associated CLT, and start to short analysis before we give the results. We scale random walk $S_{\lfloor nt \rfloor}$ to $X^{(n)}(t) = \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor}$. The process $X_t^{(n)} = \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor}$ further scales distance in such a way that one unit of distance in the new scale equals $\sqrt{n}$ spatial units used for the random walk. This is a convenient normalization since (for large n)

$$\mathbb{E}(X^{(n)}(t)) = 0, \quad \mathbb{V}(X^{(n)}(t)) = \frac{\lfloor nt \rfloor \sigma^2}{n} \approx \sigma^2 t.$$

Since the sample paths of $X^{(n)} = \{X^{(n)}(t) : t \geq 0\}$ have jumped (though small for large n) and are, therefore, discontinuous, it is technically more convenient to linearly interpolate the random walk between one jump point and the next, using the same space-time scales as used for $\{X^{(n)}(t) : t \geq 0\}$. The resulting polygonal process $\tilde{X}^{(n)} := \{\tilde{X}^{(n)} : t \geq 0\}$ is formally defined by

$$\tilde{X}^{(n)} = \frac{S_{\lfloor nt \rfloor}}{\sqrt{n}} + (nt - \lfloor nt \rfloor) \frac{S_{\lfloor nt \rfloor} + 1}{\sqrt{n}}, \quad t \geq 0.$$

In this way, just as for the limiting Brownian motion process $\{\tilde{X}^{(n)} : t \geq 0\}$ are continuous, i.e. $\tilde{X}^{(n)}$ takes its values in the same space $C([0, \infty), \mathbb{R})$ as the Brownian motion process.

Further quantifying the deviation between $\frac{1}{n} S_{\lfloor nt \rfloor}$ and ut, the result is named the Function Central Limit Theorem (FCLT), also known as Donsker's theorem; see [Whi02, Theorem 4.3.2] for the proof.

- **functional central limit theorem**
  For $t \geq 0$, define that

$$S_n(t) = \sqrt{n}\left(\frac{1}{n}S_{\lfloor nt \rfloor} - ut\right) \tag{1.6}$$

  is the normalized partial sum process in $D([0,\infty),\mathbb{R}))$, then

$$S_n(t) \Rightarrow \sigma W(t), \quad \text{as } n \to \infty,$$

  where $W(t)$ is a standard Brownian motion.

There is a natural path space large deviation result for $\frac{1}{n}S_{\lfloor nt \rfloor}$, Mogulskii's theorem, see [DZ98, Theorem 5.1.2] for the proof. Let $C([0,T];\mathbb{R})$ denote the space of continuous functions from $[0,T]$ to $\mathbb{R}$, and let $C_0([0,T];\mathbb{R}) \subseteq C([0,T];\mathbb{R})$ denote the subspace of functions with value zero at time zero.

- **large deviation principle**
  $\frac{1}{n}S_{\lfloor nt \rfloor}$ satisfy LDP in $C_0([0,T],\mathbb{R})$

$$\mathbb{P}\left(\frac{1}{n}S_{\lfloor nt \rfloor} \approx \gamma\right) \sim e^{-nI(\gamma)}, \quad t \geq 0,$$

  where the rate function $I(\gamma) = \int_0^1 I(\dot{\gamma}_t)\mathrm{d}t$. Here, $\dot{\gamma}_t$ denotes the velocity of the path γ at time t.

From FCLT, we know that Brownian motion is obtained by scaling random walks, taking the limit after linear interpolation. If we replace the random walk with the Brownian motion, we get a similar statement of limit theorems. The first one is Schilder's theorem; see [DZ98, Theorem 5.2.3].

Theorem 1.1 (Schilder's theorem). *The law of $X_n(t) = \frac{1}{\sqrt{n}}W(t)$ satisfy LDP in $C_0([0,T];\mathbb{R})$*

$$\mathbb{P}(X_n(t) \approx \gamma(t)) \sim e^{-nI(\gamma)},$$

where

$$I(\gamma) = \begin{cases} \frac{1}{2}\int_0^T |\dot{\gamma}(t)|^2\mathrm{d}t, & \textit{if } \gamma \in H_1, \\ \infty, & \textit{else} \end{cases}$$

is a good rate function, where $H_1([0,T];\mathbb{R}) \subseteq C_0([0,T];\mathbb{R})$ for the set of functions $\gamma \in C_0([0,T];\mathbb{R})$ which are absolutely continuous and which satisfy $\int_0^T |\dot{\gamma}(t)|^2\mathrm{d}t < \infty$.

Next, we remark that Schilder's theorem can be extended to more general diffusion processes. In the following result, the limit is taken as the scale of the noise goes to zero, so these results are extremely useful in practical engineering situations where one is interested in the long time behavior of dynamical systems in the small noise regime. For the sake of completeness, we state it now:

Theorem 1.2 (Freidlin-Wentzell, Theorem 5.6.3 in [DZ98]). *Let $u : \mathbb{R} \to \mathbb{R}$ be a Lipschitz function, write $\{X_n(t)\}_{t\in[0,T]}$ for the solution to the SDE*

$$\mathrm{d}X_n(t) = u(X_n(t))\mathrm{d}t + \frac{1}{\sqrt{n}}\mathrm{d}W(t). \tag{1.7}$$

Then, the law of $X_n(t)$ satisfy LDP in $C_0([0,T];\mathbb{R})$

$$\mathbb{P}(X_n(t) \approx \gamma(t)) \sim e^{-nI_u(\gamma)},$$

where

$$I_u(\gamma) = \begin{cases} \frac{1}{2}\int_0^T |\dot\gamma(t) - u(\gamma(t))|^2 \mathrm{d}t, & \text{if } \gamma \in H_1 \\ \infty, & \text{else} \end{cases}$$

is a good rate function.

From the results of the Schilder and Fredlin-Wentzell theorems, we derive rate functions of having the general form

$$I(\gamma) = \int_0^T \mathcal{L}(\gamma(t), \dot\gamma(t))\mathrm{d}t. \tag{1.8}$$

The map $\mathcal{L} : \mathbb{R}\times\mathbb{R} \to [0,\infty]$ that appears in the rate function is called the Lagrangian. If a process $X_n(t)$ satisfies a large deviation principle with a Lagrangian rate function, then its limiting dynamics can be determined by solving $\mathcal{L}(x(t), \dot x(t)) = 0$. For example, $\mathcal{L}(x,v) = \frac{1}{2}|v|^2$ is independent of x in Schilder's theorem, and $\mathcal{L}(x,v) = \frac{1}{2}|v - u(x)|^2$ is used in the Fredlin-Wentzell theorem. The limiting dynamics are $\dot x(t) = 0$ and $\dot x(t) = u(x)$, respectively.

Initially, the Schilder and Fredlin-Wentzell theorems studied the large deviations of simple stochastic differential equations. Motivated by this, many researchers have extended their work to more complex stochastic differential equations. In the following, we will further extend the discussion to the case of slow-fast processes.

1.4 Large deviations

1.4.1 Large deviations for simple slow-fast processes

In the previous section, we have seen that the study of large deviations for complex stochastic differential equations is quite mature. However, many natural phenomena vary over multiple time scales where an interplay between slow and fast processes generates complex behavior. Some examples can be seen in weather modeling, the stock market, physiological rhythms, and pituitary cells. These examples refer to systems with multiple time scales, which motivated us to study the large deviation of slow-fast processes.

To perform mathematical analysis, we introduce a switch in (1.7), we study a special class of dynamical systems of the form:

$$\mathrm{d}X_n(t) = u(X_n(t), \Lambda_n(t))\mathrm{d}t + \frac{1}{\sqrt{n}}\mathrm{d}W(t) \tag{1.9}$$

with $\Lambda_n(t)$ a jump process flip between states

$$-1 \to 1, \quad 1 \to -1, \quad \text{at rate } \; n. \tag{1.10}$$

In general, the limiting dynamics of (1.9) is characterized by the averaging principle. The application of this averaging principle provides an effective method to reduce computational complexity. It can be viewed as a variant of the law of large numbers.

To explain the average principle in our setting, when $n \to \infty$, the noise in (1.9) going away, we expect $X_n(t)$ converge to the solution of an ordinary differential equation (ODE). Because of the separation of time scale, we assume that the fast process $\Lambda_n(t)$ equilibrates at $\pi^* = \frac{1}{2}(\delta_{-1} + \delta_1)$, and then have $X_n(t) \to x$ where $x = x(t)$ is the solution of an ODE satisfying that

$$\dot{x}(t) = \frac{1}{2}u(x(t), -1) + \frac{1}{2}u(x(t), 1).$$

To characterize that $X_n(t)$ goes beyond the limit x, we do LDP. Without loss of generality, suppose that $\Lambda_n(t)$ remains stationary in a fixed state z. At this time, by the Freidlin-Wentzell, Theorem 1.2, we first have large deviation:

$$\mathbb{P}[X_n \approx \gamma] \sim \exp\left\{-n\int_0^\infty \frac{1}{2}|\dot{\gamma}(t) - u(\gamma(t), z)|^2\mathrm{d}t\right\}.$$

Secondly, we have Donsker-Varadhan large deviations for the empirical measure of $\Lambda_n(t)$:

$$\mathbb{P}\left[\int_0^1 \delta_{\Lambda_n(t)}\mathrm{d}t \approx \pi\right] \sim \exp\left\{-nI(\pi)\right\}$$

with the rate function

$$I(\pi) = -\inf_{\phi >> 0} \int \frac{A\phi}{\phi}\mathrm{d}\pi,$$

where

$$A\phi(1) = \phi(-1) - \phi(1), \quad A\phi(-1) = \phi(1) - \phi(-1)$$

is the generator of jump process of (1.10). Define $v : \{-1, 1\} \to \mathbb{R}$. We want to produce the speed v while being at x for the LDP of $X_n(t)$. Suppose fast process equilibrates at $\pi = (\pi(-1), \pi(1)) \neq \pi^*$, satisfying $\pi(-1)v_{-1} + \pi(1)v_1 = v$.

For $z = -1$, we know that the rate is $\frac{1}{2}\pi(-1)|v - u(x, -1)|^2$, and for $z = 1$, we know that the rate is $\frac{1}{2}\pi(1)|v - u(x, 1)|^2$. So the cost would be

$$\frac{1}{2}\pi(-1)|v - u(x, -1)|^2 + \frac{1}{2}\pi(1)|v - u(x, 1)|^2.$$

Then we assume that if at -1 produce v_{-1}, if at 1 produce v_1. In this case, let $\frac{1}{2}v_{-1} + \frac{1}{2}v_1 = v$. Then the cost of obtaining v allows us to optimize over all choices v_{-1}, v_1:

$$\inf_{\frac{1}{2}v_{-1}+\frac{1}{2}v_1=v} \left\{ \frac{1}{2}\pi(-1)|v - u(x,-1)|^2 + \frac{1}{2}\pi(1)|v - u(x,1)|^2 \right\}.$$

Finally, at the Freidlin-Wentzell large and Donsker-Varadhan large deviations compete at the same scale, we have

$$\mathbb{P}[X_n(t) \approx \gamma(t)] \sim e^{-nJ(\gamma)}$$

where the rate function is

$$J(\gamma) = \int_0^\infty \inf_\pi \{\mathcal{L}_\pi(\gamma(t), \dot\gamma(t)) + I(\pi)\}\mathrm{d}t, \tag{1.11}$$

with

$$\mathcal{L}_\pi(x,v) = \inf \left\{ \frac{1}{2} \int |v(z) - u(x,z)|^2 \pi(\mathrm{d}z) \middle| \int v(z)\pi(\mathrm{d}z) = v \right\}.$$

Based on the above idea, we can further study the LDP of more complex slow-fast systems in the thesis.

(1) The Cox–Ingersoll–Ross processes with fast switching

Let us first mention some works related to the large deviation of slow and fast systems. The remarkable work of Feng and Kurtz [FK06] consists of combining the tools of probability, analysis, and control theory used in the works of de Acosta [dA97], Dupuis and Ellis [DE97], Evans and Ishii [EI85], Fleming [Fle78], Fleming [Fle85], Fleming [MYZ99], Puhalskii [Puh94], and others to propose a general strategy for the study of large deviations of processes. Feng, Forde, and Fouque in [FFF10] studied the LDP of the Heston stochastic volatility model in the regime in which the maturity is small but large compared to the mean-reversion time of the stochastic volatility factor. Subsequently, Feng, Fouque, and Kumar in [FFK12] established a large deviation principle for general stochastic volatility models in the two regimes of fast and ulta-fast mean-reversion, and we derive asymptotic smiles/skews. After that, Huang, Mandjes, and Spreij studied in [HMS16] large deviations for Markov-modulated diffusion processes with rapid switching. In [PS24], Peletier and Schlottke proved the pathwise LDP of switching Markov processes by exploiting the connection between Hamilton-Jacobi (HJ) equations and Hamilton-Jacobi-Bellman (HJB) equations. In [KS20], Kraaij and Schlottke studied the LDP for the slow-fast system under regular conditions, where the fast process is a switching process. For the proof, they used the Bootstrapping procedure, which is a technology for comparison principle of the HJB equation. Later, Della Corte and Kraaij [DCK24] continued to explore LDP in the context of molecular motors modeled by a diffusion process driven by the gradient of a weakly periodic potential that depends on an internal degree of freedom. The switch of the internal state, which can freely be interpreted as a molecular switch, is modeled as a Markov jump process that depends on the location of the motor.

However, as mentioned earlier, all the work mainly considered LDP in a compact setting. We now consider a singular case that causes the noncompact domain of the slow process. The basic Cox-Ingersoll-Ross (CIR) model is a diffusion equation on $(0,\infty)$ with singularity at 0. Because of this, Euclidean techniques to study the large deviation principle fail, as has been exhibited in [DFL11]. Instead, the authors in [DFL11] take a Riemannian point of view to perform the analysis of the associated Hamilton-Jacobi equation. Thus, working with the more involved model with a switch taking possibly many values, the analysis is expected to involve techniques arising from Riemannian geometry.

Therefore, in the Chapter 3, the following CIR process with fast switching is treated to investigate large deviations of solutions.

Let $E=(0,\infty)$ and $S=\{1,2,\ldots,N\}$, $N<\infty$. The CIR processes with fast switching on $E\times S$ are described as

$$\begin{cases} \mathrm{d}X_n^\varepsilon(t) = \eta(\mu(\Lambda_n^\varepsilon(t)) - X_n^\varepsilon(t))\mathrm{d}t + n^{-\frac{1}{2}}\theta\sqrt{X_n^\varepsilon(t)}\mathrm{d}W(t), \\ (X_n^\varepsilon(0), \Lambda_n^\varepsilon(0)) = (x_0, k_0) \in E\times S, \end{cases} \tag{1.12}$$

where the fast process $\Lambda_n^\varepsilon(t)$ is a switching process with transition rate $\frac{1}{\varepsilon}q_{ij}(x)$ on a set S,

$$\mathbb{P}(\Lambda_n^\varepsilon(t+\triangle)=j \mid \Lambda_n^\varepsilon(t)=i, X_n^\varepsilon(t)=x) = \begin{cases} \frac{1}{\varepsilon}q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j\neq i, \\ 1+\frac{1}{\varepsilon}q_{ij}(x)\triangle+\circ(\triangle), & \text{if } j=i, \end{cases} \tag{1.13}$$

for small $\Delta>0$, $i,j\in S$, $x\in E$, and $\varepsilon>0$ is a small parameter. Obviously, (1.12) and (1.13) together is a slow-fast system.

Then, formally the large deviation principle with speed n holds for $X_n^\varepsilon(t)$ with a good rate function J having the same form as in (1.11),

$$J(\gamma) = \int_0^\infty \inf_\pi\{\mathcal{L}_\pi(\gamma(t),\dot\gamma(t)) + I(\pi)\}\mathrm{d}t \tag{1.14}$$

and

$$\mathcal{L}_\pi(x,v) = \inf\left\{ \frac{|v-\eta(\mu(i)-x(t))|^2}{2\theta^2 x(t)}\pi(\mathrm{d}z) \middle| \int v(z)\pi(\mathrm{d}z) = v \right\}.$$

where

$$I(\pi) = -\inf_{g\gg 0}\int \frac{R_x\phi}{\phi}\mathrm{d}\pi,$$

where R_x is the generator of a state-dependent switching:

$$R_x g(z) = \sum_{j\in S} q_{zj}(x)\left(\phi(j)-\phi(z)\right).$$

(2) The diffusion processes with fast switchings on complete Riemannian manifolds

Until now, the extensive results on LDPs for slow-fast systems are on Euclidean space, there is not much work on the Riemannian manifold. Röckner and Zhang [RZ04] studied sample path large deviations for diffusion processes on configuration spaces over a Riemannian manifold. Kraaij, Redig, and Versendaal [KRV19] generalized classical large deviation theorems to the setting for complete, smooth Riemannian manifolds. However, they focused on the simple setting of random walks rather than slow-fast systems. Versendaal [Ver21] studied large deviations for Brownian motion in evolving Riemannian manifolds. Based on existing research results, in Chapter 4, we further prove the LDPs of the diffusion processes with fast switchings on complete Riemannian manifolds. If the Riemannian manifold is not complete, the CIR process with fast switching is a special case. This can be explored in future work.

Let M be a Riemannian manifold. If we do not care about the completeness of Riemannian manifold. Define $\mathcal{E} : E \times S \to (-\infty, +\infty]$. The functional $\mathcal{E}$ is smooth and finite on E. Working in the natural global chart, we can define a Riemannian metric using $g(x) = x^{-1}$, or equivalently $\langle v, w \rangle_{g(x)} = \frac{1}{x} vw$ on the tangent bundle at x. In this setting, The model we deal with below is an extension of (1.12). That is, when fixed z, we get

$$b(x, z) := \mathrm{grad}\mathcal{E}(x, z) = g^{-1}(x)\mathcal{E}'(x, z) = \eta(\mu(z) - x),$$

if we set $\mathcal{E}(x, z) = \eta[\mu(z)\log(x) - x + \mu(z) - \mu(z)\log\mu(z)]$.

Based on the above analysis, in Chapter 4, we are considering a stochastic differential equation on $M \times S$ with initial value (x_0, k_0):

$$\mathrm{d}X_n^\varepsilon(t) = \frac{1}{\sqrt{n}} u_n^\varepsilon(t) \circ \mathrm{d}W(t) + b(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t))\mathrm{d}t,$$

where $\Lambda_n^\varepsilon(t)$ is a switching process with transition rate $\frac{1}{\varepsilon} q_{ij}(x)$ on a finite set S,

$$\mathbb{P}(\Lambda_n^\varepsilon(t + \triangle) = j \mid \Lambda_n^\varepsilon(t) = i, X_n^\varepsilon(t) = x) = \begin{cases} \frac{1}{\varepsilon} q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j \neq i, \\ 1 + \frac{1}{\varepsilon} q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j = i, \end{cases}$$

for small $\Delta > 0$, $i, j \in S$, $x \in M$, and $\varepsilon > 0$ is a small parameter. $u_n^\varepsilon(\cdot)$ is a unique element such that $X_n^\varepsilon(t) = \mathbf{p} u_n^\varepsilon(t)$, where $\mathbf{p} : O(M) \to M$ is a projection map. Precise details and conditions of this system will be specified later. Then we get LDP with the rate function having the form (1.14). But the proof on a Riemannian manifold should be careful, especially if we want to get some global properties on M.

1.5 Parameter estimations

The asymptotic properties of stochastic differential equations are studied to understand the long time behavior of the solutions. As one of the asymptotic properties, large deviations have been discussed earlier. However, if there are unknown parameters in the coefficients of stochastic differential equations, it is necessary to study the asymptotic properties of the parameters. For example, we assume that u is an unknown parameter in the OU process:

$$\mathrm{d}X(t) = -\theta X(t)\mathrm{d}t + \mathrm{d}W(t). \tag{1.15}$$

Many scholars have studied the asymptotic properties of parameter estimators from the perspective of statistics. Specifically, they focus on two key asymptotic properties of parameter estimators: consistency and asymptotic normality.

Before we give the definition of consistency, we introduce why consistency is crucial. In general, θ is commonly used to represent parameters, and the set of all possible values of parameter θ is called the parameter space, which is represented by Θ.

As we all know, a point estimate is a statistic, so it is a random variable. Under the condition of a certain sample size, it is impossible to ask it to be exactly equal to the true value of the parameter. But if we have enough observations, according to the Glivenko-Cantelli theorem, as the sample size increases, the empirical distribution functions approximate the true distribution functions. Therefore, it is perfectly reasonable to ask the estimator to approximate the true value of the parameter as the sample size continues to increase. This is consistency, which is defined as follows:

Definition 1.3 (Consistency)**.** Let $\theta \in \Theta$ is a unknown parameter, $\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \ldots, x_n)$ is a parameter estimator of θ, n is sample size. We call $\hat{\theta}_n$ consistency for θ if any $\varepsilon \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0.$$

Consistency is considered to be one of the most basic requirements for estimation. If an estimator fails to achieve any specified level of precision in estimating the parameter as the sample size increases, it becomes highly questionable. Generally, estimates that do not meet the compatibility requirements are disregarded.

The next question of interest concerns the order at which the discrepancy $\hat{\theta}_n - \theta$ converges to zero after having consistency. The answer depends on the specific situation, but for estimators based on n replications of an experiment the order is often $1/\sqrt{2}$. Then multiplication with the inverse of this rate creates a proper balance, and the sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution, most often a normal distribution.

Definition 1.4 (Asymptotic normality)**.** An asymptotically normal estimator $\hat{\theta}_n$ of parameter θ is said to be asymptotically normal if there exists a sequence of nonnegative constants $\sigma_n(\theta)$ tending to zero, such that $\frac{\hat{\theta}_n - \theta}{\sigma_n(\theta)}$ converges in distribution to the standard normal distribution. In this case, $\hat{\theta}_n$ is also said to follow an asymptotic normal distribution $N(\theta, \sigma^2(\theta))$.

By studying the consistency and asymptotic normality of parameter estimators, the estimation and inference of parameters in stochastic differential equations can be more fully understood, thus improving the ability to predict long-term behavior.

1.6 Least squares estimation

Based on the theory of parameter estimation of random variables, we consider the parameter estimation of stochastic differential equations. The literature primarily adopts two methods for estimating drift parameters in stochastic differential equations: the first one is maximum likelihood estimation (MLE) based on the Girsanov transformation. For example, Prakasa Rao [PR83], Liptser and Shiryaev [LS01]. The second one is the least squares method (LSE); see Le Breton [LB76] and Kasonga [Kas88]. It turns out that when the driving is Brownian motion or a general square-integrable process, both LSE and MLE are applicable, and LSE has strong consistency under certain conditions.

1.6.1 LSE for linear drift SDEs

To prepare for the questions considered in this thesis, we begin with the linear drift SDEs, the OU process (1.15), using the least squares method. For fixed $T > 0$,

$$\mathrm{d}X^\varepsilon(t) = -\theta X^\varepsilon(t)\mathrm{d}t + \varepsilon \mathrm{d}W(t), \tag{1.16}$$

where θ is the unknown parameter to be estimated. We assume that θ can be estimated be continuous observation $\{X^\varepsilon(t)\}_{0\le t\le T}$. The Brownian motion stands for the stochastic error. We set the zero initial value for later analysis.

To explain the least squares technique for (1.16) we formally write

$$\dot{X}^\varepsilon(t) = -\theta X^\varepsilon(t) + \varepsilon \dot{W}(t),$$

where $\dot{f}(t)$ denote derivative with respect to time and we minimize

$$\begin{aligned}&\frac{1}{\varepsilon^2}\int_0^T |\dot{X}^\varepsilon(t) + \theta X^\varepsilon(t)|^2 \mathrm{d}t \\ &\quad = \frac{1}{\varepsilon^2}\left[\int_0^T (\dot{X}^\varepsilon(t))^2\mathrm{d}t + 2\theta\int_0^T X^\varepsilon(t)\mathrm{d}X^\varepsilon(t) + \theta^2\int_0^T (X^\varepsilon(t))^2\mathrm{d}t\right].\end{aligned}$$

It is a quadratic function in terms of θ. The minimizer can be explicitly represented by

$$\hat{\theta}_\varepsilon = -\frac{\int_0^T X^\varepsilon(t)\mathrm{d}X^\varepsilon(t)}{\int_0^T (X^\varepsilon(t))^2\mathrm{d}t},$$

we call it the least squares estimator. To show the least squares estimator $\hat{\theta}_\varepsilon$ converges to the real value θ_0 satisfying $\mathrm{d}X^0(t) = -\theta_0 X^0(t)\mathrm{d}t$, we have

$$\begin{aligned}\hat{\theta}_\varepsilon - \theta_0 &= -\frac{\int_0^T X^\varepsilon(t)\mathrm{d}X^\varepsilon(t)}{\int_0^T (X^\varepsilon(t))^2\mathrm{d}t} - \theta_0 \\ &\overset{(1.16)}{=} \theta - \theta_0 - \frac{\varepsilon\int_0^T X^\varepsilon(t)\mathrm{d}W(t)}{\int_0^T (X^\varepsilon(t))^2\mathrm{d}t}.\end{aligned} \tag{1.17}$$

When $\varepsilon \to 0$, from (1.17) we have the consistency

$$\hat{\theta}_\varepsilon \to \theta_0.$$

We proceed to consider the asymptotic normality by rescaling $\hat{\theta}_\varepsilon - \theta_0$ by ε^{-1}, from (1.17) we have

$$\varepsilon^{-1}(\hat{\theta}_\varepsilon - \theta_0) = \varepsilon^{-1}(\theta - \theta_0) - \frac{\int_0^T X^\varepsilon(t)\mathrm{d}W(t)}{\int_0^T (X^\varepsilon(t))^2\mathrm{d}t},$$

where $\int_0^T X^\varepsilon(t)\mathrm{d}W(t)$ is a stochastic integral. When $\varepsilon \to 0$, we have

$$\varepsilon^{-1}(\hat{\theta}_\varepsilon - \theta_0) \to -\frac{\int_0^T X^0(t)\mathrm{d}W(t)}{\int_0^T (X^0(t))^2\mathrm{d}t}, \tag{1.18}$$

which means that the least squares estimator is asymptotic normality.

However, we assume that the OU process can be observed continuously in time. This assumption is not always appropriate and impossible to achieve in practice. In addition, the OU process is a stochastic differential equation with linear drift concluding unknown parameters, which can earlier get the explicit least squares estimator.

1.6.2 LSE for nonlinear drift SDEs

In this section, we will explore the behavior of least squares estimators in the context of nonlinear stochastic differential equations. We begin by introducing the following process: for fixed $T > 0$,

$$\mathrm{d}X^\varepsilon(t) = b(X^\varepsilon(t), \theta)\mathrm{d}t + \varepsilon \mathrm{d}W(t), \tag{1.19}$$

where $b : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$ is nonlinear about θ. We further restate (1.19) to

$$\dot{X}^\varepsilon(t) = b(X^\varepsilon(t), \theta) + \varepsilon \dot{W}(t).$$

In this case, the quadratic function is described by

$$\Psi_\varepsilon(\theta) = \int_0^T |\dot{X}^\varepsilon(t) - b(X^\varepsilon(t), \theta)|^2 \mathrm{d}t, \tag{1.20}$$

and is called contrast functions in parameter estimation theory. Similarly, the minimum value of Ψ_ε occurs when the derivative is zero. However, in the nonlinear system, the derivatives $\frac{\mathrm{d}\Psi_\varepsilon(\theta)}{\mathrm{d}\theta}$ are functions of both the independent variable and the parameters, and these derivative equations do not have a closed solution in general. Instead, given an initial value, then the parameters are refined iteratively, that is, the values are obtained by successive approximation.

Numerical analysis is the study of algorithms that use numerical approximation for the problems of mathematical analysis. Euler-Marayama (EM) scheme is a simple numerical method for stochastic differential equations. Without loss of generality, one may assume that there exists a sufficiently large integer $n > 0$ such that the stepsize

$$\delta := \frac{T}{n} \in (0, 1).$$

Now, for $k = 1, 2, \ldots, n$, we introduce the following EM scheme form of (1.19)

$$Y^\varepsilon((k+1)\delta) = Y^\varepsilon(k\delta) + b(Y^\varepsilon(k\delta), \theta)\delta + \Delta W_k,$$

where $\Delta W_k := W((k+1)\delta) - W(k\delta)$. Hence, the contrast function (1.20) is discreted and becomes the finite sum

$$\Psi_{n,\varepsilon}(\theta) = \delta^{-1} \sum_{k=1}^{n} F_k^*(\theta) F_k(\theta),$$

where

$$F_k(\theta) := Y^\varepsilon((k+1)\delta) - Y^\varepsilon(k\delta) - b(Y^\varepsilon(k\delta), \theta)\delta.$$

The next goal is that how to find a $\hat{\theta}_{n,\varepsilon}$ such that $\Psi_{n,\varepsilon}(\theta)$ taking the minimum value, namely,

$$\Psi_{n,\varepsilon}(\hat{\theta}_{n,\varepsilon}) = \min_{\theta \in \Theta} \Psi_{n,\varepsilon}(\theta) \quad \text{or} \quad \hat{\theta}_{n,\varepsilon} = \arg\min_{\theta \in \Theta} \Psi_{n,\varepsilon}(\theta).$$

Clearly, due to the nonlinearity of b, it is not possible to derive an explicit expression for $\hat{\theta}_{n,\varepsilon}$. This poses challenges in proving consistency and asymptotic normality. To address this issue, we first introduce theorem 5.9 in [vdV98] to prove the consistency of $\hat{\theta}_{n,\varepsilon}$. Without changing the notation, we emphasize it here.

Theorem 1.5 (Theorem 5.9 in [vdV98]). *Let Ψ_n be random vector-valued functions and let Ψ be a fixed vector-valued function of θ such that for every $\varepsilon > 0$*

$$\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{\mathbb{P}} 0, \tag{1.21}$$

$$\inf_{\theta : d(\theta, \theta_0) \geq \varepsilon} \|\Psi(\theta)\| > 0 = \|\Psi(\theta_0)\|. \tag{1.22}$$

Then any sequence of estimators $\hat{\theta}_n$ such that $\Psi_n(\hat{\theta}_n) = O_{\mathbb{P}}(1)$ converges in probability to θ_0.

Returning to our case, once we have established consistency, it becomes meaningful to expand $\Psi_{n,\varepsilon}$ around θ_0 using a Taylor expansion. Informally, we have

$$\Psi_{n,\varepsilon}(\hat{\theta}_{n,\varepsilon}) \approx \Psi_{n,\varepsilon}(\theta_0) + (\hat{\theta}_{n,\varepsilon} - \theta_0) \nabla_\theta \Psi_{n,\varepsilon}(\theta_0), \qquad \theta \in \mathbb{B}_{\theta_0}(\hat{\theta}_{n,\varepsilon}).$$

This can be rewritten as

$$\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0) \approx \frac{\varepsilon^{-1}(\Psi_{n,\varepsilon}(\hat{\theta}_{n,\varepsilon}) - \Psi_{n,\varepsilon}(\theta_0))}{\nabla_\theta \Psi_{n,\varepsilon}(\theta_0)}.$$

The numerator is asymptotically normal by the central limit theorem. The denominator is an average and can be analyzed by the law of large numbers. Together with Slutsky's lemma, these observations yield $\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0)$ is asymptotic normality. For the rigorous proof, it will be made in this thesis.

(3) The stochastic differential equations with Hölder drift

There are numerous results on parameter estimation for SDEs with regular drift under various settings to support this point. In [WS16], maximum likelihood estimation was used to study drift parameters in diffusion processes. As for more complex processes, we refer to [WWMX16] and [Lon09], who investigated the maximum likelihood estimation of McKean-Vlasov SDEs and studied the parameter estimation problem for one-dimensional Ornstein-Uhlenbeck processes with small Lévy noise, respectively. In particular, for considering a high-frequency sample of discrete observations of the diffusion processes at time points, the following related works are crucial for parameter estimation; see, [FZ89, Yos92, Kes97, Lon09, DGCL18, AHPP23] and references therein.

However, the above SDEs are all under the condition of regular coefficients, and the problem of parameter estimation with irregular drift has not been well studied. Motivated by this, in Chapter 5, we study the asymptotic property of stochastic differential equations with singular coefficients. In detail, the singular coefficient is the drift satisfying Hölder conditions as we will now describe.

We fix the time horizon $T > 0$. For the scale parameter $\varepsilon \in (0, 1)$, we are interested in the following SDE

$$\mathrm{d}X^\varepsilon(t) = b(X^\varepsilon(t), \theta)\mathrm{d}t + \varepsilon\, \sigma(X^\varepsilon(t))\mathrm{d}W(t), \tag{1.23}$$

where $b : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$, $\sigma : \mathbb{R}^d \to \mathbb{R}^d \otimes \mathbb{R}^d$ and $(W(t))_{t\geq 0}$ is a d-dimensional Brownian motion defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with the filtration $(\mathscr{F}_t)_{t\geq 0}$ satisfying the usual condition. In (1.23), we assume that the drift b and the diffusion σ are known apart from the parameter $\theta \in \Theta$. Then, as above mentioned, we establish contrast function and further get least square estimator $\hat{\theta}_{n,\varepsilon} = \arg\min_{\theta\in\Theta} \Psi(\theta)$. Under suitable conditions, the first result is that

$$\hat{\theta}_{n,\varepsilon} \to \theta_0 \quad \text{in probability.}$$

Roughly speaking, $\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0)$ is asymptotic normality, which is the second result. Please see Chapter 5 for more details.

(4) The McKean-Vlasov stochastic differential equations with delay

In recent years, McKean–Vlasov stochastic differential equations (SDEs), also called distribution-dependent or mean field SDEs, have received increasing attention for their theoretical importance in characterizing nonlinear Fokker–Planck equations from physics. On the other hand, SDEs have been developed as crucial mathematical tools for modeling economic and financial systems. In the real world, the evolution of these systems is not only driven by micro actions (drift and noise), but also relies on the macro environment (in mathematics, distribution of the systems). The McKean-Vlasov SDE

$$\mathrm{d}X^\varepsilon(t) = b(X^\varepsilon(t), \mathcal{L}_{X^\varepsilon(t)}, \theta)\mathrm{d}t + \varepsilon\,\sigma(X^\varepsilon(t), \mathcal{L}_{X^\varepsilon(t)})\mathrm{d}W(t) \tag{1.24}$$

is a kind of mathematical model, which can characterize the evolution of the phenomenon.

With deep research, people also realize some facts that many phenomena around us do not immediately generate an impact at the moment they occur and they exhibit some degree of delay: individuals infected with the coronavirus may not display symptoms such as fever and cough until two weeks later; when a driver encounters a sudden situation and initiates emergency braking, it takes some time for the car to come to a complete stop. This means that we should add delay elements in (1.24), then we get

$$\begin{aligned} \mathrm{d}X^\varepsilon(t) = b(X^\varepsilon(t), \mathscr{L}_{X^\varepsilon(t)}, X^\varepsilon(t-\tau), \mathscr{L}_{X^\varepsilon(t-\tau)}, \theta)\mathrm{d}t \\ +\varepsilon\,\sigma(X^\varepsilon(t), \mathscr{L}_{X^\varepsilon(t)}, X^\varepsilon(t-\tau), \mathscr{L}_{X^\varepsilon(t-\tau)})\mathrm{d}W(t), \end{aligned} \tag{1.25}$$

which is called McKean-Vlasov SDEs with delay. This is our aim process studying in Chapter 6 if we use a simple notation, $X^\varepsilon(\cdot)$ and $\mathcal{L}_{X^\varepsilon(\cdot)}$ are replaced by $X^\varepsilon_\cdot$ and $\mu^\varepsilon_\cdot$, respectively.

$$\mathrm{d}X^\varepsilon_t = b(X^\varepsilon_t, X^\varepsilon_{t-\tau}, \mu^\varepsilon_t, \mu^\varepsilon_{t-\tau}, \theta)\mathrm{d}t + \varepsilon\,\sigma(X^\varepsilon_t, X^\varepsilon_{t-\tau}, \mu^\varepsilon_t, \mu^\varepsilon_{t-\tau})\mathrm{d}W(t).$$

We assume that the drift term b and the diffusion term σ are known apart from the parameter $\theta \in \Theta$. Then we study the parameter estimation of θ in the same way as above. The main difference is that when we construct the contrast function, we need to discretize not only the time, but also need to discretize the distribution that appears in (1.25) through empirical distribution

$$\mu^{\varepsilon,N}_t(dx) = \frac{1}{N}\sum_{j=1}^{N} \delta_{X^{\varepsilon,j,N}_t}(dx), \qquad t \geq -\tau,$$

where $X^{\varepsilon,j,N}$, $j = 1, 2, \cdots, N$ satisfying a stochastic interacting particle system which can be described as

$$\mathrm{d}X^{\varepsilon,i,N}_t = b(X^{\varepsilon,i,N}_t, X^{\varepsilon,i,N}_{t-\tau}, \mu^{\varepsilon,N}_t, \mu^{\varepsilon,N}_{t-\tau}, \theta)\mathrm{d}t + \varepsilon\sigma(X^{\varepsilon,i,N}_t, X^{\varepsilon,i,N}_{t-\tau}, \mu^{\varepsilon,N}_t, \mu^{\varepsilon,N}_{t-\tau}, \theta)\mathrm{d}W(t).$$

1.7 Goal and overview of the thesis

The thesis is divided into two parts:

(I) Large deviations for slow-fast processes: this part starts with a mathematical introduction of the strategy of large deviations for simple stochastic processes in Chapter 2. Subsequently, it delves into the analysis of large deviations for slow-fast processes in Euclidean space and Riemannian manifold settings, respectively. This includes Chapters 3 and 4.

(II) Parameter estimations related to the study of consistency and asymptotic normality: this involves Chapters 5 and 6.

As mentioned above, in Chapter 2, we introduce a general strategy of large deviations, namely the nonlinear semigroup method, which is frequently employed throughout this thesis. The proof of this method relies on several key elements including operator convergence, exponential tightness, comparison principle, and action-integral representation.

In Chapter 3, we delve into the analysis of large deviations of Cox-Ingersoll-Ross processes with fast switching, employing the nonlinear semigroup method introduced in Chapter 2. However, the domain of CIR processes poses a challenge as it is a noncompact space and singular at 0. To address this issue, the Lyapunov function is introduced to stay away from the singularity points. Additionally, the Riemannian metric allows for the proof of comparison principle. These measures help us facilitate the analysis of large deviations.

We proceed with Chapter 4, where we consider a slow-fast system on a complete Riemannian manifold. The setting gives us a complex Hamiltonian. We first find a suitable Lyapunov function. Then the Riemannian metric with parallel transport to prove the comparison principle. Finally, we prove the existence of a global solution for a nonsmooth Hamiltonian.

Large deviation is a form of an asymptotic property. Now, we turn to consider other asymptotic behavior parameter estimators, focusing on consistency and asymptotic normality.

In Chapter 5, we explore a multi-dimensional stochastic differential equation with an unknown parameter under the Hölder drift condition. Initially, we address the Hölder drift using the Zvonkin transformation. The idea of the Zvonkin transformation is to construct a one-to-one transformation that allows us to transition from a diffusion process with a nonzero drift coefficient to a process without drift. Subsequently, we establish the consistency and asymptotic normality of the least squares estimator after using the Euler-Maruyama scheme. Additionally, we extend our analysis to the setting of stochastic functional differential equations, where the discrete method is replaced by the truncated Euler-Maruyama scheme.

In Chapter 6, we focus on parameter estimation for a McKean-Vlasov stochastic differential equation with delay. Consequently, under more general conditions, we derive the asymptotic properties of the least squares estimator.

We provide a flow diagram in Figure 1.1 to illustrate the relationship between each chapter.

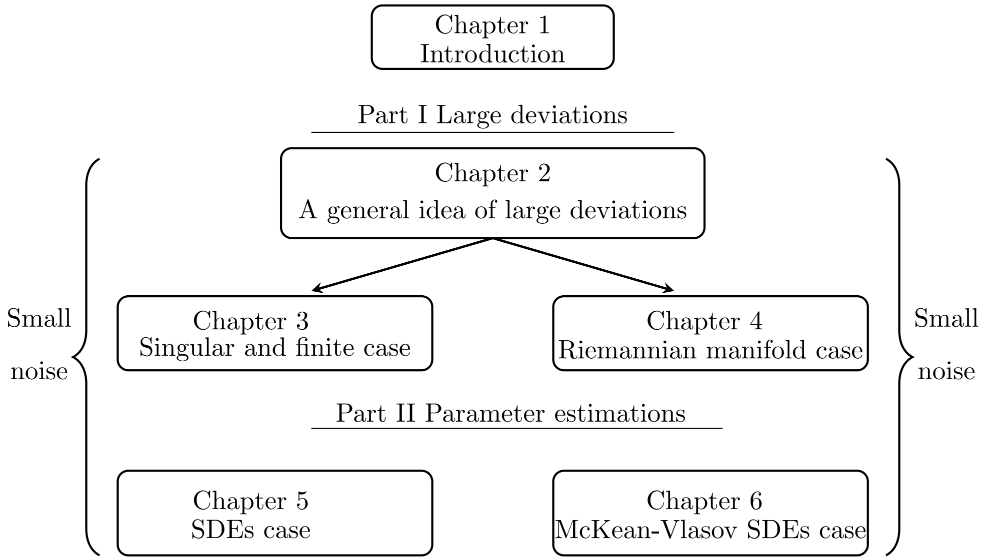

Figure 1.1: Overview of the structure of this thesis

Part I

Large deviations

Chapter 2

Large deviations of simple stochastic processes

In Chapter 1, we provide an informal analysis of the ideas presented in the thesis. Beginning with this Chapter, the content will adhere strictly to formal standards. We commence by defining large deviations in Polish space.

Definition 2.1. Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables on Polish space $\mathcal{X}$. Furthermore, consider a function $I : \mathcal{X} \to [0, \infty]$. We say that

- the function I is a *rate function* if the set $\{x \in \mathcal{X} \mid I(x) \leq c\}$ is closed for every $c \geq 0$. The function I is a *good rate function* if the set $\{x \in \mathcal{X} \mid I(x) \leq c\}$ is compact for every $c \geq 0$.

- the sequence $\{X_n\}_{n\geq 1}$ is *exponentially tight* at speed n, if for every $a \geq 0$, there exists a compact set $K_a \subseteq \mathcal{X}$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(X_n \notin K_a^c) \leq -a.$$

- the sequence $\{X_n\}_{n\geq 1}$ satisfies the *large deviation principle* with speed n and good rate function I if for every closed set $F \subseteq \mathcal{X}$, we have

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(X_n \in F) \leq -\inf_{x\in F} I(x),$$

  and, for every open set $U \subseteq \mathcal{X}$, we have

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(X_n \in U) \geq -\inf_{x\in U} I(x).$$

Recall the Freidlin-Wentzell theorem discussed in Chapter 1, which motivates the study of large deviations of stochastic differential equations. The classical proof method utilized by Freidlin-Wentzell is the continuous mapping method. In this chapter, we will explore a general method that uses semigroup theory to establish pathwise large deviation principles for Markov processes.

To accomplish this objective, we first present the Bryc theorem regarding the large deviation principles for sequences of random variables.

Theorem 2.2 (Bryc's theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of E-value random variables. Suppose that the sequence $(X_n)_{n\geq 1}$ is exponentially tight and that the limit*

$$\Lambda(f) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[e^{nf(X_n)}]$$

exists for each $f \in C_b(E)$. Then $(X_n)_{n\geq 1}$ satisfies the large deviation principle with good rate function

$$I(x) = \sup_{f\in C_b(E)} \{f(x) - \Lambda(f)\}, \quad x \in E.$$

We aim to establish connections between large deviations for sequences of random variables and the asymptotic behavior of functionals, specifically logarithmic moment generating functionals of the form $\frac{1}{n}\log \mathbb{E}[e^{nf(X_n(t))}]$. To achieve this, we introduce a sequence of Markov processes $X_n(t)$ taking values in E with generator A_n. We define the linear Markov semigroup $S_n(t)$ as follows:

$$S_n(t)f(x) := \mathbb{E}[f(X_n(t))|X_n(0) = x], \quad t \geq 0,\ x \in E,$$

which, at least formally, satisfies

$$A_n f := \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} S_n(t)f = \lim_{t\to 0} \frac{S_n(t)f - f}{t}.$$

Fleming in [Fle85] introduced the following nonlinear contraction semigroup $V_n(t)$ for Markov processes $X_n(t)$,

$$V_n(t)f(x) = \frac{1}{n}\log \mathbb{E}_x[e^{nf(X_n(t))}] = \frac{1}{n}\log S_n(t)e^{nf}(x), \quad t \geq 0,\ x \in E, \tag{2.1}$$

and large deviations for sequences $(X_n(t))_{n\geq 1}$ of Markov processes can be studied using the asymptotic behavior of the corresponding nonlinear semigroups. Again, at least formally, $V_n(t)$ should satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} V_n(t)f = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \frac{1}{n}\log S_n(t)e^{nf} = \frac{1}{n}e^{-nf}A_n e^{nf} =: H_n f. \tag{2.2}$$

Fleming and others have employed this approach to establish large deviation results for sequences $(X_n(t))_{n\geq 1}$ of Markov processes $X_n(t)$ at single time points and exit times. The book by [FK06] extends this approach further, demonstrating how convergence of the nonlinear semigroups and their generators H_n can be used to obtain both, exponential tightness and the large deviation principle for the finite dimensional distributions of the processes. Showing the large deviation principle for finite dimensional distributions of the sequence of Markov processes and using the exponential tightness gives then the full pathwise large deviation principle for the sequence of Markov processes. The approach, known as nonlinear semigroup methods, requires proving the following key points:

(a) exponential tightness;

(b) the existence of an H such that $H_n \to H$;

(c) H generates a semigroup $V(t)$.

However, during the proof, establishing that H determines a limiting semigroup $V(t)$ can be challenging. In the example that follows, the existence of such a semigroup from the nonlinear H is asserted without formal proof.

2.1 An illustrative example

Recall the Freidlin-Wentzell theorem, which studies the large deviation behavior of solutions to stochastic differential equations with small noise.

Theorem 2.3 (Freidlin-Wentzell, Theorem 5.6.3 in [DZ98])**.** *Let $u : \mathbb{R} \to \mathbb{R}$ be a Lipschitz function, write $\{X_n(t)\}_{t\in[0,T]}$ for the solution to the SDE*

$$\mathrm{d}X_n(t) = u(X_n(t))\mathrm{d}t + \frac{1}{\sqrt{n}}\mathrm{d}W(t).$$

Then, the law of $X_n(t)$ satisfy LDP in $C_0([0,T];\mathbb{R})$

$$\mathbb{P}(X_n(t) \approx \gamma(t)) \sim e^{-nI_u(\gamma)},$$

where

$$I_u(\gamma) = \begin{cases} \frac{1}{2}\int_0^T |\dot{\gamma}(t) - u(\gamma(t))|^2 \mathrm{d}t, & \textit{if } \gamma \in H_1 \\ \infty, & \textit{else} \end{cases}$$

is a good rate function.

The proof is by the continuous mapping method. Here, we want to use the nonlinear semigroup method. The important step is that if there is a semigroup $V(t)$ generated by H. The answer is yes. In the following, we provide an informal proof.

To simplify, we assume that item (a) naturally holds when working in a compact space.

To obtain item (b), first, the generator of $X_n(t)$ is given by

$$A_n f(x) = u(x)\partial_x f(x) + \frac{1}{2n}\partial_{xx}(x)f(x), \quad x \in \mathbb{R}.$$

Therefore, according to $H_n f = \frac{1}{n} e^{-nf} A_n e^{nf}$, we have

$$H_n f(x) = u(x)\partial_x f(x) + \frac{1}{2n}\partial_{xx} f(x) + \frac{1}{2}|\partial_x f(x)|^2,$$

which represents a nonlinear generator of $X_n(t)$. Let $n \to \infty$ in the above equation, it leads to a limit Hamiltonian:

$$Hf(x) = u(x)\partial_x f(x) + \frac{1}{2}|\partial_x f(x)|^2. \tag{2.3}$$

We have completed the proof of item (b). For item (c), we claim that H generates the semigroup $V(t)$ is

$$V(t)f(x) = \sup_{\gamma,\gamma(0)=x} \left\{ f(\gamma(t)) - \int_0^t \frac{1}{2}|\dot{\gamma}(s) - u(x)|^2 \mathrm{d}s \right\}. \tag{2.4}$$

Informally, we can justify this claim by considering the derivative with respect to t at $t = 0$:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} V(t)f(x) &= \sup_{\gamma,\gamma(0)=x} \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \left\{ f(\gamma(t)) - \int_0^t \frac{1}{2}|\dot{\gamma}(s) - u(x)|^2 \mathrm{d}s \right\} \\
&= \sup_{\gamma,\gamma(0)=x} \left\{ \partial_x f(\gamma(0))\dot{\gamma}(0) - \frac{1}{2}|\dot{\gamma}(0) - u(x)|^2 \right\} \\
&= \sup_{v} \left\{ \partial_x f(x) v - \frac{1}{2}|v - u(x)|^2 \right\} \\
&= u(x)\partial_x f(x) + \frac{1}{2}(\partial_x f(x))^2 \\
&= Hf(x).
\end{aligned} \tag{2.5}$$

Thus, $V(t)$ indeed corresponds to the semigroup generated by H.

Next, we turn to the large deviation principle associated with this setup gives a rate function

$$I(x) = \int_0^T \frac{1}{2}|\dot{\gamma}(t) - u(x)|^2 \mathrm{d}t.$$

In general, the integrand of the rate function is called Lagrangian $\mathcal{L}$. In this case, $\mathcal{L}(\dot{\gamma}(t)) = \frac{1}{2}|\dot{\gamma}(t) - u(x)|^2$. The Lagrangian $\mathcal{L}$ can be characterized in two ways.

First, Lagrangian $\mathcal{L}$ is the Legendre transform of Hamiltonian H. To do it, let $p = \partial_x f(x)$, from (2.3) we define

$$\mathcal{H}(p) = up + \frac{1}{2}p^2,$$

which is a convex function of p. We change the function of p to a function of the rate v based on the Legendre transform of $\mathcal{H}$, obtaining

$$\mathcal{L}(v) = \sup_{p \in \mathbb{R}} \{pv - \mathcal{H}(p)\} = \frac{1}{2}|v - u(x)|^2. \tag{2.6}$$

By convex duality in Euclidean space, it also follows that

$$\mathcal{H}(p) = \sup_{v\in\mathbb{R}}\{pv - \mathcal{L}(v)\}.$$

Second, Lagrangian $\mathcal{L}$ comes from an optimization problem. Modifying $V(t)$ in (2.4) using $\mathcal{L}$, we get

$$V(t)f(x) = \sup_{\gamma,\gamma(0)=x}\left\{f(\gamma(t)) - \int_0^t \mathcal{L}(\dot\gamma(s))\mathrm{d}s\right\}.$$

Thus, Lagrangian $\mathcal{L}$ is characterized both by its role as the Legendre transform of the Hamiltonian $\mathcal{H}$ and by its appearance in the optimization formulation of the semigroup $V(t)$.

We continue to analyze (2.5). The semigroup $V(t)$ is a solution of the following parabolic partial differential equation (PDE):

$$\begin{cases}\partial_t g(x,t) - \mathcal{H}(\partial_x g(x,t)) = 0, & \text{if } t>0,\\ g(x,0) = g_0(x), & \text{if } t=0.\end{cases} \tag{2.7}$$

The above PDE can be solved rigorously in the sense of viscous solutions. However, in Feng and Kurtz's book [FK06], instead of (2.7), they studied the following Hamilton-Jacobi (HJ) equation to derive the large deviation principle:

$$f - \lambda Hf = h \tag{2.8}$$

where H is defined in (2.3), $\lambda>0$ and $h\in C_b(\mathbb{R})$. The resolvent

$$\mathbf{R}(\lambda)h(x) = \sup_{\substack{\gamma\in\mathcal{AC}\\ \gamma(0)=x}}\int_0^\infty \lambda^{-1}\mathrm{e}^{-\lambda^{-1}t}\left(h(\gamma(t)) - \int_0^t\frac{1}{2}|\dot\gamma(s)-u|^2\mathrm{d}s\right)\mathrm{d}t,$$

is a solution of (2.8). By Lemma 8.18 of [FK06], we have

$$\lim_{m\to\infty}|\mathbf{R}(t/m)^m f(x) - \mathbf{V}(t)f(x)| = 0.$$

In the thesis, we follow the approach using the second HJB equation, $f-\lambda Hf = h$, to prove large deviations.

The background of HJ equations originates from control theory. While solving the HJ equation at all points in the classical sense is challenging, an alternative approach involves considering the following: suppose $\overline{f}\in C^\infty(\mathbb{R})$ such that $f-\overline{f}$ has a local maximum at x. This leads to

$$\overline{f}(x) - \lambda H\overline{f}(x) \le h(x),$$

we call $\overline{f}$ is a subsolution of the HJ equation under this setting. Similarly, consider $\underline{f} \in C^\infty(\mathbb{R})$ such that $f - \underline{f}$ has a local minimum at x. Then we have

$$\underline{f}(x) - \lambda H\overline{f}(x) \geq h(x),$$

we call $\overline{f}$ is a supersolution of the HJ equation.

It is worthwhile taking a moment to observe what is going on here. If $f - \overline{f}$ has a local maximum at x, we have $f(x) = \overline{f}(x)$ and $f(y) \leq \overline{f}(y)$ for all y near x. Thus, the graph of $\overline{f}$ touches the graph of f from above at the point x, $\overline{f}$ must be a subsolution of the HJ equation. Similarly, the graph of $\underline{f}$ touches the graph of f from above at the point x, $\underline{f}$ must be a supersolution of the HJ equation. In the following, we give Figure 2.1 to further understand.

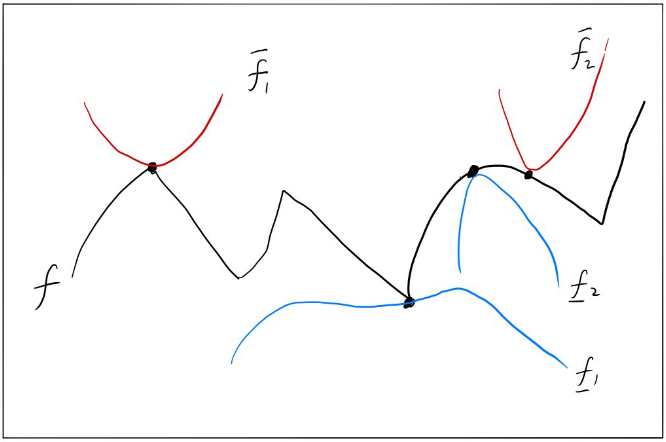

Figure 2.1: An illustration of test functions touching a nonsmooth function u from above and below. The functions $\overline{f}_1$ and $\overline{f}_2$, drawn in red, touch u from above, while $\underline{f}_1$ and $\underline{f}_2$, drawn in blue, touch f from below

Finally, for the strict definition of viscosity solution in the coming section.

2.1.1 Viscosity solutions

We give a strict definition of viscosity solution.
Definition 2.4 (Viscosity solutions)**.** Let $H \subseteq C_b(E) \times C_b(E)$ be a multivalued operator. We denote $\mathcal{D}(H)$ for the domain of H and $\mathcal{R}(H)$ for the range of H. Let $\lambda > 0$ and $h \in C_b(E)$. Consider the Hamilton-Jacobi equation

$$f - \lambda Hf = h. \tag{2.9}$$

Classical solutions We say that u is a classical subsolution of (2.9) if there is a function g such that $(u, g) \in H$ and $u - \lambda g \leq h$. We say that v is a classical supersolution of (2.9) if there is a function g such that $(v, g) \in H$ and $v - \lambda g \geq h$. We say that u is a classical solution if it is both a subsolution and a supersolution.

Viscosity subsolutions We say that u is a (viscosity) subsolution of (2.9) if u is bounded, upper semicontinuous, and if for every $(f, g) \in H$ there exists a sequence $x_n \in E$ such that

$$\lim_{n\to\infty} u(x_n) - f(x_n) = \sup_x u(x) - f(x),$$

$$\limsup_{n\to\infty} u(x_n) - \lambda g(x_n) - h(x_n) \leq 0.$$

Viscosity supersolutions We say that v is a (viscosity) supersolution of (2.9) if v is bounded, lower semicontinuous, and if for every $(f, g) \in H$ there exists a sequence sequence $x_n \in E$ such that

$$\lim_{n\to\infty} v(x_n) - f(x_n) = \inf_x v(x) - f(x),$$

$$\liminf_{n\to\infty} v(x_n) - \lambda g(x_n) - h(x_n) \geq 0.$$

Viscosity solutions We say that u is a (viscosity) solution of (2.9) if it is both a subsolution and a supersolution to (2.9).

Remark 2.5. Consider the definition of subsolutions. Suppose that the test function $(f, g) \in H$ has compact sublevel sets, then instead of working with a sequence x_n, we can pick x_0 such that

$$u(x_0) - f(x_0) = \sup_x u(x) - f(x),$$

$$u(x_0) - \lambda g(x_0) - h(x_0) \leq 0.$$

Similarly, a simplification holds in the case of supersolutions. This is used in the proof Lemma 3.13 below.

Definition 2.6 (Comparison principle)**.** We say that (2.9) satisfies the comparison principle if for every viscosity subsolutions u and viscosity supersolutions v to (2.9), we have $u \leq v$.

Remark 2.7 (Uniqueness)**.** The comparison principle implies the uniqueness of viscosity solutions. Suppose that u and v are both viscosity solutions, then the comparison principle yields that $u \leq v$ and $v \leq u$, implying that $u = v$.

2.2 Nonlinear semigroup methods

In the last section, we introduced viscosity solutions and HJ equations using simple stochastic differential equations. In the current section, we proceed to present a general strategy for large deviations based on viscosity solutions and HJ equations.

Before presenting the strategy, we will first provide some definitions that will be used later. Based on Lagrangian $\mathcal{L}$, we give the definition of Nisio semigroup $V(t)$ and resolvent $\mathbf{R}(\lambda)$.

Definition 2.8 (Nisio semigroup)**.** Define the Nisio semigroup for measurable functions f on E:

$$\mathbf{V}(t)f(x) = \sup_{\substack{\gamma\in\mathcal{AC}\\ \gamma(0)=x}} \left\{ f(\gamma(t)) - \int_0^t \mathcal{L}(\gamma(s),\dot{\gamma}(s))\mathrm{d}s \right\}. \tag{2.10}$$

Definition 2.9 (Resolvent $\mathbf{R}(\lambda)$)**.** For $\lambda > 0$ and $h \in C_b(E)$, define the resolvent $\mathbf{R}(\lambda)h : E \to \mathbb{R}$ by

$$\mathbf{R}(\lambda)h(x) = \sup_{\substack{\gamma\in\mathcal{AC}\\ \gamma(0)=x}} \int_0^\infty \lambda^{-1}\mathrm{e}^{-\lambda^{-1}t}\left(h(\gamma(t)) - \int_0^t \mathcal{L}(\gamma(s),\dot{\gamma}(s))\mathrm{d}s \right)\mathrm{d}t. \tag{2.11}$$

We continue to define an action-integral representation of the rate function using absolutely continuous curves.

Definition 2.10 (Absolutely continuous)**.** We denote by $\mathcal{AC}(E)$ the space of absolutely continuous curves in E. A curve $\gamma : [0,T] \to E$ is absolutely continuous if there exists a function $g \in L^1[0,T]$ such that for $t \in [0,T]$ we have $\gamma(t) = \gamma(0) + \int_0^t g(s)\mathrm{d}s$. We write $g = \dot{\gamma}$.

A curve $\gamma : [0,\infty) \to E$ is absolutely continuous, i.e. $\gamma \in \mathcal{AC}(E)$, if the restriction to $[0,T]$ is absolutely continuous for every $T > 0$.

Definition 2.11 (Exponential compact containment of $X_n(t)$)**.** We say that a process $X_n(t)$ satisfies the *exponential compact containment condition* at speed n, for every all compact $K_0 \subseteq E$, $T > 0$ and $a \geq 0$, there exists a compact set $K_{a,T} \subseteq E$ such that

$$\limsup_{n\to\infty} \sup_{x_0\in K_0} \frac{1}{n} \log \mathbb{P}\left[X_n(t) \notin K_{a,T} \text{ for some } t \leq T \,|\, X_n(0) = x_0\right] \leq -a.$$

Definition 2.12 (Action-integral representation of rate function)**.** We say that a rate function $I : \mathcal{D}_E(\mathbb{R}^+) \to [0,\infty]$ is of the action-integral form if there is a non-trivial convex map $\mathcal{L} : E \times E \to [0,\infty]$ with which

$$I(x) = \begin{cases} I_0(x(0)) + \int_0^\infty \mathcal{L}(x(t),\dot{x}(t))\mathrm{d}t, & \text{if } x \in \mathcal{AC}(E), \\ \infty, & \text{otherwise,} \end{cases}$$

where $I_0 : E \to [0,\infty]$ is a rate function. We refer to the map $\mathcal{L}$ as the Lagrangian, i.e. $v \mapsto \mathcal{L}(x,v)$ is convex and $(x,v) \mapsto \mathcal{L}(x,v)$ is lower semicontinuous.

Definition 2.13 (Extended limit, Definition A.12 in [FK06])**.** For every $n \geq 1$, let $H_n \subset C_b(E) \times C_b(E)$ be an operator. The *extended limit* $\mathrm{ex}-\lim_{n\to\infty} H_n$ is defined as the collection $(f,g) \in C_b(E) \times C_b(E)$ for which there exists a sequence $\{f_n\}_{n\geq 1}$ with $f_n \in \mathcal{D}(H_n)$ such that

$$\lim_{n\to\infty} (\|f_n - f\| + \|H_n f_n - g\|) = 0.$$

An operator H is said to be contained in $\text{ex} - \lim_{n\to\infty} H_n$ if the graph $\{(f, Hf) | f \in \mathcal{D}(H)\}$ is a subset of $\text{ex} - \lim_{n\to\infty} H_n$.

Now, we are ready to provide the main proposition for proving for proving large deviation using nonlinear semigroup methods.

Proposition 2.14 (Adaptation of Theorem 5.15, Theorem 8.27 and Corollary 8.28 in [FK06] to our context)**.** *Let $X_n(t)$ be Markov processes on E. Suppose that*

(a) $X_n(0)$ *satisfies large deviation principle;*

(b) *there exists an operator $H \subset ex - \lim_{n\to\infty} H_n$ in the sense Definition 2.13;*

(c) *we have exponential compact containment of the process $X_n(t)$;*

(d) *for all $\lambda > 0$ and $h \in C_b(E)$, the comparison principle holds for $f - \lambda H f = h$.*

Then the following hold:

(i) (Limit of nonlinear semigroup) There exists a unique operator semigroup $V(t)$ such that

$$\lim_{n\to\infty} \|V_n(t) f_n - V(t) f\| = 0 \tag{2.12}$$

and there exists a unique $R(\lambda)f$ such that

$$\lim_{m\to\infty} \|R(t/m)^m f - V(t) f\| = 0, \tag{2.13}$$

whenever $f \in \overline{\mathcal{D}(H)}$, $f_n \in C_b(E)$, and $\|f_n - f\| \to 0$.

(ii) (Large deviation principle) $X_n(t)$ satisfies the large deviation principle with good rate function I given by

$$I(x) = I_0(x(t_0)) + \sup_{k\in\mathbb{N}} \sup_{0=t_0<t_1<\cdots<t_k<\infty} \sum_{i=0}^{k} I^V_{t_{i+1}-t_i}(x(t_{i+1}) \mid x(t_i)), \tag{2.14}$$

where for $\Delta t = t_{i+1} - t_i > 0$ and $x(t_{i+1})$, $x(t_i) \in E$, the conditional rate functions $I^V_{\Delta t}(x(t_{i+1}) \mid x(t_i))$ are

$$I^V_{\Delta t}(x(t_{i+1}) \mid x(t_i)) = \sup_{f\in C_b(E)} [f(x(t_{i+1})) - V(\Delta t) f(x(t_i))].$$

Suppose in addition that

(e) $V(t) = \mathbf{V}(t)$ with $\mathbf{V}$ as in (2.10).

Then we have the action-integral representation following the representation of rate function (2.14)*:*

$$I(\gamma) = \begin{cases} I_0(\gamma(0)) + \int_0^\infty \mathcal{L}(\gamma(s), \dot\gamma(s))ds, & \text{if } \gamma \in \mathcal{AC}(E), \\ \infty, & \text{otherwise.} \end{cases} \tag{2.15}$$

2.2.1 Outline of the ideas

In this subsection, we sketch why Proposition 2.14 is true. That is, we check the assumptions (a)-(e) of Proposition 2.14 in detail.

Step 1: For the Markov process $X_n(t)$, we aim to prove LDP on the path space. The approach is based on a variant of the projective limit theorem, Lemma 2.15 as below. Namely, if a sequence of the processes is exponentially tight in the Skorokhod space, then it suffices to establish the LDP of finite-dimensional distributions. Moreover, the rate function is given in the projective limit form: it is given as the supremum over the rate functions of the finite-dimensional distributions.

Step 2: We show the exponential tightness of $X_n(t)$ and the LDP for the finite dimensional distributions. For the proof of exponential tightness, we do not enter a detailed discussion and refer to Corollary 4.19 in [FK06], it suffices in our context to establish the exponential compact containment condition. When the state space E is compact, exponential tightness is usually easier to verify.

We are left to prove LDP for finite-dimensional distributions, which is established via Bryc's theorem, Lemma 2.16 as below. For this one needs to prove convergence of log expectations.

For simplicity, we consider the log expectation for $k = 1$ with f_0, $f_1 \in C_b(E)$ and $0 = t_0 < t_1$ only, and have

$$\begin{aligned} \Gamma_n &= \Gamma_n(f_0, f_1) \\ &:= \frac{1}{n} \log \mathbb{E}\big(\mathrm{e}^{nf_0(X_n(t_0)) + nf_1(X_n(t_1))}\big) \\ &= \frac{1}{n} \log \mathbb{E}[\mathbb{E}(\mathrm{e}^{nf_0(X_n(t_0)) + nf_1(X_n(t_1))} \mid X_n(t_0) = x_0] \\ &= \frac{1}{n} \log \mathbb{E}\big(\mathrm{e}^{nf_0(x_0) + nV_n(t_1)f_1(x_0)}\big), \end{aligned} \tag{2.16}$$

where in the third equality we used the Markov property and the conditional log-Laplace transform $V_n(t_1)$ is defined in (2.1). Moreover, (2.16) reduces to proving

(a) the LDP for $X_n(t_0)$ with rate function I_0, which is item (a) of Proposition 2.14;

(b) the convergence of the conditional log-Laplace transform $V_n(t_1)$.

To proceed, we defer the proof of item (b) to Step 3. Given that item (b) is true, there exists a limit Γ of Γ_n as $n \to \infty$. We note that the limit is $f_0(z) + V(t_1)f_1(z)$. Combining this limit, we first use Lemma 2.17 and take over sup about f_0 and f_1 via Lemma 2.16 in the first equality below. The rate function for $(X_n(t_0), X_n(t_1))$ by Bryc's theorem is given by

$$\begin{aligned}
&I_{t_0,t_1}(x(t_0), x(t_1)) \\
&\quad = \underbrace{\sup_{f_0,f_1}\Big(f_0(x(t_0)) + f_1(x(t_1)) - \underbrace{\sup_z \big(f_0(z) + V(t_1)f_1(z) - I_0(z)\big)}_{\text{Lem 2.17}}\Big)}_{\text{Lem 2.16}} \\
&\quad = \sup_{g_0,f_1}\Big(g_0(x(t_0)) - V(t_1)f_1(x(t_0)) + f_1(x(t_1)) - \sup_z \big(g_0(z) - I_0(z)\big)\Big) \\
&\quad = \sup_{g_0}\inf_z \big(g_0(x(t_0)) - g_0(z) + I_0(z)\big) + \sup_{f_1}\big(f_1(x(t_1) - V(t_1)f_1(x(t_0))\big) \\
&\quad =: I_0(x(t_0)) + I^V_{\Delta t}(x(t_1) \mid x(t_0)),
\end{aligned}$$

where in the second equality we define $g_0 := f_0 + V(t_1)f_1$ for simplicity, and in the last equality we have

$$I_0(x(t_0)) = \sup_{g_0}\inf_z \big(g_0(x(t_0)) - g_0(z) + I_0(z)\big)$$

and

$$I^V_{\Delta t}(x(t_1)|x(t_0)) = \sup_{f_1}\big(f_1(x(t_1) - V(t_1)f_1(x(t_0))\big), \quad \Delta t = t_1 - t_0.$$

By induction, we get LDP for the finite-dimensional distributions of $(X_n(0), X_n(t_1), \ldots, X_n(t_k))$ with rate function with the projective limit form:

$$I^V_{t_0\ldots t_k}(x(t_0), \ldots, x(t_k)) = I^V_0(x(t_0)) + \sum_{i=0}^{k} I^V_{t_{i+1}-t_i}(x(t_{i+1})|x(t_i)).$$

Step 3: We are left to establish item (b) of step 2, $\|V_n(t)f - V(t)f\| \to 0$ for any $t \geq 0$ and $n \to \infty$. To do this, we achieve the goal by the Trotter-Kato-Kurtz theorem, Lemma 2.18 as below. From the theorem, we need to check the following conditions for any $f \in C_b(E)$:

(a) H_n is the generator of semigroup V_n;

(b) H is the generator of semigroup V;

(c) $\lim_{n\to\infty} \|V_n(t)f - (\mathbb{1} - \frac{t}{n}H_n)^{-n}f\| = 0$;

(d) $\lim_{n\to\infty} \|V(t)f - (\mathbb{1} - \frac{t}{n}H)^{-n} f\| = 0$;

(e) $\lim_{n\to\infty} \|H_n - H\| = 0$.

The statement of items (a) and (c) is similar to items (b) and (d), but the proof is completely different. Item (a), it is obtained in (2.2). For item (c), we obtain it by the semigroup generation theorem.

We cite the Crandall-Liggett theorem, Lemma 2.19 as below, to show item (b) and (d) by modifying item (e). Two conditions need to be verified, dissipativity and the range condition. For the first one, the dissipativity of H holds since the operator H_n is dissipative.

The second one is the range condition: for sufficiently many $h \in C_b(E)$ and all $\lambda > 0$ one can find an $f \in \mathcal{D}(H)$ that solves the equation $f - \lambda Hf = h$ in the classical sense. Moreover, if H is dissipative, then such a solution is unique.

However, for nonlinear equations, the verification of the range condition is very hard and it was observed early on [CL83] that viscosity solutions can be used to replace classical solutions. By weakening the type of solution needed for $(\mathbb{1} - \lambda H)f = h$, we have to require a strong form of uniqueness condition known as the comparison principle. Informally, this principle states that, if upper semicontinuous $\overline{f}$ and lower semicontinuous $\underline{f}$ satisfy

$$(\mathbb{1} - \lambda H)\overline{f} \leq h \quad \text{and} \quad (\mathbb{1} - \lambda H)\underline{f} \geq h, \tag{2.17}$$

then $\overline{f} \leq \underline{f}$. The $\overline{f}$ and $\underline{f}$ are called, respectively, a viscosity subsolution and a viscosity supersolution, and are not necessarily in the domain of H.

This provides an opportunity for further relaxation of conditions. If (2.17) holds, we can introduce two more operators: H_0, H_1 such that $Hf \leq H_0 f$ and $Hf \geq H_1 f$ for all $f \in \mathcal{D}(H) \cap \mathcal{D}(H_i)$. Then

$$(\mathbb{1} - \lambda H_0)\overline{f} \leq h \quad \text{and} \quad (\mathbb{1} - \lambda H_1)\underline{f} \geq h.$$

Later on, formulate H_0, H_1 in terms of a Lyapunov function to restrict further analysis to compact sets. It suffices to establish the comparison principle for H_0, H_1 in the sense of viscosity solutions.

Next, we turn to the existence of viscosity solutions using the Barles-Perthame procedure. The construction of $\underline{f}$, $\overline{f}$ by the Barles–Perthame procedure then reveals that $\overline{f} = \underline{f} = f \in C_b(E)$. Hence, each h uniquely corresponds to an $f \in C_b(E)$, and we can denote it by $f = R(\lambda)h$. Consequently, at least formally, $R(\lambda) = (\mathbb{1} - \lambda H)^{-1}$. In other words, H_0, H_1 implicitly determine H through its resolvent. We can now completely avoid using item (e) $\lim_{n\to\infty} \|H_n - H\| = 0$, and we replace it by: for each

$$H_1 f \leq \liminf_{n\to\infty} H_n f_n, \quad \limsup_{n\to\infty} H_n f_n \leq H_0 f \quad \text{some} \quad f_n \to f,$$

in the strongly uniform sense.

From the above analysis, the conditions items (a) - (e) reduce to establish that viscosity solutions of $f-\lambda H_n f = h$ converge to viscosity solutions of $f-\lambda H f = h$.

To do this, by the Barles-Perthame procedure, for every $x \in E$, there exists a sequence $x_n \in E$ such that $\lim_{n\to\infty} x_n = x$, one can show that

$$u(x) := \sup\left\{ \limsup_{n\to\infty} R_n(\lambda)h(x_n) \mid \lim_{n\to\infty} x_n = x \right\},$$

$$v(x) := \inf\left\{ \liminf_{n\to\infty} R_n(\lambda)h(x_n) \mid \lim_{n\to\infty} x_n = x \right\},$$

are a viscosity subsolution and a viscosity supersolution to $f - \lambda H f = h$. It is obvious that $u \geq v$ from the construction. In addition, from item (d) of Proposition 2.14: the comparison principle is satisfied, we obtain that $u = v = R(\lambda)h$ which is the unique viscosity solution. Next, using this solution, we can extend the domain of the operator H by adding all pairs of the form $(R(\lambda)h, \lambda^{-1}(R(\lambda)h - h))$ to the graph of H to obtain a new operator $\hat{H}$:

$$\hat{H} = \left\{ \left(R(\lambda)h, \lambda^{-1}(R(\lambda)h - h)\right) \,\middle|\, \lambda > 0, h \in C_b(X) \right\}.$$

Furthermore, we prove this extension operator $\hat{H}$ satisfying the conditions of the Crandall-Liggett theorem. Firstly, the range condition is held by construction. Secondly, $\hat{H}$ is a dissipative operator as it satisfies the positive maximum principle, Lemma 2.22 as below, and [Kra22] have proven it on proposition 4.10.

Using the Crandall-Liggett theorem once again, $\hat{H}$ generates a semigroup $V(t)$. Subsequently, thanks to the Trotter-Kato-Kurtz theorem, we achieve the goal $\|V_n f(t) - V f(t)\| \to 0$ for any $t \geq 0$, $f \in \overline{\mathcal{D}(\hat{H})}$, as $n \to \infty$.

Step 4: From steps 1, 2, and 3, we obtain that the LDP is satisfied and with a rate function in the projective limit form (2.14). However, the rate function is only implicitly characterized by $V(t)$. This is why we establish a Lagrangian form rate function (2.15) based on (2.14). To do so it is sufficient to prove that for any $f \in C_b(E)$,

$$V(t)f = \mathbf{V}(t)f \tag{2.18}$$

by the method of resolvent approximation, where $\mathbf{V}(t)$ is defined in (2.10).

To do this, we connect the variational semigroup to the resolvent. We first recall the Nisio semigroup of (2.11):

$$\mathbf{R}(\lambda)h(x) = \sup_{\substack{\gamma\in\mathcal{AC}\\ \gamma(0)=x}} \int_0^\infty \lambda^{-1} \mathrm{e}^{-\lambda^{-1}t}\big(h(\gamma(t)) - \int_0^t \mathcal{L}(\gamma(s),\dot{\gamma}(s))\mathrm{d}s\big)\mathrm{d}t.$$

If conditions 8.9, 8.10, and 8.11 in [FK06] hold, we obtain the following important results.

- For $\lambda > 0$, we have $\mathbf{R}(\lambda)C_b(E) \subseteq C_b(E)$ and $\mathbf{R}(\lambda)h$ is a viscosity solution of $f - \lambda\mathbf{H}f = h$, following the first part of the proof of Lemma 2.23;
- for $f \in C_b(E)$,
$$\lim_{n\to\infty} \|\mathbf{R}(t/n)^n f - \mathbf{V}(t)f\| = 0, \tag{2.19}$$
thanks to Lemma 2.24.

In addition, by combining the comparison principle, we get
$$\mathbf{R}(\lambda)h = R(\lambda)h, \tag{2.20}$$
which is proposition 2.14 of Proposition 2.14.

Subsequently, from (2.13), (2.19) and (2.20), we obtain $V(t)f = \mathbf{V}(t)f$.

In summary, it is proven that conditions 8.9, 8.10, and 8.11 in [FK06] are prerequisites for obtaining the action-integral rate function.

Next, we give the auxiliary lemmas used in Steps 1-4 in the order of occurrence.

Lemma 2.15 (Projective limit theorem, Theorem 4.28 in [FK06])**.** *Assume that $\{X_n\}$ is exponentially tight in $\mathcal{D}_E[0,\infty)$ and that for each $0 \le t_1 < t_2 < \cdots < t_m$, $\{(X_n(t_1),\ldots,X_n(t_m))\}$ satisfies the large deviation principle in E^m with rate function $I_{t_1,\ldots,t_m}$. Then $\{X_n\}$ satisfies the large deviation principle in $\mathcal{D}_E[0,\infty)$ with good rate function*
$$I(x) = \sup_{\{t_i\}\subset\Delta_x^c} I_{t_1,\ldots,t_m}((x(t_1),\ldots,x(t_m))),$$
where $\{t_i\}$ is shorthand for all sets of the form $\{t_1,t_2,\ldots,t_m\}$ and Δ_x is the set of times where x is discontinuous.

Lemma 2.16 (Bryc's theorem, Proposition 3.25 in [FK06])**.** *Suppose $\{(X_n,Y_n)\}$ is exponentially tight in the product space $(S_1\times S_2, d_1+d_2)$. Let $\mu_n \in \mathcal{P}(S_1\times S_2)$ be the distribution of (X_n,Y_n) and let $\mu_n(\mathrm{d}x\times\mathrm{d}y) = \eta_n(\mathrm{d}y|x)\mu_n^1(\mathrm{d}x)$, that is, μ_n^1 is the S_1-marginal of μ_n and η_n gives the conditional distribution of Y_n given X_n. Suppose that for each $f \in C_b(S_2)$*
$$\Lambda_2(f|x) = \lim_{n\to\infty}\frac{1}{n}\log\int_{S_2} \mathrm{e}^{nf(y)}\eta_n(\mathrm{d}y|x)$$
exists, that the convergence is uniform for x in compact subsets of S_1, and that $\Lambda_2(f|x)$ is a continuous function of x. For $x \in S_1$ and $y \in S_2$, define
$$I_2(y|x) = \sup_{f\in C_b(S_2)} (f(y) - \Lambda_2(f|x)).$$
If $\{X_n\}$ satisfies the large deviation principle with good rate function I_1, then $\{(X_n,Y_n)\}$ satisfies the large deviation principle with good rate function
$$I(x,y) = I_1(x) + I_2(y|x).$$

Lemma 2.17 (Varadhan's Lemma, Theorem III.13 in [dH08])**.** *Let (P_n) satisfy the LDP on $\mathcal{X}$ with rate n and with rate function I. Let $F_n : \mathcal{X} \to \mathbb{R}$ be a continuous function that is bounded from above and $\|F_n - F\| \to 0$ when $n \to \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{X}} \mathrm{e}^{nF_n(x)} P_n(\mathrm{d}x) = \sup_{x\in\mathcal{X}} (F(x) - I(x)).$$

The following theorem is a simplification of Proposition 5.5 in [FK06].

Lemma 2.18 (Trotter-Kato-Kurtz theorem, Proposition 5.5 in [FK06])**.** *Let E be a Polish space and let $H_n : C_b(E) \to C_b(E)$ and $H : \mathcal{D}(H) \subseteq C_b(E) \to C_b(E)$ be dissipative operators that satisfy the range condition with the same λ. Let $V_n(t)$ and $V(t)$ be the corresponding generated semigroups in the Crandall-Liggett sense. Suppose that the following:*

For each $f \in \mathcal{D}(H)$, there exist $f_n \in \mathcal{D}(H_n)$ such that

$$\|f - f_n\| \overset{n\to\infty}{\longrightarrow} 0 \quad \text{and} \quad \|Hf - H_n f_n\| \overset{n\to\infty}{\longrightarrow} 0.$$

Then for any $f \in \overline{\mathcal{D}(H)}$ and $f_n \in C_b(E)$ such that $\|f - f_n\| \to 0$, we have

$$\|V(t)f - V_n(t)f_n\| \overset{n\to\infty}{\longrightarrow} 0.$$

Lemma 2.19 (Crandall-Liggett theorem in [CL71])**.** *Let H be an operator on a Banach space X. Suppose that*

(a) *H is dissipative. We say $H \subseteq C_b(E) \times C_b(E \times S)$ is dissipative if for all (f_1, g_1), $(f_2, g_2) \in H$ and $\lambda > 0$ we have*

$$\|f_1 - \lambda g_1 - (f_2 - \lambda g_2)\| \geq \|f_1 - f_2\|;$$

(b) *H satisfies the range condition. We say $H \subseteq C_b(E) \times C_b(E \times S)$ satisfies the range condition if for all $\lambda > 0$ we have: the uniform closure of $\mathcal{D}(H)$ is a subset of $\mathcal{R}(\mathbb{1} - \lambda H)$.*

We denote by $R(\lambda) = (\mathbb{1} - \lambda H)^{-1}$. Then there is a strongly continuous contraction semigroup $V(t)$ defined on the uniform closure of $\mathcal{D}(H)$ and for all $t \geq 0$ and f in the uniform closure of $\mathcal{D}(H)$

$$\lim_{n\to\infty} \|R(t/n)^n f - V(t)f\| = 0.$$

Condition 2.20 (Condition 3.1 in [Kra20])**.** $A_n \subset C_b(E) \times C_b(E)$ is an operator such that the martingale problem for $A_n \subset C_b(E) \times C_b(E)$ is well-posed. Denote by $\mathbb{P}_x \in \mathcal{P}(D_E(\mathbb{R}^+))$ the solution that satisfies $X(0) = x$, $\mathbb{P}_x$ almost surely. The map $x \mapsto \mathbb{P}_x$ is assumed to be continuous for the weak topology on $\mathcal{P} = \mathcal{P}(D_E(\mathbb{R}^+))$.

Lemma 2.21 (Theorem 3.6 in [Kra20]). *Let Condition 2.20 be satisfied. For each $h \in C_b(E)$ and $\lambda > 0$ the function $R_n(\lambda)h$ is a viscosity solution to $f - \lambda H_n f = h$.*

Lemma 2.22 (The positive maximum principle for nonlinear generator in [Kra16]). *For any (f_1, g_1), $(f_2, g_2) \in H$ and $\lambda > 0$, there exist sequences $x_n \in C_b(E)$ satisfied dissipativity condition it is equivalent to (a) and (b).*

(a) *If $x_0 \in E$ is such that*

$$f_1(x_0) - f_2(x_0) = \sup_{x_n \in E} f_1(x_n) - f_2(x_n),$$

then $g_1(x_0) - g_2(x_0) \leq 0$;

(b) *If $x_0 \in E$ is such that*

$$f_1(x_0) - f_2(x_0) = \inf_{x_n \in E} f_1(x_n) - f_2(x_n),$$

then $g_1(x_0) - g_2(x_0) \geq 0$.

Lemma 2.23 (Theorem 8.27 in [FK06]). *Let (E, r) and (U, q) be complete, separable metric spaces. Suppose that $A \subset C_b(E) \times C(E \times U)$ and $\mathcal{L} : E \times U \to [0, \infty]$ satisfy Conditions 8.9, 8.10, and 8.11 of [FK06]. Define*

$$\mathbf{H}f(x) = \sup_{u \in \Gamma_x} (Af(x, \mu) - \mathcal{L}(x, u))$$

with $\mathcal{D}(H) = \mathcal{D}(A)$. Then

(a) *For each $h \in D_\alpha$,*

$$\begin{aligned}\mathbf{R}_\alpha h(x_0) := \sup_{\substack{\gamma \in \mathcal{AC} \\ \gamma(0) = x_0}} \{ \int_0^\infty & \alpha^{-1} \mathrm{e}^{-\alpha^{-1} t} h(\gamma(t)) \mathrm{d}t \\ & - \int_0^\infty \alpha^{-1} \mathrm{e}^{-\alpha^{-1} t} \int_0^t \mathcal{L}(\gamma(s), \dot{\gamma}(s)) \mathrm{d}s \mathrm{d}t \}\end{aligned}$$

is continuous and is a solution of $f - \alpha \mathbf{H} f = h$.

Lemma 2.24 (Lemma 8.18 in [FK06]). *Suppose Conditions 8.9 and 8.10 hold, and let $\{\mathbf{V}(t)\}$ be defined by (8.10). Then for each $f \in C_b(E)$ and each $x_0 \in E$,*

$$\mathbf{V}(t) f(x_0) = \lim_{n \to \infty} \mathbf{R}(t/n)^n f(x_0).$$

Chapter 3

Large deviations with finite fast switching

In this chapter, we study the large deviations for Cox-Ingersoll-Ross (CIR) processes with small noise and state-dependent fast switching via associated Hamilton-Jacobi-Bellman equations. As the separation of time scales, when the noise goes to 0 and the rate of switching goes to ∞, we get a limit equation characterized by the averaging principle. Moreover, we prove the large deviation principle with an action-integral form rate function to describe the asymptotic behavior of such systems. The new ingredient is establishing the comparison principle in the singular context. The proof is carried out using the nonlinear semigroup method introduced in Chapter 2.

This chapter is based on

[HKX23] Yanyan Hu, Richard C. Kraaij, and Fubao Xi. Large deviations for Cox-Ingersoll-Ross processes with state-dependent fast switching, 2023

3.1 Introduction

The classical Cox-Ingersoll-Ross (CIR) process was proposed by John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross in [CIR05, CIR85]. It is an important tool for modeling the stochastic evolution of interest rates and has widespread applications in the field of finance, especially in the stock market. In general, it is described as

$$\mathrm{d}X(t) = \eta(\mu - X(t))\mathrm{d}t + \theta\sqrt{X(t)}\mathrm{d}W(t),$$

where $X(t)$ stands for the instantaneous interest rate at time t; $\eta > 0$ is the rate of mean reversion; μ represents the mean of the interest rate; $\theta > 0$ is the standard deviation of the interest rate; $(W(t))_{t\geq 0}$ is a real value Brownian motion.

In the real world, motivated by the increasing demand for modeling complex systems, in which structural changes, small fluctuations as well as big spikes coexistence are intertwined, we realize that the classical CIR process is lacking the desired complexity. Moreover, an instructive example in a stock market is that equity investors can be classified as belonging to two categories, long-term investors and short-term investors. Long-term investors consider a relatively long time horizon and make decisions based on the weekly or monthly performance of the stock, whereas short-term investors, such as day traders, focus on returns in the short term, daily, or even shorter periods. Their time scales are in sharp contrast, and we call it the *two time-scale systems* or *slow-fast systems*. Hence, we add switching to CIR processes that can have mutual impacts, and if we adjust the frequency of the switching may cause a separation of scale. In this chapter, for $E = (0, \infty)$ and $S = \{1, 2, \dots, N\}$, $N < \infty$, we study CIR processes with small noise and fast switching on $E \times S$

$$\begin{cases} \mathrm{d}X_n^\varepsilon(t) = \eta(\mu(\Lambda_n^\varepsilon(t)) - X_n^\varepsilon(t))\mathrm{d}t + \frac{1}{\sqrt{n}}\theta\sqrt{X_n^\varepsilon(t)}\mathrm{d}W(t), \\ (X_n^\varepsilon(0), \Lambda_n^\varepsilon(0)) = (x_0, k_0) \in E \times S, \end{cases} \tag{3.1}$$

where the fast process $\Lambda_n^\varepsilon(t)$ is a jumping process on S satisfying

$$\mathbb{P}(\Lambda_n^\varepsilon(t+\triangle) = j \mid \Lambda_n^\varepsilon(t) = i, X_n^\varepsilon(t) = x) = \begin{cases} \frac{1}{\varepsilon}q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j \neq i, \\ 1 + \frac{1}{\varepsilon}q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j = i, \end{cases} \tag{3.2}$$

for $\Delta > 0$, $i, j \in S$, $x \in E$, and $\varepsilon > 0$ is a small parameter.

The key feature of such slow-fast systems is that the fast process reaches its equilibrium state at much shorter time scales while the slow system effectively remains unchanged. The local equilibration phenomenon allows the approximation of the properties of the slow system by averaging out the coefficients over the local stationary distributions of the fast process. Such approximations yield a significant model simplification and are mathematically justified by establishing an appropriate *averaging principle*. Hence, when $n \to \infty$ and $\varepsilon \to 0$, the system (3.1) and (3.2) is averaged under the law of large number scaling, we can identify that the limit equation is

$$\mathrm{d}\overline{X}(t) = \eta\bigg(\sum_{i\in S}\mu(i)\pi_i^x(t) - \overline{X}(t)\bigg)\mathrm{d}t,$$

where $\pi^x(t) = (\pi_i^x(t))_{i\in S}$ is the stationary distribution of fast processes depending on the position of $X_n^\varepsilon(t) = x$.

In this setting, we need quantification of how well, the averaging principle applies to a specific problem. One of the ways to quantify this approximation is *large deviation principle* (LDP) of the Markov processes $(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t))$. In the following, we first conduct an informal analysis.

Due to the scale separation phenomenon of slow-fast systems, there are two types of LDP. We first have the Donsker-Varadhan LDP for the occupation measures of the fast process $\Lambda_n^\varepsilon(t)$ around $\pi(t)$ when $X_n^\varepsilon(t)$ is close to $x(t)$:

$$\mathbb{P}\bigg(\frac{1}{\mathrm{d}t}\int_t^{t+\mathrm{d}t}\delta_{\Lambda_n^\varepsilon(s)}\mathrm{d}s|_{t\geq 0}\approx\pi(t)|_{t\geq 0}\bigg|X_n^\varepsilon(t)|_{t\geq 0}=x(t)|_{t\geq 0}\bigg)\backsim \mathrm{e}^{-\frac{1}{\varepsilon}\tilde{I}_x(\pi)}, \tag{3.3}$$

where

$$\tilde{I}_x(\pi)=-\inf_{g\gg 0}\int\frac{R_x g}{g}\mathrm{d}\pi,$$

where R_x is the generator of a state-dependent switching:

$$R_x g(z)=\sum_{j\in S}q_{zj}(x)\left(g(j)-g(z)\right),$$

where $x(t):=x$.

Furthermore, we find that the Freidlin-Wentzel LDP of the slow process $X_n^\varepsilon(t)$ is obtained under the condition that the fast process reaches $\pi(t)$, and has

$$\mathbb{P}\bigg(\dot{X}_n^\varepsilon(t)|_{t\geq 0}\approx\dot{\rho}(t)|_{t\geq 0}\bigg|\frac{1}{\mathrm{d}t}\int_t^{t+\mathrm{d}t}\delta_{\Lambda_n^\varepsilon(s)}\mathrm{d}s|_{t\geq 0}\approx\pi(t)|_{t\geq 0},X_n^\varepsilon(s)|_{s\in(0,t]}=x(s)|_{s\in(0,t]}\bigg)\backsim e^{-n\hat{I}(\rho|\pi)}, \tag{3.4}$$

where

$$\hat{I}(\rho|\pi)=\min_{\dot{\rho}(t)=\sum_{i=1}^N v_i\pi_i(t)}\sum_{i=1}^N\frac{|v_i-\eta(\mu(i)-x(t))|^2}{2\theta^2 x(t)}\pi_i(t).$$

To analyze the system (3.1) and (3.2) from the point of view of a long-term investor, we need to consider the LDP of both fast and slow processes $(X_n^\varepsilon(t),\Lambda_n^\varepsilon(t))$ at time t to maximize profits. That is formally

$$\begin{aligned}
&\mathbb{P}\left(\dot{X}_n^\varepsilon(t)|_{t\geq 0}\approx\dot{\rho}(t)|_{t\geq 0},\frac{1}{\mathrm{d}t}\int_t^{t+\mathrm{d}t}\delta_{\Lambda_n^\varepsilon(s)}\mathrm{d}s|_{t\geq 0}\approx\pi(t)|_{t\geq 0}\bigg|X_n^\varepsilon(s)|_{s\in(0,t]}=x(s)|_{s\in(0,t]}\right)\\
&=\mathbb{P}\left(\dot{X}_n^\varepsilon(t)|_{t\geq 0}\approx\dot{\rho}(t)|_{t\geq 0}\bigg|\frac{1}{\mathrm{d}t}\int_t^{t+\mathrm{d}t}\delta_{\Lambda_n^\varepsilon(s)}\mathrm{d}s|_{t\geq 0}\approx\pi(t)|_{t\geq 0},X_n^\varepsilon(s)|_{s\in[0,t]}=x(s)|_{s\in[0,t]}\right)\\
&\quad\times\mathbb{P}\left(\frac{1}{\mathrm{d}t}\int_t^{t+\mathrm{d}t}\delta_{\Lambda_n^\varepsilon(s)}\mathrm{d}s|_{t\geq 0}\approx\pi(t)|_{t\geq 0}\bigg|X_n^\varepsilon(s)|_{s\in(0,t]}=x(s)|_{s\in(0,t]}\right)\\
&=\exp\left\{-\left(n\hat{I}(\rho|\pi)+\frac{1}{\varepsilon}\tilde{I}_x(\pi)\right)\right\},
\end{aligned} \tag{3.5}$$

where in the last line, we use the results of Donsker-Varadhan LDP (3.3) and Freidlin-Wentzel LDP (3.4). Moreover, from (3.5) and the contraction principle [DZ98, Theorem 4.2.1], we obtain

$$\mathbb{P}\left(\dot{X}_n^\varepsilon(t)|_{t\geq 0}\approx\dot{\rho}(t)|_{t\geq 0}\bigg|X_n^\varepsilon(s)|_{s\in(0,t]}=x(s)|_{s\in(0,t]}\right)\backsim\exp\left\{-\inf_\pi\left(n\hat{I}(\rho|\pi)+\frac{1}{\varepsilon}\tilde{I}_x(\pi)\right)\right\}.$$

For ease of analysis in subsequent steps, we define a set

$$G=\left\{\dot{X}_n^{\varepsilon}(t)|_{t\geq 0}\approx\dot{\rho}(t)|_{t\geq 0}\middle|X_n^{\varepsilon}(s)|_{s\in(0,t]}=x(s)|_{s\in(0,t]}\right\}.$$

Hence, we get

$$-\log\mathbb{P}(G)\backsim\inf_{\pi}\left(n\hat{I}(\rho|\pi)+\frac{1}{\varepsilon}\ \tilde{I}_x(\pi)\right). \tag{3.6}$$

In (3.1) the intensity of the multiplicative noise and in (3.2) the frequency of the fast random switching may have different ratios n^{-1}/ε when $n\to\infty$ or $\varepsilon\to 0$. To determine the optimal LDP's convergence speed of $(X_n^{\varepsilon}(t),\Lambda_n^{\varepsilon}(t))$ at time t, it is necessary to analyse (3.6) in three different scenarios:

Case 1 $\varepsilon\ll\frac{1}{n}$: we have a Donsker-Varadhan type LDP

$$-\varepsilon\log\mathbb{P}(G)\backsim\inf_{\pi}\left(n\varepsilon\hat{I}(\rho|\pi)+\tilde{I}_x(\pi)\right)\to\inf_{\pi}\tilde{I}_x(\pi);$$

Case 2 $\frac{1}{n}\ll\varepsilon$: we have a Freidlin-Wentzell type LDP

$$-\frac{1}{n}\log\mathbb{P}(G)\backsim\inf_{\pi}\left(\hat{I}(\rho\mid\pi)+\frac{1}{n\varepsilon}\tilde{I}_x(\pi)\right)\to\inf_{\pi}\hat{I}(\rho\mid\pi);$$

Case 3 $\varepsilon=\frac{1}{n}$: we have the combination of Donsker-Varadhan LDP and Freidlin-Wentzell LDP

$$\begin{aligned}-\varepsilon\log\mathbb{P}(G)&\backsim\inf_{\pi}\left(\hat{I}(\rho\mid\pi)+\tilde{I}_x(\pi)\right)\\&=\inf_{\pi}\left(\min_{\dot{\rho}(t)=\sum_{i=1}^N v_i\pi_i(t)}\sum_{i=1}^N\frac{|v_i-\eta(\mu(i)-x(t))|^2}{2\theta^2x(t)}\pi_i(t)+\tilde{I}_x(\pi)\right).\end{aligned}$$

In this chapter, we treat the most complex case, namely Case 3, $\varepsilon=\frac{1}{n}$, which the two LDP's are completed at the same scale.

For the proof, we use Feng and Kurtz's method based on Hamilton-Jacobi theory and control theory, which has been developed to study LDP associated with a sequence of Markov processes. Firstly, the advantage of this method is that the operator convergence treats both the classical Freidlin-Wentzell theory and the Donsker-Varadhan theory within one framework. Secondly, Feng and Kurtz's method deals with the difficulties caused by nonlinear operators using viscosity solutions. Inspired by this approach, Peletier and Schlottke [PS24] studied a stochastic differential equation with finite state fast switching on the flat torus, but the diffusion coefficient was additive noise. Subsequently, Kraaij and Schlottke [KS20] investigated the LDP of the slow-fast system by giving the abstract generator with uniformly elliptic conditions. In addition, Feng, Fouque and Kumar [FFK12] studied the small-time large deviation for fast mean-reverting stochastic volatility models using the simplified Feng and Kurtz's method due to this system living in a compact space. Huang, Mandjes, and Spreij [HMS16] were also interested in the joint sample-path large deviations principle for the Markov-modulated diffusion process and the occupation measure of the Markov chain. However, as mentioned earlier, all the work mainly considered LDP in a regular setting, we now consider a singular setting which causes the non-compact domain of the slow process. Because of this, Euclidean techniques to study the large deviation principle fail. Alternatively, the authors in [DFL11] take a Riemann point of view to analyze the associated Hamilton-Jacobi equations, and we extend their insights to the two time-scale contexts by adding switching.

To conclude, we investigate the LDP for CIR processes with state-dependent fast-switching, by associated Hamilton-Jacobi and Hamilton-Jacobi-Bellman equations, which are the primary tools we need. Our specific technical roadmap is as follows: we begin with using Skorokhod's representation to give an integral form of the fast-switching process and obtain strong nonnegative solutions of the CIR processes with fast switching by pathwise splicing. Then, we modify the technique introduced in the book [FK06] of Feng and Kurtz to

(a) verify convergence of the sequence of nonlinear operators H_n to a multi-valued limit operator H. We reduce H to $\mathbf{H}$ by solving an eigenvalue problem, in which we effectively find an optimal stationary measure most notably;

(b) verify exponential tightness on the "path-space" as the CIR process is a diffusion process equation on $(0,\infty)$ with a singularity at 0 leading to a non-compact space.

(c) verify the comparison principle for the nonlinear multi-valued limiting operator H, which is hard to prove but plays a prominent role. We achieve it by connecting viscosity solutions for H to those for $\mathbf{H}$ and prove comparison principle for $f - \lambda \mathbf{H} f = h$, $\lambda > 0$.

(d) construct a variational representation for $\mathbf{H}$, which gives the rate function with the action-integral from.

3.1.1 CIR processes with finite state-dependent fast switching

Through the introduction, we have learned about the process to be studied and the questions to be explored. Below, we will provide a detailed explanation. First, we will introduce the process: CIR processes with finite state-dependent fast switching on $E \times S$

$$\begin{cases} \mathrm{d}X_n^\varepsilon(t) = \eta(\mu(\Lambda_n^\varepsilon(t)) - X_n^\varepsilon(t))\mathrm{d}t + n^{-\frac{1}{2}}\theta\sqrt{X_n^\varepsilon(t)}\mathrm{d}W(t), \\ (X_n^\varepsilon(0), \Lambda_n^\varepsilon(0)) = (x_0, k_0) \in E \times S, \end{cases} \tag{3.7}$$

where the fast process $\Lambda_n^\varepsilon(t)$ is a jumping-process on S satisfying

$$\mathbb{P}(\Lambda_n^\varepsilon(t+\triangle) = j \mid \Lambda_n^\varepsilon(t) = i, X_n^\varepsilon(t) = x) = \begin{cases} \frac{1}{\varepsilon}q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j \neq i, \\ 1 + \frac{1}{\varepsilon}q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j = i, \end{cases} \tag{3.8}$$

for $\Delta > 0$, $i, j \in S$, $x \in E$, and $\varepsilon > 0$ is a small parameter. The system $(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t))$ is a Markov process.

The slow process $X_n^\varepsilon(t)$ is mean-reverting singular diffusion. $X_n^\varepsilon(t)$ stands for the instantaneous interest rate at time t; $\eta > 0$ is the rate of mean reversion; for any $i \in S$, $\mu(i)$ represents the mean of the interest rate; $\theta > 0$ is the standard deviation of the interest rate. $(W(t))_{t\geq 0}$ is a real value Brownian motion defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with the filtration $(\mathscr{F}_t)_{t\geq 0}$ satisfying the usual condition (i.e., $\mathscr{F}_0$ contains all $\mathbb{P}$-null sets and $\mathscr{F}_t = \mathscr{F}_{t+} := \bigcap_{s>t}\mathscr{F}_s$).

The fast process $\Lambda_n^\varepsilon(t)$ is a finite state-dependent switching process. In particular, if $S = \{1\}$, (3.7) is often used to characterize the interest rate in finance which is called classical CIR processes without switching.

Before studying the process $(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t))$, we first give a result on the existence and uniqueness of the process.
Proposition 3.1 (Existence and uniqueness)**.** *For every i, assume that $2\eta\mu(i) \geq \theta^2$, then the systems* (3.7) *and* (3.8) *have a nonnegative unique strong solution $(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t))$ with initial value $(X_n^\varepsilon(0), \Lambda_n^\varepsilon(0)) = (x_0, k_0)$, and $(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t))$ is non-explosive.*

The proof of Proposition 3.1 is deferred to Section 3.6.

3.1.2 Main results

Before giving a main result, we set the assumptions that will be necessary for the main result.
Assumption 3.1. *Let $\varepsilon = \frac{1}{n}$, this shows that small disturbance and fast switching have the same rate.*
Assumption 3.2. *For any $x \in E$, $(q_{ij}(x))_{i,j\in S}$ is a conservative, irreducible transition rate matrix, and $\sup_{i\in S}\sum_{j\in S, j\neq i} q_{ij}(x) < \infty$.*
Assumption 3.3. *There exists a constant $C > 0$ such that*

$$|q_{ij}(x) - q_{ij}(y)| \leq C|x-y|, \quad x, y \in E, \ i, j \in S.$$

Remark 3.2. If Assumption 3.1 is satisfied, (3.7) and (3.8) become

$$\begin{cases} \mathrm{d}X_n(t) = \eta(\mu(\Lambda_n(t)) - X_n(t))\mathrm{d}t + n^{-\frac{1}{2}}\theta\sqrt{X_n(t)}\mathrm{d}W(t), \\ (X_n(0), \Lambda_n(0)) = (x_0, k_0) \in E \times S, \end{cases} \tag{3.9}$$

and

$$\mathbb{P}(\Lambda_n(t+\triangle) = j \mid \Lambda_n(t) = i,\ X_n(t) = x) = \begin{cases} nq_{ij}(x)\triangle + \circ(\triangle), & \text{if } j \neq i, \\ 1 + nq_{ij}(x)\triangle + \circ(\triangle), & \text{if } j = i. \end{cases} \tag{3.10}$$

From now on, except for the Section 3.6 we use $(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t))$ instead of $(X_n(t), \Lambda_n(t))$. Assumption 3.2 guarantees that there exists a unique stationary distribution $\pi^x(t) = (\pi_i^x(t))_{i\in S}$ for the fast process $\Lambda_n(t)$ if slow process is fixed at x. Moreover, the system (3.9) and (3.10) will be averaged according to the law of large numbers limit, and the limit equation is

$$\mathrm{d}\overline{X}(t) = \eta\bigg(\sum_{i\in S} \mu(i)\pi_i^x(t) - \overline{X}(t)\bigg)\mathrm{d}t.$$

We need Assumption 3.3 to prove the comparison principle for technical reasons.

Here, we state the path large deviation principles of the Markov process $(X_n(t), \Lambda_n(t))$, which is the main result in this chapter.

Theorem 3.3 (Large deviations for slow processes)**.** *Let $(X_n(t), \Lambda_n(t))$ be the Markov processes* (3.9) *and* (3.10) *on $E \times S$. Suppose that*

- *the large deviation principle holds for $X_n(0)$ on E with speed n and a good rate function I_0;*
- *Assumptions 3.1, 3.2 and 3.3 are satisfied.*

Then, the large deviation principle with speed n holds for $X_n(t)$ on $\mathcal{D}_E(\mathbb{R}^+)$ with a good rate function I having action-integral representation,

$$I(\gamma) = \begin{cases} I_0(\gamma(0)) + \int_0^\infty \mathcal{L}\left(\gamma(s), \dot\gamma(s)\right)\mathrm{d}s, & \textit{if } \gamma \in \mathcal{AC}(E), \\ \infty, & \textit{otherwise} \end{cases}$$

with $\mathcal{L}(x,v) := \sup_{p\in\mathbb{R}}\{\langle p, v\rangle - \mathcal{H}(x,p)\}$ which is the Legendre dual of $\mathcal{H}$ given by

$$\mathcal{H}(x,p) = \sup_{\pi\in\mathcal{P}(S)} \left\{ \int B_{x,p}(z)\pi(dz) - \mathcal{I}(x,\pi) \right\}, \tag{3.11}$$

where

$$B_{x,p}(i) = \eta(\mu(i) - x)p + \frac{1}{2}\theta^2 xp^2$$

coming from the slow process $X_n(t)$ and Donsker–Varadhan function

$$\mathcal{I}(x,\pi) = -\inf_{g>0} \int \frac{R_x g(z)}{g(z)}\pi(dz),$$

where R_x is the generator corresponding to the fast process $\Lambda_n(t)$ defined by

$$R_x g(z) = \sum_{j\in S} q_{zj}(x)\left(g(j) - g(z)\right).$$

The verification of Theorem 3.3 is based on extending Proposition 2.14 from the single process to slow-fast systems. In the process of implementation, we need to prove operator convergence, exponential tire tightness, comparison principle, and action-integral representation of rate function. We start with the operator convergence.

3.2 Operator convergence and principal-eigenvalue problem

In this section, we discuss how to verify the convergence of the nonlinear operator H_n and study the principal-eigenvalue problem.

3.2.1 Operator convergence

Let $C^2(E \times S; \mathbb{R}^+)$ denote the family of all nonnegative functions which are twice differentiable in the spatial variable. By Assumption 3.1, we focus on $(X_n(t), \Lambda_n(t))$, which is a Markov process whose generator is given by

$$\begin{aligned} A_n f(x,i) = \eta\left(\mu(i) - x\right)\partial_x f(x,i) + \frac{1}{2n}\theta^2 x \partial_{xx} f(x,i) \\ + n \sum_{j \in S} q_{ij}(x)\left(f(x,j) - f(x,i)\right). \end{aligned} \tag{3.12}$$

Based on the relation

$$H_n f = \frac{1}{n}\mathrm{e}^{-nf} A_n \mathrm{e}^{nf},$$

which is a Fleming's nonlinear generator; see [Fle78], we have

$$\begin{aligned} H_n f(x,i) = \eta\left(\mu(i) - x\right)\partial_x f(x,i) + \frac{1}{2}\theta^2 x \left(\partial_x f(x,i)\right)^2 + \frac{1}{2n}\theta^2 x \partial_{xx} f(x,i) \\ + \sum_{j \in S} q_{ij}(x)\big(\mathrm{e}^{n(f(x,j) - f(x,i))} - 1\big). \end{aligned}$$

For any $i \in S$, $H_n f(x,i)$ does not converge due to the divergence of the fourth term as $n \to \infty$. Hence, we can choose a suitable function sequence to deal with the divergent term. Let

$$f_n(x,i) = f(x) + \frac{1}{n}\phi(x,i), \quad \forall\ f \in C_b^2(E) \text{ and } \phi \in C_b^2(E \times S). \tag{3.13}$$

then we have

$$\begin{aligned} H_n f_n(x,i) = \eta\big(\mu(i) - x\big)\big(\partial_x f(x) + \frac{1}{n}\partial_x \phi(x,i)\big) + \frac{1}{2}\theta^2 x\big(\partial_x f(x) + \frac{1}{n}\partial_x \phi(x,i)\big)^2 \\ + \frac{1}{2n}\theta^2 x\big(\partial_{xx} f(x) + \frac{1}{n}\partial_{xx}\phi(x,i)\big) + \sum_{j \in S} q_{ij}(x)\big(\mathrm{e}^{\phi(x,j) - \phi(x,i)} - 1\big). \end{aligned} \tag{3.14}$$

Hence, there exists a limiting function $H_{f,\phi}(x,i)$ such that, for all $f \in \mathcal{D}(H)$ and $\phi \in C_b^2(E \times S)$

$$\lim_{n\to\infty} \|H_n f_n - H_{f,\phi}\| = 0.$$

where

$$\begin{aligned} H_{f,\phi}(x,i) = \eta\,(\mu(i)-x)\,\partial_x f(x) + \frac{1}{2}\theta^2 x\big(\partial_x f(x)\big)^2 \\ + \sum_{j\in S} q_{ij}(x)\big(\mathrm{e}^{\phi(x,j)-\phi(x,i)} - 1\big). \end{aligned} \tag{3.15}$$

We gather the important results in the following proposition.

Proposition 3.4 (Multi-valued limit Hamiltonian). *Let $(X_n(t), \Lambda_n(t))$ be a Markov process on $E \times S$ with generator A_n in (3.12). Set $H_n f_n$ as in (3.14) and*

$$H := \Big\{(f, H_{f,\phi}) \mid f \in C_b^2(E), H_{f,\phi} \in C_b(E\times S) \text{ and } \phi \in C_b^2(E\times S)\Big\}, \tag{3.16}$$

where

$$H_{f,\phi}(x,i) = b(x,i)\mathrm{d}f(x) + \frac{1}{2}(\mathrm{d}f(x))^2 + \sum_{j\in S} q_{ij}(x)[e^{\phi(x,j)-\phi(x,i)} - 1]. \tag{3.17}$$

Then, $H \subset ex-\lim_{n\to\infty} H_n$ in the sense of Definition 2.13.

3.2.2 Principal-eigenvalue problem

Building on the preparation in the previous subsection, we obtain a multi-valued limit H. We proceed to solve a principal eigenvalue problem to prove the comparison principle in Lemma 3.13. The eigenvalue problem is one in terms of fast processes.

Consider Equation (3.17) of Proposition 3.4, we have the following decompose: a function depending on i

$$B_{x,\partial_x f(x)}(i) := \eta(\mu(i)-x)\partial_x f(x) + \frac{1}{2}\theta^2 x(\partial_x f(x))^2$$

and the jump operator R_x acting on the state i,

$$R_x \mathrm{e}^{\phi(x,i)} := \sum_{j\in S} q_{ij}(x)(\mathrm{e}^{\phi(x,j)} - \mathrm{e}^{\phi(x,i)}).$$

We seek a $\overline{\phi}$ such that there is a constant $\mathcal{H}(x, \partial_x f(x))$ such that

$$\mathcal{H}(x,\partial_x f(x)) := B_{x,\partial_x f(x)}(i) + e^{-\overline{\phi}(x,i)} R_x e^{\overline{\phi}(x,i)}$$

is independent of i. Rewriting this equation in terms of $\overline{g} = e^{\overline{\phi}}$, we thus aim to find $\overline{g}$ and $\mathcal{H}(x, \mathrm{d}f(x))$ such that

$$\big(R_x + B_{x,\partial_x f(x)}\big)\,\overline{g}(i) = \mathcal{H}(x,\partial_x f(x))\overline{g}(i).$$

In other words, we aim to find the principal eigenfunction and eigenvalue for the operator $R_x + B_{x,\partial_x f(x)}$ in terms of i, which can be carried out using the Perron-Frobenius theorem and leads to the representation (3.11).

Proposition 3.5 (Principal-eigenvalue problem). *Let Assumptions 3.2 be satisfied.*

For each $(x, \partial_x f(x))$, there exist $\overline{g} > 0$ and a unique eigenvalue $\mathcal{H}(x, \partial_x f(x)) \in \mathbb{R}$ such that

$$\left(R_x + B_{x,\partial_x f(x)}\right)\overline{g} = \mathcal{H}(x, \partial_x f(x))\overline{g}, \tag{3.18}$$

with $\mathcal{H}(x, \partial_x f(x))$ given by

$$\begin{aligned} \mathcal{H}(x, \partial_x f(x)) &= \sup_{\pi\in\mathcal{P}(S)} \inf_{g>0} \int \frac{\left(R_x + B_{x,\partial_x f(x)}\right) g(i)}{g(i)} \pi(\mathrm{d}i) \\ &= \sup_{\pi\in\mathcal{P}(S)} \left\{ \int B_{x,\partial_x f(x)}(i)\pi(\mathrm{d}i) - \mathcal{I}(x,\pi) \right\} \end{aligned} \tag{3.19}$$

where

$$\mathcal{I}(x,\pi) = -\inf_{g>0} \int \frac{R_x g(i)}{g(i)} \pi(\mathrm{d}i). \tag{3.20}$$

Proof. Using Assumptions 3.2, from the Perron-Frobenius theorem in [DV75], we can obtain there exists a unique eigenvalue with associated eigenfunction which have the representation (3.19). □

3.3 Exponential tightness

In this section, we prove exponential tightness by applying [FK06, Corollary 4.17], which establishes exponential tightness based on the exponential compact containment condition and the convergence of the sequence H_n. The definition of the exponential compact containment condition of $(X_n(t), \Lambda_n(t))$ is similar to Definition 2.11, but should consider the fast process.

Definition 3.6 (Exponential compact containment of $(X_n(t), \Lambda_n(t))$). We say that a process $(X_n(t), \Lambda_n(t))$ satisfies the *exponential compact containment condition* at speed n, for every all compact $K_0 \subseteq E$, $T > 0$ and $a \geq 0$, there exists a compact set $K_{a,T} \subseteq E$ such that

$$\limsup_{n\to\infty} \sup_{(x_0,k_0)\in K_0\times S} \frac{1}{n} \log \mathbb{P}\left[(X_n(t), \Lambda_n(t)) \notin K_{a,T} \times S \text{ for some } t \leq T\right] \leq -a.$$

Here, we prove that $(X_n(t), \Lambda_n(t))$ satisfies the exponential compact containment condition. To do this, we need to find a containment function Υ, which plays the role of a Lyapunov function and allows our analysis to be restricted to compact regions in E. Here we give the rigorous definition and take a specific function Υ in our case.

Definition 3.7 (Containment function)**.** We say that a function $\Upsilon : E \to [0, \infty)$ is a containment function for $B_{x,p}$ if $\Upsilon \in C^1(E)$ and it is such that

- for every $C > 0$, the set $\{x \mid \Upsilon(x) \leq C\}$ is compact;
- $\sup_{x,i} B_{x,\partial_x \Upsilon(x)}(x, i) < \infty$.

Lemma 3.8. *The function*

$$\Upsilon(x) := -\log(x) + \log(1 + \frac{1}{2}x^2) - \log\sqrt{2} \tag{3.21}$$

is a containment function for $B_{x,p}$.

Proof. Firstly, we prove that Υ has compact sub-level sets. Note that 0 and ∞ are the boundary of E and the function $x \mapsto \Upsilon(x)$ goes to ∞ at the boundary points 0 and ∞, respectively. Regarding the second property, for any $x \in E$, we have

$$H((x,i), \partial_x \Upsilon(x)) = -\frac{1}{x}(\eta\mu(i) - \frac{\theta^2}{2}) + \eta < \infty, \tag{3.22}$$

and which boundedness condition follows with the constant

$$\sup_{x,i} -\frac{1}{x}(\eta\mu(i) - \frac{\theta^2}{2}) < \infty$$

and $2\eta\mu(i) \geq \theta^2$ for all $i \in S$. From (3.22), it follows that

$$C_\Upsilon := \sup_{x,i} B_{x,\partial_x \Upsilon(x)}(x, i) < \infty. \tag{3.23}$$

The proof is completed. □

Here, we are ready to give the following proposition that $(X_n(t), \Lambda_n(t))$ satisfies the exponential compact containment condition.

Proposition 3.9. *Let* $(X_n(t), \Lambda_n(t))$ *be a Markov process corresponding to* A_n. Υ *is a containment function in* (3.21)*. Suppose that the sequence* $(X_n(0), \Lambda_n(0))$ *is exponentially tight with speed* n*. Then the sequence* $(X_n(t), \Lambda_n(t))$ *satisfies the exponential compact containment condition with speed* n *as in Definition 3.6.*

Proof. We use the proof method coming from [FK06, Lemma 4.22]. The details are as follows.

Fix $a \geq 0$ and $T > 0$. S is a finite state space, S is also a compact set. We construct a compact set $K' \times S$ by Tychonoff's theorem [Eng89, Theorem 3.2.4] such that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[(X_n(t), \Lambda_n(t)) \notin K' \times S \text{ for some } t \leq T\right] \leq -a.$$

As $(X_n(0), \Lambda_n(0)$ is exponentially tight with speed n, we can find compact K_0 so that

$$\limsup_{n\to\infty} \frac{1}{n}\log\mathbb{P}\left[(X_n(0), \Lambda_n(0)) \notin K_0 \times S\right] \leq -a,$$

Then, in virtue of the convergence of the operator, we can find $(f_n, g_n) \in H_n$, a compact $K \times S$ and an open set $G \times S$, and define

$$\beta(K, G, S) := \liminf_{n\to\infty} \left(\inf_{(x,i)\in G^c\times S} f_n(x,i) - \sup_{(x,i)\in K\times S} f_n(x,i) \right)$$

and

$$\gamma(G, S) = \limsup_{n\to\infty} \sup_{x\in G\times S} g_n(x,i)$$

such that $\beta(K, G, S) + T\gamma(G, S) \leq -a$.

Set $\gamma := \sup_{(x,i)\in E\times S} H((x,i), \partial_x\Upsilon(x))$ and $c_1 := \sup_{(x,i)\in K_0\times S} \Upsilon(x)$. Observe that $\gamma < \infty$ by (3.22) and $c_1 < \infty$ by compactness. Now choose c_2 such that

$$-(c_2 - c_1) + T\gamma = -a \tag{3.24}$$

and take $K = \{(x,i) \in E \times S \mid \Upsilon(x) \leq c_1\}$ and $G = \{(x,i) \in E \times S \mid \Upsilon(x) < c_2\}$.

Let $\theta : [0,\infty) \to [0,\infty)$ be a compactly supported smooth function with the property that $\theta(x) = x$ for $x \leq c_2$. For each n, define $f_n := \theta \circ \Upsilon$ and $g_n := H_n f_n$. By the convergence of operator, $g_n \to Hf$ and moreover, by construction $\beta(K, G, S) = c_2 - c_1$ and $\gamma(G, S) = \gamma$. Thus by (3.24) and [FK06, Lemma 4.22] we obtain

$$\limsup_{n\to\infty} \frac{1}{n}\log\mathbb{P}[(X_n(t), \Lambda_n(t)) \notin G \times S \text{ for some } t \leq T] \leq -a$$

and the compact containment condition holds with $K_{a,T} = \overline{G}$. □

3.4 Comparison principle

In this section, we prove the comparison principle for the Hamilton-Jacobi equation $f - \lambda Hf = h$ with the help of the single valued operator $\mathbf{H}f(x) := \mathcal{H}(x, \partial_x f(x))$ in (3.19) as defined.

We argue by first encoding the containment function Υ into the domain of our operators. This allows us to work with optimizes as in Remark 2.5, and the strategy is summarized in Figure 3.1 below. We begin with the definition of the operators $H_1, H_2, H_\dagger$, and $H_\ddagger$ using $\Upsilon(x) = -\log(x) + \log\left(1 + \frac{1}{2}x^2\right) - \log\sqrt{2}$ with $C_\Upsilon = \sup_{x,i} B_{x,\partial_x\Upsilon(x)}(x,i) < \infty$. Denote by $C_l^\infty(E)$ the set of smooth functions on E that have a lower bound and by $C_u^\infty(E)$ the set of smooth functions on E that have an upper bound.

Definition 3.10. (Single valued operators)

- For $f \in C_l^\infty(E)$ and $\delta \in (0,1)$ set
$$f_\dagger^\delta := (1-\delta)f + \delta\Upsilon,$$
$$H_{\dagger,f}^\delta(x) := (1-\delta)\mathbf{H}f(x) + \delta C_\Upsilon,$$
and set
$$H_\dagger := \left\{\left(f_\dagger^\delta, H_{\dagger,f}^\delta\right) \middle| f \in C_l^\infty(E), \delta \in (0,1)\right\}.$$
- For $f \in C_u^\infty(E)$ and $\delta \in (0,1)$ set
$$f_\ddagger^\delta := (1+\delta)f - \delta\Upsilon,$$
$$H_{\ddagger,f}^\delta(x) := (1+\delta)\mathbf{H}f(x) - \delta C_\Upsilon,$$
and set
$$H_\ddagger := \left\{\left(f_\ddagger^\delta, H_{\ddagger,f}^\delta\right) \middle| f \in C_u^\infty(E), \delta \in (0,1)\right\}.$$

Definition 3.11. (Multi-valued operators)

- For $f \in C_l^\infty(E)$, $\delta \in (0,1)$ and $\phi \in C_b^2(E \times S)$. Set
$$f_1^\delta := (1-\delta)f + \delta\Upsilon,$$
$$H_{1,f,\phi}^\delta(x,i) := (1-\delta)H_{f,\phi}(x,i) + \delta C_\Upsilon,$$
and set
$$H_1 := \left\{\left(f_1^\delta, H_{1,f,\phi}^\delta\right) \middle| f \in C_l^\infty(E), \delta \in (0,1), \phi \in C_b^2(E \times S)\right\}.$$
- For $f \in C_u^\infty(E)$, $\delta \in (0,1)$ and $\phi \in C_b^2(E \times S)$. Set
$$f_2^\delta := (1+\delta)f - \delta\Upsilon,$$
$$H_{2,f,\phi}^\delta(x,i) := (1+\delta)H_{f,\phi}(x,i) - \delta C_\Upsilon,$$
and set
$$H_2 := \left\{\left(f_2^\delta, H_{2,f,\phi}^\delta\right) \middle| f \in C_u^\infty(E), \delta \in (0,1), \phi \in C_b^2(E \times S)\right\}.$$

Based on the above preparations, we are ready to state the most important proposition of this section.

Proposition 3.12 (Comparison principle)**.** *Let Assumptions 3.2 and 3.3 be satisfied. Let h_1, $h_2 \in C_b(E)$ and $\lambda > 0$. Let u be any subsolution to $f - \lambda Hf = h_1$ and let v be any supersolution to $f - \lambda Hf = h_2$. Then we have that*
$$\sup_x u(x) - v(x) \leq \sup_x h_1(x) - h_2(x).$$

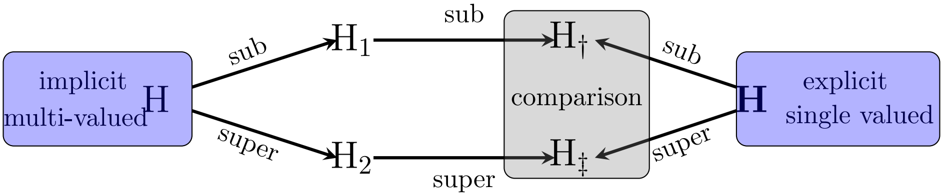

Figure 3.1: An arrow connecting an operator A with operator B with subscript 'sub' means that viscosity subsolutions of $f - \lambda A f = h$ are also viscosity subsolutions of $f - \lambda B f = h$. Similarly, we get the description for arrows with a subscript 'super'. The middle gray box around the operators $H_\dagger$ and $H_\ddagger$ indicates that the comparison principle holds for subsolutions of $f - \lambda H_\dagger f = h$ and supersolutions of $f - \lambda H_\ddagger f = h$. The left blue box indicates that H is an implicit and multi-valued operator. The right blue box indicates $\mathbf{H}$ is an explicit single value operator.

3.4.1 Strategy of proof of Proposition 3.12

The argument of Proposition 3.12 is inspired by the methods of [FK06, Chapter 11] and [KS21, Section 5] and is carried out by establishing the Figure 3.1. We first establish the two horizontal arrows in Figure 3.1.

Lemma 3.13. *Fix $\lambda > 0$ and $h \in C_b(E)$.*

(a) *Every subsolution to $f - \lambda H_1 f = h$ is also a subsolution to $f - \lambda H_\dagger f = h$.*

(b) *Every supersolution to $f - \lambda H_1 f = h$ is also a supersolution to $f - \lambda H_\ddagger f = h$.*

Proof. Let u be a subsolution to $f - \lambda H_1 f = h$. We show it is also a subsolution to $f - \lambda H_\dagger f = h$. To do it, for one thing, we find a unique optimizer x_0 in the compact level sets of Υ for the definition of viscosity solutions due to the existence of Υ. For another, we find a corrector using x_0.

Step 1: we show there exists x_0 such that

$$u(x_0) - f_1^\delta(x_0) = \sup_x u(x) - f_1^\delta(x). \tag{3.25}$$

First of all, note that u and $-f_1^\delta$ are upper semicontinuous. As Υ has compact sub-level sets, there exists x_0 such that

$$u(x_0) - f_1^\delta(x_0) = \sup_x u(x) - f_1^\delta(x).$$

Next, let $\hat{f} \in C_l^\infty(E)$ such that $\hat{f}(x_0) = f(x_0)$ and $\hat{f}(x) > f(x)$ if $x \neq x_0$ so that x_0 is the unique optimizer in

$$u(x_0) - f_1^\delta(x_0) = \sup_x u(x) - f_1^\delta(x)$$

and in addition $f'(x_0) = \hat{f}'(x_0)$.

Step 2: we consider the corrector. The corrector $\phi_{x_0} = \tau(x_0, p)$ existing by Proposition 3.5 is such that

$$H^{\delta}_{1,\hat{f},\phi_{x_0}}(x_0, i)$$

does not depend on i, and we have

$$H^{\delta}_{1,\hat{f},\phi_{x_0}}(x_0, i) = (1-\delta)\mathbf{H}f(x_0) + \delta C_{\Upsilon}.$$

due to Proposition 3.5. As u is a subsolution to $f - \lambda H_1 f = h$, there are (x_n, i_n) such that

$$\lim_n u(x_n) - \hat{f}^{\dagger}_{\delta}(x_n) = \sup_x u(x) - \hat{f}^{\dagger}_{\delta}(x)$$

and

$$\limsup_n u(x_n) - \lambda H^{\delta}_{1,\hat{f},\phi_{x_0}}(x_n, i_n) - h(x_n) \leq 0.$$

As x_0 is the unique optimizer of $\sup(u - \hat{f}^{\dagger}_{\delta})$, and as S is compact, there exists $i_0 \in S$ such that along a subsequence we have $(x_n, i_n) \to (x_0, i_0)$ as $n \to \infty$. We conclude that

$$u(x_0) - \lambda H^{\delta}_{\dagger,f}(x_0) - h(x_0) = u(x_0) - \lambda H^{\delta}_{1,\hat{f},\phi_{x_0}}(x_0, i) - h(x_0) \leq 0.$$

So in combination with (3.25), we have obtained the two desired properties for each pair of functions in $H_{\dagger}$. We conclude that u is a subsolution to $f - \lambda H_{\dagger} f = h$. $\square$

Lemma 3.14. *Fix $\lambda > 0$ and $h \in C_b(E)$.*

(a) *Every subsolution to $f - \lambda \mathbf{H} f = h$ is also a subsolution to $f - \lambda H_{\dagger} f = h$.*

(b) *Every supersolution to $f - \lambda \mathbf{H} f = h$ is also a supersolution to $f - \lambda H_{\ddagger} f = h$.*

Proof. Fix $\lambda > 0$ and $h \in C_b(E)$. Let u be a subsolution to $f - \lambda \mathbf{H} f = h$. We prove it is also a subsolution to $f - \lambda H_{\dagger} f = h$.

Fix $\delta > 0$, $f \in C_l^{\infty}(E)$ such that $(f^{\delta}_{\dagger}, H^{\delta}_{\dagger,f}) \in H_{\dagger}$. We will prove that there is a sequence $x_n \in E$ such that

$$\lim_{n\to\infty} u(x_n) - f^{\delta}_{\dagger}(x_n) = \sup_{x \in E} u(x) - f^{\delta}_{\dagger}(x), \tag{3.26}$$

$$\limsup_{n\to\infty} u(x_n) - \lambda H^{\delta}_{\dagger,f}(x_n) - h(x_n) \leq 0. \tag{3.27}$$

As the function $[u - (1-\delta)f]$ is bounded from above and Υ has compact sublevel sets, the sequence x_n along which the first limit is attained can be assumed to lie in the compact set

$$K := \left\{ x \mid \Upsilon(x) \leq \delta^{-1} \sup_x (u(x) - (1-\delta)f(x)) \right\}.$$

Set $M = \delta^{-1}\sup_x(u(x) - (1-\delta)f(x))$. Let $\gamma : \mathbb{R} \to \mathbb{R}$ be a smooth increasing function such that

$$\gamma(r) = \begin{cases} r, & \text{if } r \leq M, \\ M+1, & \text{if } r \geq M+2. \end{cases}$$

Let f_δ be a function on E defined by

$$f_\delta(x) := \gamma((1-\delta)f(x) + \delta\Upsilon(x)) = \gamma(f_\dagger^\delta(x)).$$

By construction, f_δ is smooth and constant outside of a compact set and thus lies in $\mathcal{D}(H) = C_{cc}^\infty(E)$. As u is a viscosity subsolution for $f - \lambda Hf = h$ there exists a sequence $x_n \in K \subseteq E$ (by our choice of K) with

$$\lim_n u(x_n) - f_\delta(x_n) = \sup_{x\in E} u(x) - f_\delta(x), \tag{3.28}$$

$$\limsup_n u(x_n) - \lambda \mathbf{H} f_\delta(x_n) - h(x_n) \leq 0. \tag{3.29}$$

As f_δ equals $f_\dagger^\delta$ on K, we have from (3.28) that also

$$\lim_n u(x_n) - f_\dagger^\delta(x_n) = \sup_{x\in E} u(x) - f_\dagger^\delta(x),$$

establishing (3.26). Convexity of $p \to \mathcal{H}(x,p)$ yields for arbitrary points $x \in K$ the estimate

$$\begin{aligned} \mathbf{H} f_\delta(x) &= \mathcal{H}(x, \partial_x f_\delta(x)) \\ &\leq (1-\delta)\mathcal{H}(x, \partial_x f(x)) + \delta\mathcal{H}(x, \partial_x \Upsilon(x)) \\ &\leq (1-\delta)\mathcal{H}(x, \partial_x f(x)) + \delta C_\Upsilon = H_{\dagger,f}^\delta(x). \end{aligned}$$

Combining this inequality with (3.29) yields

$$\begin{aligned} \limsup_n & u(x_n) - \lambda H_{\dagger,f}^\delta(x_n) - h(x_n) \\ &\leq \limsup_n u(x_n) - \lambda \mathbf{H} f_\delta(x_n) - h(x_n) \leq 0, \end{aligned}$$

establishing (3.27). This concludes the proof. $\square$

Lemma 3.15. *Fix $\lambda > 0$ and $h \in C_b(E)$.*

(a) *Every subsolution to $f - \lambda Hf = h$ is also a subsolution to $f - \lambda H_1 f = h$.*

(b) *Every supersolution to $f - \lambda Hf = h$ is also a supersolution to $f - \lambda H_2 f = h$.*

Proof. This proof has the same idea as Lemma 3.14, but we need to make appropriate modifications. To maintain integrity and readability, we give its proof in the following.

Fix $\lambda > 0$ and $h \in C_b(E)$. Let u be a subsolution to $f - \lambda Hf = h$. We prove it is also a subsolution to $f - \lambda H_1 f = h$. Fix $\delta \in (0,1)$, $\phi \in C_b^2(E \times S)$ and $f \in C_l^\infty(E)$, such that $(f_1^\delta, H_{1,f,\phi}^\delta) \in H_1$. We will prove that there are (x_n, i_n) such that

$$\lim_n u(x_n) - f_1^\delta(x_n) = \sup_x u(x) - f_1^\delta(x) \tag{3.30}$$

$$\limsup_n u(x_n) - \lambda H_{1,f,\phi}^\delta(x_n, i_n) - h(x_n) \leq 0. \tag{3.31}$$

We have that $M := \delta^{-1} \sup_x (u(x) - (1-\delta) f(x)) < \infty$ as u is bounded and $f \in C_l(E)$. It follows that the sequence x_n along which the limit in (3.30) is attained is contained in the compact set $K := \{x \mid \Upsilon(x) \leq M\}$.

Let $\gamma : \mathbb{R} \to \mathbb{R}$ be a smooth increasing function such that

$$\gamma(r) = \begin{cases} r, & \text{if } r \leq M, \\ M+1, & \text{if } r \geq M+2. \end{cases}$$

Denote by f_δ the function on E defined by

$$f_\delta(x) := \gamma((1-\delta) f(x) + \delta \Upsilon(x)) = \gamma(f_1^\delta(x)).$$

By construction, f_δ is smooth and constant outside of a compact set and thus lies in $\mathcal{D}(H) = C_{cc}^\infty(E)$. As $\mathrm{e}^\phi \in C_b^2(E \times S)$, we also have $\mathrm{e}^{(1-\delta)\phi} \in C^2(E \times S)$. We conclude that $(f_\delta, H_{f_\delta,(1-\delta)\phi}) \in H$. As u is a viscosity subsolution for $f - \lambda Hf = h$, there exist $x_n \in K \subseteq E$ (by our choice of K) and $i_n \in S$ with

$$\lim_n u(x_n) - f_\delta(x_n) = \sup_x u(x) - f_\delta(x), \tag{3.32}$$

$$\limsup_n u(x_n) - \lambda H_{f_\delta,(1-\delta)\phi}(x_n, i_n) - h(x_n) \leq 0. \tag{3.33}$$

As f_δ equals f_1^δ on K, we have from (3.32) that also

$$\lim_n u(x_n) - f_1^\delta(x_n) = \sup_x u(x) - f_1^\delta(x),$$

establishing (3.30). For arbitrary sequences (x_n, i_n) the elementary estimate

$$\begin{aligned}
&H_{f_\delta,(1-\delta)\phi}(x_n, i_n) \\
&\quad = B_{x,\partial_x f_\delta(x)}(x_n, i_n) + \mathrm{e}^{-(1-\delta)\phi(x_n,i_n)} R_x \mathrm{e}^{(1-\delta)\phi(x_n,i_n)} \\
&\quad \leq (1-\delta) B_{x,\partial_x f(x)}(x_n, i_n) + \delta B_{x,\partial_x \Upsilon(x)}(x_n, i_n) + (1-\delta) \mathrm{e}^{-\phi(x_n,i_n)} R_x \mathrm{e}^{\phi(x_n,i_n)} \\
&\quad = (1-\delta) \left(B_{x,\partial_x f(x)}(x_n, i_n) + \mathrm{e}^{-\phi(x_n,i_n)} R_x \mathrm{e}^{\phi(x_n,i_n)} \right) + \delta B_{x,\partial_x \Upsilon(x)}(x_n, i_n) \\
&\quad \leq H_{1,f,\phi}^\delta(x_n, i_n),
\end{aligned}$$

In the first inequality, we use that $B_{x,p}$ is convex concerning p. In virtue of (3.23), the last inequality holds. Combining above inequality with (3.33) yields

$$\begin{aligned}\limsup_n \big[u(x_n) - \lambda H^{\delta}_{1,f,\phi}(x_n, i_n) - h(x_n)\big] \\ \leq \limsup_n \left[u(x_n) - \lambda H_{f_\delta,\phi}(x_n, i_n) - h(x_n)\right] \leq 0,\end{aligned}$$

establishing (3.31). This concludes the proof. □

The following lemma is to verify the comparison principle for Hamilton-Jacobi-Bellman equations involving the Hamiltonian $H_\dagger$ and $H_\ddagger$.

Lemma 3.16. *Assumptions 3.2 and 3.3 hold. Let h_1, $h_2 \in C_b(E)$ and $\lambda > 0$. Let u be any subsolution to $f - \lambda H_\dagger f = h_1$ and let v be any supersolution to $f - \lambda H_\ddagger f = h_2$. Then we have*

$$\sup_x u(x) - v(x) \leq \sup_x h_1(x) - h_2(x).$$

A key step in the proof is the doubling of variables procedure as e.g. explained in [CIlL92]. We first give the the definition of penalization function for introducing the auxiliary lemma below that is often used in proving Lemma 3.16.

Definition 3.17 (Penalization function)**.** We say that $d : E \times E \to [0, \infty)$ is a penalization function if $d \in C(E \times E)$ and if $x = y$ if and only if $d(x, y) = 0$.

Lemma 3.18 (Lemma A.10 in [CK17])**.** *Let u be bounded and upper semicontinuous, let v be bounded and lower semicontinuous, let the distance function $d : E \times E \to \mathbb{R}^+$ be good penalization function and let Υ be a good containment function.*

Fix $\delta > 0$. For every $m > 0$ there exist points $x_{\delta,m}$, $y_{\delta,m} \in E$, such that

$$\begin{aligned}&\frac{u(x_{\delta,m})}{1-\delta} - \frac{v(y_{\delta,m})}{1+\delta} - md^2(x_{\delta,m}, y_{\delta,m}) - \frac{\delta}{1-\delta}\Upsilon(x_{\delta,m}) - \frac{\delta}{1+\delta}\Upsilon(y_{\delta,m}) \\ &= \sup_{x,y\in E}\Big\{\frac{u(x)}{1-\delta} - \frac{v(y)}{1+\delta} - md^2(x, y) - \frac{\delta}{1-\delta}\Upsilon(x) - \frac{\delta}{1+\delta}\Upsilon(y)\Big\}.\end{aligned}$$

Additionally, for every $\delta > 0$ we have that

(a) *The set $\{x_{\delta,m}, y_{\delta,m} \mid m > 0\}$ is relatively compact in E.*

(b) *All limit points of $\{(x_{\delta,m}, y_{\delta,m})\}_{m>0}$ are of the form (z, z) and for these limit points we have*

$$u(z) - v(z) = \sup_{x\in E} u(x) - v(x).$$

(c) *We have*

$$\lim_{m\to\infty} md^2(x_{\delta,m}, y_{\delta,m}) = 0.$$

Remark 3.19. For the good penalization function d in Lemma 3.18, we take

$$d^2(x,y) = \frac{2}{\theta^2}(\sqrt{x} - \sqrt{y})^2 \tag{3.34}$$

based on [DFL11, Section 2.2] in the proof of Lemma 3.16.

Here, we give the proof of Lemma 3.16.

Proof of Lemma 3.16. For $0 < \delta < 1$ and $m > 1$, let

$$\Phi_{\delta,m}(x,y) := \frac{u(x)}{1-\delta} - \frac{v(y)}{1+\delta} - md^2(x,y) - \frac{\delta}{1-\delta}\Upsilon(x) - \frac{\delta}{1+\delta}\Upsilon(y),$$

where $d(\cdot,\cdot)$ is given in (3.34) and $\Upsilon(\cdot)$ is given in (3.21). Since $\Upsilon(\cdot)$ has compact level sets, there exists $(x_{\delta,m}, y_{\delta,m}) \in E \times E$ satisfying

$$\Phi_{\delta,m}(x_{\delta,m}, y_{\delta,m}) = \sup_{(x,y)\in E\times E} \Phi_{\delta,m}(x,y). \tag{3.35}$$

Let $\varphi_1^{\delta,m} \in \mathcal{D}(H_\dagger)$ be defined as

$$\begin{aligned}\varphi_1^{\delta,m}(x) :&= (1-\delta)\big(\frac{v(y_{\delta,m})}{1+\delta} + md^2(x, y_{\delta,m}) + \frac{\delta}{1-\delta}\Upsilon(x) + \frac{\delta}{1+\delta}\Upsilon(y_{\delta,m})\big) \\ &\quad + (1-\delta)(x - x_{\delta,m})^2,\end{aligned}$$

where adding $(1-\delta)(x - x_{\delta,m})^2$ in $\varphi_1^{\delta,m}(x)$ implies that $u - \varphi_1^{\delta,m}$ attains its a unique supremum at $x = x_{\delta,m}$, namely

$$\sup_{x\in E} u(x) - \varphi_1^{\delta,m}(x) = u(x_{\delta,m}) - \varphi_1^{\delta,m}(x_{\delta,m}).$$

By the viscosity subsolution property of u one has

$$u(x_{\delta,m}) - \lambda\left[(1-\delta)\mathcal{H}(x_{\delta,m}, p^1_{\delta,m}) + \delta C_\Upsilon\right] \le h_1(x_{\delta,m}), \tag{3.36}$$

where

$$P^1_{\delta,m} := m\partial_x d^2(x_{\delta,m}, y_{\delta,m}) = \frac{2m}{\theta^2}\big(1 - \frac{\sqrt{y_{\delta,m}}}{\sqrt{x_{\delta,m}}}\big). \tag{3.37}$$

Similarly, let $\varphi_2^{\delta,m} \in \mathcal{D}(H_\ddagger)$ be defined as

$$\begin{aligned}\varphi_2^{\delta,m}(y) :&= (1+\delta)\big(\frac{u(x_{\delta,m})}{1-\delta} - md^2(x, y_{\delta,m}) - \frac{\delta}{1-\delta}\Upsilon(x_{\delta,m}) - \frac{\delta}{1+\delta}\Upsilon(y)\big) \\ &\quad - (1+\delta)(y - y_{\delta,m})^2.\end{aligned}$$

Therefore, we obtain the supersolution inequality

$$v(x_{\delta,m}) - \lambda\left[(1+\delta)\mathcal{H}(y_{\delta,m}, p^2_{\delta,m}) - \delta C_\Upsilon\right] \ge h_2(y_{\delta,m}), \tag{3.38}$$

where

$$p^2_{\delta,m} := -m\partial_y d^2(x_{\delta,m}, y_{\delta,m}) = -\frac{2m}{\theta^2}\big(1 - \frac{\sqrt{x_{\delta,m}}}{\sqrt{y_{\delta,m}}}\big). \tag{3.39}$$

By Lemma 3.18, we have

$$\lim_{m\to\infty} md^2(x_{\delta,m}, y_{\delta,m}) = 0. \tag{3.40}$$

Combining (3.36), (3.38) and (3.40) we get

$$\begin{aligned}\sup_{x\in E} u(x) - v(x) &\le \liminf_{\delta\to 0}\liminf_{m\to\infty}\big(\frac{u(x_{\delta,m})}{1-\delta} - \frac{v(y_{\delta,m})}{1+\delta}\big)\\ &\le \liminf_{\delta\to 0}\liminf_{m\to\infty}\big\{\frac{h_1(x_{\delta,m})}{1-\delta} - \frac{h_2(y_{\delta,m})}{1+\delta} &(3.41)\\ &\quad + \frac{\delta}{1-\delta}C_\Upsilon + \frac{\delta}{1+\delta}C_\Upsilon &(3.42)\\ &\quad + \lambda\big(\mathcal{H}(x_{\delta,m}, p^1_{\delta,m}) - \mathcal{H}(y_{\delta,m}, p^2_{\delta,m})\big)\big\}, &(3.43)\end{aligned}$$

where in the first inequality we use (3.35) and drop the non-negative functions $d^2(\cdot,\cdot)$ and $\Upsilon(\cdot)$. The term in (3.42) vanishes as C_Υ in (3.23) is a constant.

Based on Lemma 3.18, for fixed δ and varying m, the sequence $(x_{\delta,m}, y_{\delta,m})$ takes its values in a compact set and, hence, admits converging subsequences. By (b) of Lemma 3.18, these subsequences converge to points of the form (x, x). Therefore, we can deal with (3.41). Then, by the above analysis, we can get

$$\begin{aligned}\sup_{x\in E} u(x) - v(x) \le{}& \lambda \liminf_{\delta\to 0}\liminf_{m\to\infty}\big(\mathcal{H}(x_{\delta,m}, p^1_{\delta,m}) - \mathcal{H}(y_{\delta,m}, p^2_{\delta,m})\big)\\ &+ \sup_{x\in E} h_1(x) - h_2(x).\end{aligned}$$

It follows that the comparison principle holds for $f - \lambda H_\dagger f = h_1$ and $f - \lambda H_\ddagger f = h_2$ whenever for any $\delta > 0$

$$\liminf_{m\to\infty}\big(\mathcal{H}(x_{\delta,m}, p^1_{\delta,m}) - \mathcal{H}(y_{\delta,m}, p^2_{\delta,m})\big) \le 0. \tag{3.44}$$

To that end, recall $\mathcal{H}(x,p)$ in (3.19):

$$\mathcal{H}(x,p) = \sup_{\pi\in\mathcal{P}(S)}\left\{\int B_{x,p}(z)\pi(\mathrm{d}z) - \mathcal{I}(x,\pi)\right\},$$

where $\pi \mapsto \int B_{x_{\delta,m},p^i_{\delta,m}}(z)\pi(\mathrm{d}z)$ is bounded and continuous and $\mathcal{I}(x_{\delta,m},\cdot)$ has compact sub-level sets in $\mathcal{P}(S)$. Thus, there exists an optimizer $\pi_{\delta,m} \in \mathcal{P}(S)$ such that

$$\mathcal{H}(x_{\delta,m}, p^1_{\delta,m}) = \int B_{x_{\delta,m},p^1_{\delta,m}}(z)\pi_{\delta,m}(\mathrm{d}z) - \mathcal{I}(x_{\delta,m}, \pi_{\delta,m}) \tag{3.45}$$

and

$$\mathcal{H}(y_{\delta,m}, p^2_{\delta,m}) \geq \int B_{x_{\delta,m}, p^2_{\delta,m}}(z)\pi_{\delta,m}(\mathrm{d}z) - \mathcal{I}(x_{\delta,m}, \pi_{\delta,m}). \tag{3.46}$$

Combining (3.45) and (3.46), we obtain

$$\begin{aligned} &\mathcal{H}(x_{\delta,m}, p^1_{\delta,m}) - \mathcal{H}(y_{\delta,m}, p^2_{\delta,m}) \\ &\quad \leq \int \big(B_{x_{\delta,m}, p^1_{\delta,m}}(z) - B_{y_{\delta,m}, p^2_{\delta,m}}(z)\big)\pi_{\delta,m}(\mathrm{d}z) && (3.47) \\ &\qquad + \mathcal{I}(y_{\delta,m}, \pi_{\delta,m}) - \mathcal{I}(x_{\delta,m}, \pi_{\delta,m}). && (3.48) \end{aligned}$$

It is sufficient to prove that (3.47) and (3.48) are sufficiently small. For (3.47), by calculating the difference of integrand $B_{x,p}$ in detail, for any $z \in E$ and $i \in S$, one has

$$\begin{aligned} &B_{x_{\delta,m}, p^1_{\delta,m}}(z) - B_{y_{\delta,m}, p^2_{\delta,m}}(z) \\ &= \big(\eta(\mu(z) - x_{\delta,m})p^1_{\delta,m} + \frac{1}{2}\theta^2 x_{\delta,m}(p^1_{\delta,m})^2\big) - \big(\eta(\mu(z) - y_{\delta,m})p^2_{\delta,m} + \frac{1}{2}\theta^2 y_{\delta,m}(p^2_{\delta,m})^2\big) \\ &= \big(\eta(\mu(z) - x_{\delta,m})p^1_{\delta,m} - \eta(\mu(z) - y_{\delta,m})p^2_{\delta,m}\big) + \big(\frac{1}{2}\theta^2 x_{\delta,m}(p^1_{\delta,m})^2 - \frac{1}{2}\theta^2 y_{\delta,m}(p^2_{\delta,m})^2\big) \\ &= \frac{2m\eta}{\theta^2}(\mu(z) - y_{\delta,m})\big(1 - \frac{\sqrt{x_{\delta,m}}}{\sqrt{y_{\delta,m}}}\big) + \frac{2m\eta}{\theta^2}(\mu(z) - x_{\delta,m})\big(1 - \frac{\sqrt{y_{\delta,m}}}{\sqrt{x_{\delta,m}}}\big) \\ &= \frac{2m\eta}{\theta^2}\big[(\mu(z) - y_{\delta,m})\big(1 - \frac{\sqrt{x_{\delta,m}}}{\sqrt{y_{\delta,m}}}\big) + (\mu(z) - x_{\delta,m})\big(1 - \frac{\sqrt{y_{\delta,m}}}{\sqrt{x_{\delta,m}}}\big)\big] \\ &= -m\eta\mu(z)\frac{d^2(x_{\delta,m}, y_{\delta,m})}{\sqrt{x_{\delta,m}y_{\delta,m}}} - m\eta d^2(x_{\delta,m}, y_{\delta,m}) \leq 0, \end{aligned}$$

where in the third equality we use (3.37) and (3.39). For (3.48), we utilize the equicontinuity of $\mathcal{I}(\cdot, \pi)$ established in Lemma 3.20 below for the spatial variable. We are left with stating and verifying Lemma 3.20. This finishes the proof of (3.44) and the comparison principle for $H_\dagger$ and $H_\ddagger$. □

Lemma 3.20. *Let Assumption 3.3 be satisfied. Recall* (3.20)*:*

$$\mathcal{I}(x, \pi) = -\inf_{g>0} \int \frac{R_x g(z)}{g(z)}\pi(\mathrm{d}z).$$

For any compact set $G \subseteq E$, then the collection $\{x \mapsto J(x, \pi) \mid \pi \in \mathcal{P}(S)\}$ is equicontinuity.

Proof. Let ρ be some metric on the topology of E. We will prove that for any compact sets $G \subseteq E$ and $\varepsilon > 0$, there is some $\delta > 0$ such that for all $x, y \in G$ with $\rho(x, y) \leq \delta$ and for all $\pi \in \mathcal{P}(S)$, we have

$$|\mathcal{I}(x, \pi) - \mathcal{I}(y, \pi)| \leq \varepsilon. \tag{3.49}$$

In virtue of the definition of $\mathcal{I}$, there exists a function $\phi \in C(S)$ independent of x such that e^{ϕ} in the domain of R_x, and

$$\begin{aligned}\mathcal{I}(x,\pi) &= -\inf_{g>0}\int \frac{R_x g(z)}{g(z)}\pi(\mathrm{d}z)\\ &= \sup_{\phi\in C(S)}\sum_{i,j\in S} q_{ij}(x)\pi_i\big(1-\mathrm{e}^{\phi(j)-\phi(i)}\big).\end{aligned}$$

Let $x,y \in G$. By continuity of the transition rate $q_{ij}(x)$, the $\mathcal{I}(x,\cdot)$ are uniformly bounded for $x\in G$:

$$0\le \mathcal{I}(x,\pi)\le \sum_{i,j,i\neq j} q_{ij}(x)\pi_i \le \sum_{i,j,i\neq j} q_{ij}(x)\le \sum_{i,j,i\neq j}\overline{q}_{ij},\ \overline{q}_{ij} := \sup_{x\in G} q_{ij}(x).$$

For any $n\in N$, there exists $\phi^n \in C(S)$ such that

$$0\le \mathcal{I}(x,\pi)\le \sum_{i,j,i\neq j} q_{ij}(x)\pi_i(1-\mathrm{e}^{\phi^n(j)-\phi^n(i)})+\frac{1}{n}.$$

By reorganizing, we find for all pairs (k,l) the bound

$$\pi_k \mathrm{e}^{\phi^n(l)-\phi^n(k)} \le \frac{1}{r_G(k,l)}\Big(\sum_{i,j,i\neq j} q_{ij}(x)\pi_i+\frac{1}{n}\Big)\le \frac{1}{r_G(k,l)}\Big(\sum_{i,j,i\neq j}\overline{q}_{ij}+\frac{1}{n}\Big),$$

where $r_G(k,l) := \inf_{x\in G} q_{kl}(x)$. Thereby, evaluating in $\mathcal{I}(y,\pi)$ the same function ϕ^n to estimate the supremum,

$$\begin{aligned}&\mathcal{I}(x,\pi)-\mathcal{I}(y,\pi)\\ &\le \frac{1}{n}+\sum_{k,l,k\neq l} q_{kl}(x)\pi_k\big(1-\mathrm{e}^{\phi^n(l)}-\mathrm{e}^{\phi^n(k)}\big)-\sum_{k,l,k\neq l} q_{kl}(y)\pi_k\big(1-\mathrm{e}^{\phi^n(l)}-\mathrm{e}^{\phi^n(k)}\big)\\ &\le \frac{1}{n}+\sum_{k,l,k\neq l}|q_{kl}(x)-q_{kl}(y)|\pi_k+\sum_{k,l,k\neq l}|q_{kl}(y)-q_{kl}(x)|\,\pi_k\mathrm{e}^{\phi^n(l)-\phi^n(k)}\\ &\le \frac{1}{n}+\sum_{k,l,k\neq l}|q_{kl}(x)-q_{kl}(y)|\Big[1+\frac{1}{r_G(k,l)}\Big(\sum_{k,l,k\neq l}\overline{q}_{kl}+1\Big)\Big].\end{aligned}$$

We take $n\to\infty$ and use that the rates $x\mapsto q_{kl}(x)$ are Lipschitz continuous under Assumption 3.3, and hence uniformly continuous on compact sets, to obtain (3.49). Hence, the proof of the lemma is concluded. □

3.4.2 Proof of Proposition 3.12

We now prove Proposition 3.12; that is, the verification of the comparison principle for the Hamilton-Jacobi equations $f-\lambda Hf = h$. The proof follows the strategy of Figure 3.1 combined with Lemma 3.13, Lemma 3.14, Lemma 3.15 and Lemma 3.16. We thus obtain Figure 3.2 via adding these lemmas in Figure 3.1 as below for an easy understanding of the proof strategy of Proposition 3.12.

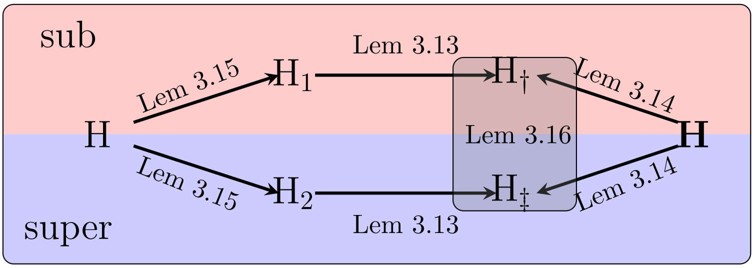

Figure 3.2: Add lemmas in Figure 3.1

Proof of Proposition 3.12. Fix $h_1,\ h_2 \in C_b(E)$ and $\lambda > 0$. Let u be a viscosity subsolution to $(1-\lambda H)f = h_1$ and v be a viscosity supersolution to $(1-\lambda H)f = h_2$. By Lemma 3.15 and Lemma 3.13, the function u is a viscosity subsolution to $(1-\lambda H_\dagger)f = h_1$ (see red part on Figure 3.2) and v is a viscosity supersolution to $(1-\lambda H_\ddagger)f = h_2$ (see blue part on Figure 3.2). Hence by the comparison principle for $H_\dagger$, $H_\ddagger$ established in Lemma 3.16, we get $\sup_x u(x) - v(x) \leq \sup_x h_1(x) - h_2(x)$. This finished the proof. □

3.5 Proof of action-integral representation of the rate function

In this section, we will prove our main result Theorem 3.3. According to the strategy in Section 2.2, the proof is based on three main parts that we have proven:

- operator convergence;
- exponential tightness;
- comparison principle.

Hence, $X_n^\varepsilon(t)$ satisfies large deviation principles with projective limit form rate function (2.14).

We are left to prove that (2.14) has the action-integral form rate function (2.15), and put it in the proof of Theorem 3.3 below. To achieve the aim, according to step 4 in Section 2.2.1, we should first prove the lemma below which is necessary to obtain (2.15).

Lemma 3.21. *Let $\mathcal{H} : E \times E \to \mathbb{R}$ be the map given in* (3.19) *and the operator* $\mathbf{H}f(x) := \mathcal{H}(x, \partial_x f(x))$*. Then, the operator* $\mathbf{H}$ *satisfies Conditions 8.9, 8.10, and 8.11 of [FK06].*

Proof. We first show that the following conditions imply Conditions 8.9, 8.10, and 8.11 in a non-compact setting. These ideas come from the proof of Proposition 6.1-(i) in [PS24]. We begin with modifying the conditions adapted to our setting.

(a) The function $\mathcal{L} : E \times E \to [0, \infty]$ is lower semicontinuous and for every $C \geq 0$, the level set $\{(x, v) \in E \times E \mid \mathcal{L}(x, v) \leq C\}$ is relatively compact in $E \times E$.

(b) For all $f \in \mathcal{D}(\mathbf{H})$ there exists a right continuous, non-decreasing function $\psi_f : [0, \infty) \to [0, \infty)$ such that for all $(x, v) \in E \times E$,

$$|\partial_x f(x) \cdot v| \leq \psi_f(\mathcal{L}(x, v)) \quad \textit{and} \quad \lim_{r \to \infty} r^{-1} \psi_f(r) = 0.$$

(c) For each $x_0 \in E$ and every $f \in \mathcal{D}(\mathbf{H})$, there exists an absolutely continuous path $x : [0, \infty) \mapsto E$ such that

$$\int_0^t \mathcal{H}(x(s), \partial_x f(x(s)))\mathrm{d}s = \int_0^t \partial_x f(x(s)) \cdot \dot{x}(s) - \mathcal{L}(x(s), \dot{x}(s))]\mathrm{d}s.$$

Then, we will use (1), (2), and (3) to prove Condition 8.9, 8.10, and 8.11. Regarding Condition 8.9, the operator $Af(x, v) := \partial_x f(x) \cdot v$ on the domain $\mathcal{D}(A) = \mathcal{D}(H)$ satisfies Condition 8.9.1 in [FK06]. For Condition 8.9.2 in [FK06], we can choose $\Gamma = E \times E$, and for $x_0 \in E$, take the pair (x, λ) with $x(t) = x_0$ and $\lambda(\mathrm{d}v \times \mathrm{d}t) = \delta_0(\mathrm{d}v) \times \mathrm{d}t$. Condition 8.9.3 in [FK06] is a consequence of Condition 8.9.1 in [FK06] from above. Condition 8.9.4 in [FK06] can be verified as follows. Let be Υ the containment function used in (3.21) and note that the sub-level sets of Υ are compact. Let $\gamma \in \mathcal{AC}(E)$ with $\gamma(0) \in K$ and such that the control

$$\int_0^T \mathcal{L}(\gamma(s)), \dot{\gamma}(s)) \leq M$$

implies $\gamma(t) \in \hat{K}$ for all $t \leq T$, where $\hat{K}$ is a compact set. Then,

$$\begin{aligned}
\Upsilon(\gamma(t)) &= \Upsilon(\gamma(0)) + \int_0^t \langle \partial_x \Upsilon(\gamma(s)), \dot{\gamma}(s) \rangle \mathrm{d}s \\
&\leq \Upsilon(\gamma(0)) + \int_0^t \mathcal{L}(\gamma(s)), \dot{\gamma}(s)) + \mathcal{H}(\gamma(s), \partial_x \Upsilon(\gamma(s)))\mathrm{d}s \\
&\leq \sup_{y \in K} \Upsilon(y) + M + \int_0^T \sup_z \mathcal{H}(z, \partial_x \Upsilon(z))\mathrm{d}s \\
&:= C < \infty.
\end{aligned}$$

Hence, we can take $\hat{K} = \{z \in E \mid \Upsilon(z) \leq C\}$. Condition 8.9.5 in [FK06] is implied by Condition 8.9.2 in [FK06] from above.

Condition 8.10 [FK06] is implied by Condition 8.11 [FK06] with the fact that $\mathbf{H}1 = 0$ [FK06, Remark 8.12-(e)].

Finally, Condition 8.11 in [FK06] is implied by (3) above, with the control $\lambda(\mathrm{d}v \times \mathrm{d}t) = \delta_{\dot{x}(t)}(\mathrm{d}v) \times \mathrm{d}t$. □

In the rest of this section, we prove Theorem 3.3.

Proof of Theorem 3.3. From Proposition 2.14, we have proven LDP with a projective limit form rate function. Then we rewrite the rate-function on the Skorohod space in an action-integral form.

We first show that the Lagrangian $\mathcal{L}$ is superlinear, which means $(\mathcal{L}(x,v)/|v|) \to \infty$ as $|v| \to \infty$. To do it, for any $c > 0$ we have

$$\begin{aligned}\frac{\mathcal{L}(x,v)}{|v|} &= \sup_{p\in\mathbb{R}} \left[p \cdot \frac{v}{|v|} - \frac{\mathcal{H}(x,p)}{|v|}\right] \\ &\geq \sup_{|p|=c} \left[p \cdot \frac{v}{|v|} - \frac{\mathcal{H}(x,p)}{|v|}\right] \\ &\geq c - \frac{1}{|v|} \sup_{|p|=c} \mathcal{H}(x,p).\end{aligned}$$

The convex Hamiltonian is continuous, and therefore $\sup_{|p|=c} \mathcal{H}(x,p)$ is finite. Hence for arbitrary $c > 0$, we have $\mathcal{L}(x,v)/|v| > c/2$ for all $|v|$ large enough.

Let $x : [0,T] \to E$ be absolutely continuous and are two arbitrary $0 = t_0 < t_1$. We show that

$$I^V_{t_1-t_0}(x(t_1) \mid x(t_0)) = \inf_{\substack{\gamma(t_0)=x(t_0)\\ \gamma(t_1)=x(t_1)}} \int_0^t \mathcal{L}(\gamma(s),\dot{\gamma}(s))\mathrm{d}s, \tag{3.50}$$

where the infimum is taken over absolutely continuous paths $\gamma : [t_0,t_1] \to E$. Once we have this equality established, we obtain for arbitrary $k \in \mathbb{N}$ and points in time $0 = t_0, t_1, \ldots, t_k = T$ the estimate

$$\begin{aligned}&I^V_{t_1-t_0}(x(t_1) \mid x(t_0)) + I^V_{t_2-t_1}(x(t_2) \mid x(t_1)) + \cdots + I^V_{t_k-t_{k-1}}(x(t_k) \mid x(t_{k-1}) \\ &\quad \leq \int_0^T \mathcal{L}(x(s),\dot{x}(s))\mathrm{d}s,\end{aligned}$$

since $x(\cdot)$ satisfies the begin- and endpoint constraints. For the reverse inequality, we note that adding time points increases the two-point rate functions since we add a condition on the paths; for $t_0 < t_1 < t_2$,

$$\begin{aligned}&I^V_{t_2-t_0}(x(t_2) \mid x(t_0)) \\ &\quad = \inf_{\substack{\gamma(t_0)=x(t_0)\\ \gamma(t_2)=x(t_2)}} \left[\int_{t_0}^{t_1} \mathcal{L}(\gamma(s),\dot{\gamma}(s))\mathrm{d}t + \int_{t_1}^{t_2} \mathcal{L}(\gamma(s),\dot{\gamma}(s))\mathrm{d}t\right] \\ &\quad \leq \inf_{\substack{\gamma(t_0)=x(t_0)\\ \gamma(t_1)=x(t_1)}} \left[\int_{t_0}^{t_1} \mathcal{L}(\gamma(s),\dot{\gamma}(s))\mathrm{d}t\right] + \inf_{\substack{\gamma(t_1)=x(t_1)\\ \gamma(t_2)=x(t_2))}} \left[\int_{t_1}^{t_2} \mathcal{L}(\gamma(s),\dot{\gamma}(s))\mathrm{d}t\right] \\ &\quad = I^V_{t_2-t_1}(x(t_2) \mid x(t_1)) + I^V_{t_1-t_0}(x(t_1)|x(t_0)).\end{aligned}$$

The partitions of a time interval $[0,T]$ give rise to a monotonically increasing sequence. In the limit, we obtain

$$\sup_k \sup_{t_i} \sum_{i=0}^{k} I^V_{t_{i+1}-t_i}(x(t_{i+1}) \mid x(t_i)) = \int_0^T \mathcal{L}(x(s),\dot{x}(s))\mathrm{d}s.$$

We do not show that here but refer to [Vil09, Definition 7.11, Example 7.12]. We now show how (3.50) follows from the compact sub-level sets.

For $f \in C_b(E)$ and $x(t_0) \in E$, starting from

$$\begin{aligned} V(t_1)f(x(t_0)) &= \mathbf{V}(t_1)f(x(t_0)) \\ &= \sup_{\substack{\gamma(t_0)=x(t_0)\\ \gamma(t_1)=x(t_1)}} \left\{ f(\gamma(t_1)) - \int_{t_0}^{t_1} \mathcal{L}(\gamma(s),\dot{\gamma}(s))\mathrm{d}s \right\}. \\ &= -\inf_{\substack{\gamma(t_0)=x(t_0)\\ \gamma(t_1)=x(t_1)}} \left\{ \int_{t_0}^{t_1} \mathcal{L}(\gamma(s),\dot{\gamma}(s))\mathrm{d}s - f(\gamma(t_1)) \right\}. \end{aligned} \tag{3.51}$$

We have

$$\begin{aligned} &I^V_{t_1-t_0}(x(t_1) \mid x(t_0)) \\ &\quad= \sup_{f\in C_b(E)} \left(f(x(t_1)) - \mathbf{V}(t_1)f(x(t_0))\right) \\ &\quad\overset{(3.51)}{=} \sup_{f\in C_b(E)} \inf_{\substack{\gamma(t_0)=x(t_0)\\ \gamma(t_1)=x(t_1)}} \left[f(x(t_1)) - f(\gamma(t_1)) + \int_{t_0}^{t_1} \mathcal{L}(\gamma(s),\dot{\gamma}(s)\mathrm{d}s\right]. \end{aligned} \tag{3.52}$$

For any $f \in C_b(E)$,

$$\inf_{\gamma(t_0)=x(t_0)} \left[f(x(t_1)) - f(\gamma(t_1)) + \int_{t_0}^{t_1} \mathcal{L}(\gamma(s),\dot{\gamma}(s)\mathrm{d}s\right] \leq \inf_{\substack{\gamma(t_0)=x(t_0)\\ \gamma(t_1)=x(t_1)}} \int_{t_0}^{t_1} \mathcal{L}(\gamma(s),\dot{\gamma}(s)\mathrm{d}s,$$

since $\{\gamma : \gamma(t_0) = x(t_0)\}$ contains $\{\gamma : \gamma(t_0) = x(t_0), \gamma(t_1) = x(t_1)\}$. Taking the supremum over all f

$$\begin{aligned} &\sup_{f\in C_b(E)} \inf_{\gamma(t_0)=x(t_0)} \left[f(x(t_1)) - f(\gamma(t_1)) + \int_{t_0}^{t_1} \mathcal{L}(\gamma(s),\dot{\gamma}(s)\mathrm{d}s\right] \\ &\qquad\leq \inf_{\substack{\gamma(t_0)=x(t_0)\\ \gamma(t_1)=x(t_1)}} \int_{t_0}^{t_1} \mathcal{L}(\gamma(s),\dot{\gamma}(s)\mathrm{d}s, \end{aligned}$$

shows the inequality "$\leq$" of (3.50).

For the reverse, let $f \in C_b(E)$. There are curves γ_m satisfying $\gamma_m(t_0) = x(t_0)$ and

$$\inf_{\gamma(t_0)=x(t_0)}\left[f(x(t_1))-f(\gamma(t_1))+\int_{t_0}^{t_1}\mathcal{L}(\gamma(s),\dot\gamma(s)\mathrm{d}s\right]+\frac{1}{m}$$
$$\geq f(x(t_1))-f(\gamma_m(t_1))+\int_{t_0}^{t_1}\mathcal{L}(\gamma_m(s),\dot\gamma_m(s)\mathrm{d}s.$$

Since f is bounded, this implies $\limsup_{m\to\infty}\int_{t_0}^{t_1}\mathcal{L}(\gamma_m(s),\dot\gamma_m(s)\mathrm{d}s<\infty$. By compactness of sublevel sets, we can pass to a converging subsequence (denoted as well by γ_m). If $\gamma_m(t_1)\nrightarrow x(t_1)$, then $I_t(x(t_1)\mid x(t_0))=\infty$, and the desired estimate holds. If $\gamma_m(t_1)\to x(t_1)$, then by lower semicontinuity of $\gamma\to\int_{t_0}^{t_1}\mathcal{L}(\gamma(s),\dot\gamma(s)\mathrm{d}s$,

$$\inf_{\gamma(t_0)=x(t_0)}\left[f(x(t_1))-f(\gamma(t_1))+\int_{t_0}^{t_1}\mathcal{L}(\gamma(s),\dot\gamma(s)\mathrm{d}s\right]$$
$$\geq\liminf_{m\to\infty}f(x(t_1))-f(\gamma_m(t_1))+\int_{t_0}^{t_1}\mathcal{L}(\gamma_m(s),\dot\gamma_m(s)\mathrm{d}s$$
$$\geq\int_{t_0}^{t_1}\mathcal{L}(\gamma(s),\dot\gamma(s)\mathrm{d}s\geq\inf_{\substack{\gamma(t_0)=x(t_0)\\ \gamma(t_1)=x(t_1)}}\int_{t_0}^{t_1}\mathcal{L}(\gamma(s),\dot\gamma(s)\mathrm{d}s,$$

and the reverse inequality "$\geq$" of (3.50) follows. □

3.6 Proof of existence and uniqueness

In order to prove Proposition 3.1, we use the methods of Skorokhod's representation and pathwise splicing. In the following, we start with introducing Skorokhod's representation of fast process $\Lambda_n^\varepsilon(t)$.

3.6.1 Skorokhod's representation

The role of Skorokhod's representation is to represent the evolution of the discrete component $\Lambda_n^\varepsilon(t)$ in the form of a stochastic integral with respect to a Poisson random measure (see, for example, [Sha18]). Precisely, for each $x\in\mathbb{R}$, construct a family of intervals $\{\Gamma_{ij}^\varepsilon(x):i,j\in S\}$ on the half line in the following manner:

$$\Gamma_{12}^\varepsilon(x)=\big[0,\ \frac{1}{\varepsilon}q_{12}(x)\big)$$
$$\Gamma_{13}^\varepsilon(x)=\big[\frac{1}{\varepsilon}q_{12}(x),\ \frac{1}{\varepsilon}[q_{12}(x)+q_{13}(x)]\big)$$
$$\vdots$$
$$\Gamma_{1N}^\varepsilon(x)=\big[\frac{1}{\varepsilon}\sum_{j=1}^{N-1}q_{1j}(x),\ \frac{1}{\varepsilon}q_1(x)\big)$$
$$\Gamma_{21}^\varepsilon(x)=\big[\frac{1}{\varepsilon}q_1(x),\ \frac{1}{\varepsilon}[q_1(x)+q_{21}(x)]\big)$$

$$\Gamma_{23}^{\varepsilon}(x) = \Big[\frac{1}{\varepsilon}[q_1(x) + q_{21}(x)],\ \frac{1}{\varepsilon}[q_1(x) + q_{21}(x) + q_{23}(x)]\Big)$$
$$\vdots$$

and so on. Therefore, we obtain a sequence of consecutive, left-closed, right-open intervals $\Gamma_{ij}^{\varepsilon}(x)$ of $\mathbb{R}^+$, each having length $\frac{1}{\varepsilon}q_{ij}(x)$. For convenience of notation, we set $\Gamma_{ii}^{\varepsilon}(x) = \emptyset$ and $\Gamma_{ij}^{\varepsilon}(x) = \emptyset$ if $q_{ij}(x) = 0$, $i \neq j$. Define a function $h^{\varepsilon} : \mathbb{R} \times S \times M^{\varepsilon} \to \mathbb{R}$ by

$$h^{\varepsilon}(x, i, z) = \sum_{l \in S} (l - i) \mathbb{1}_{\Gamma_{il}^{\varepsilon}(x)}(z).$$

That is, with the partition $\{\Gamma_{ij}^{\varepsilon}(x) : i, j \in S \text{ with } i \neq j\}$ used and for each $i \in S$, if $z \in \Gamma_{ij}^{\varepsilon}(x), h^{\varepsilon}(x, i, z) = j - i$; otherwise $h^{\varepsilon}(x, i, z) = 0$. Then (3.8) is equivalent to

$$\mathrm{d}\Lambda_n^{\varepsilon}(t) = \int_{[0, M^{\varepsilon}]} h^{\varepsilon}(X_n^{\varepsilon}(t), \Lambda^{\varepsilon}(t-), z) N(\mathrm{d}t, \mathrm{d}z), \tag{3.53}$$

where $M^{\varepsilon} = N(N-1)H^{\varepsilon}$ with $H^{\varepsilon} := \max_{i,j \in S} \sup_{x \in \mathbb{R}} \frac{1}{\varepsilon} q_{ij}(x) < \infty$. $N(\mathrm{d}t, \mathrm{d}z)$ is a Poisson random measure (corresponding to a stationary point process $p(t)$) with intensity $\mathrm{d}t \times \mathbf{m}(\mathrm{d}z)$, and $\mathbf{m}(\mathrm{d}z)$ is the Lebesgue measure on $[0, M^{\varepsilon}]$. Note that $N(\cdot, \cdot)$ does not depend on ε, because the function $h^{\varepsilon}(\cdot, \cdot, \cdot)$ contains all information about $\frac{1}{\varepsilon}q(\cdot)$, see [XZ17, Proporsition 2.4]. $N(\cdot, \cdot)$ is independent of the Brownian motion $W(\cdot)$. Due to the finiteness of $\mathbf{m}(\cdot)$ on $[0, M^{\varepsilon}]$, there is only a finite number of jumps of the process $p(t)$ in each finite time interval. Let $\sigma_1 < \sigma_2 < \ldots < \ldots$ be the enumeration of all elements in the domain D_p of the stationary point process $p(t)$ corresponding to the above Poisson random measure $N(\mathrm{d}t, \mathrm{d}z)$. It follows that $\lim_{n \to \infty} \sigma_n = \infty$ almost surely.
Remark 3.22. (3.8) describes the evolution of the jump process, but it is difficult to study the existence and uniqueness of the system solution directly by using (3.8). Skorokhod's representation is a good approach to express the phenomenon (3.8) with the integral equation, and the information contained is not lost.

3.6.2 Proof of Proposition 3.1

For each $k \in S$, when we fixed a state, (3.7) becomes a classical CIR process

$$\mathrm{d}X_n^{\varepsilon,(k)}(t) = \eta(\mu(k) - X_n^{\varepsilon,(k)}(t))\mathrm{d}t + n^{-\frac{1}{2}}\theta\sqrt{X_n^{\varepsilon,(k)}(t)}\mathrm{d}W(t). \tag{3.54}$$

There exist a unique nonnegative strong solution of (3.54) with $2\eta u(i) \geq \theta^2$ due to [KS91, Proposition 5.2.13 and Corollary 5.3.23] by Yamada-Watanabe theorem.

Proof of Proposition 3.1. The idea of the proof comes from the stationary increments of Brownian motion and the poisson point process, and we segment the path-space through stopping time, see Figure 3.3.

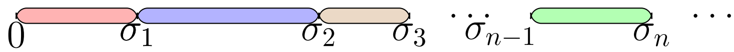

Figure 3.3: The time interval

Step 1: Let us first consider the solution in the time interval $[0, \sigma_1]$, where σ_1 is the stopping time. For any $t \in [0, \sigma_1)$ and any path $\{(X_n^\varepsilon(s), \Lambda_n^\varepsilon(s)) : 0 \le s \le t\}$, we always have

$$\int_0^t \int_{[0,M^\varepsilon]} h^\varepsilon(X_n^\varepsilon(s-), \Lambda_n^\varepsilon(s-), z) N_1(\mathrm{d}s, \mathrm{d}z) \equiv 0, \tag{3.55}$$

and then $\Lambda_n^\varepsilon(t) \equiv \Lambda_n^\varepsilon(0) = k$. Hence, on the interval $[0, \sigma_1)$, (3.7) is equivalent to (3.54) which has a unique strong solution $X_n^{\varepsilon,(k)}(t)$ with $X_n^{\varepsilon,(k)}(0) = x$, so $(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t)) = (X_n^{\varepsilon,(k)}(t), k)$ for $0 \le t < \sigma_1$. From (3.53) we have

$$\Lambda_n^\varepsilon(\sigma_1) = k + \sum_{l \in S} (l - k) \mathbf{1}_{\Gamma_{kl}^\varepsilon(X_n^{\varepsilon,(k)}(\sigma_1))}(p(\sigma_1)).$$

Then, on the time interval $[0, \sigma_1]$, set

$$(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t)) = \begin{cases} (X_n^{\varepsilon,(k)}(t), k), & 0 \le t < \sigma_1, \\ (X_n^{\varepsilon,(k)}(\sigma_1), \Lambda_n^\varepsilon(\sigma_1)), & t = \sigma_1 \end{cases} \tag{3.56}$$

Next, set $\tilde{\xi} = X_n^\varepsilon(\sigma_1)$, $\tilde{W}(t) = W(t + \sigma) - W(t)$ and $\tilde{p}(t) = p(t + \sigma_1)$.

Step 2: Similarly, we consider the solution on the interval $[0, \sigma_2 - \sigma_1]$ with respect to $(\tilde{\xi}, \Lambda_n^\varepsilon(\sigma_1))$ as above, and define

$$(\tilde{X}_n^\varepsilon(t), \tilde{\Lambda}_n^\varepsilon(t)) = (X_n^{\varepsilon,(\Lambda_n^\varepsilon(\sigma_1))}(t), \Lambda_n^\varepsilon(\sigma_1)) \quad \text{for} \quad 0 \le t < \sigma_2 - \sigma_1,$$

$$\tilde{X}_n^\varepsilon(\sigma_2 - \sigma_1) = X_n^{\varepsilon,(\Lambda^\varepsilon(\sigma_1))}(\sigma_2 - \sigma_1),$$

$$\tilde{\Lambda}_n^\varepsilon(\sigma_2 - \sigma_1) = \Lambda_n^\varepsilon(\sigma_1) + \sum_{l \in S} (l - \Lambda_n^\varepsilon(\sigma_1)) \mathbf{1}_{\tilde{A}_n^\varepsilon(l)}(\tilde{p}(\sigma_2 - \sigma_1)),$$

where

$$\tilde{A}_n^\varepsilon(l) = \Gamma_{\Lambda_n^\varepsilon(\sigma_1) l}^\varepsilon (X_n^{\varepsilon,(\Lambda_n^\varepsilon(\sigma_1))}(\sigma_2 - \sigma_1)-).$$

Furthermore, we define

$$(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t)) = (\tilde{X}_n^\varepsilon(t - \sigma_1), \tilde{\Lambda}_n^\varepsilon(t - \sigma_1)) \quad t \in [\sigma_1, \sigma_2],$$

which and (3.56) together give the solution on the time interval $[0, \sigma_2]$. Continuing this procedure inductively, $(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t))$ is determined uniquely on the time interval $[0, \sigma_n]$ for every n and thus $(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t))$ is determined globally because $\lim_{n\to\infty} \sigma_n = \infty$ almost surely.

Step 3: Consequently, we have proved the existence of a unique strong solution to the systems (3.7) and (3.53).

The proof is completed. □

Chapter 4

Large deviations on Riemannian manifolds

In Chapter 3, we have studied the large deviations of the slow-fast processes on Euclidean space. In this chapter, we consider a class of slow-fast processes on a connected complete Riemannian manifold M. The limiting dynamics as the scale separation goes to ∞ is governed by the averaging principle. Around this limit, we prove large deviation principles with an action-integral rate function for the slow process by nonlinear semigroup methods together with Hamilton-Jacobi-Bellman (HJB) equation techniques.

Our main innovation is solving the comparison principle for viscosity solutions for the HJB equation on M and the construction of a variational viscosity solution for the non-smooth Hamiltonian, which lies at the heart of deriving the action integral representation for the rate function.

This chapter is based on

[HKX24] Yanyan Hu, Richard C. Kraaij, and Fubao Xi. Large deviations for slow-fast processes on connected complete Riemannian manifolds. *Stochastic Process. Appl.*, 2024

4.1 Introduction

In this paper, let M be a d-dimensional connected complete Riemannian manifold and $S = \{1, 2, \dots, N\}$, $N < \infty$. We consider a stochastic differential equation consisting of Riemannian Brownian motion with a switching drift on $M \times S$ with an initial value

(x_0, k_0):

$$\mathrm{d}X_n^\varepsilon(t) = \frac{1}{\sqrt{n}} U_n^\varepsilon(t) \circ \mathrm{d}W(t) + b(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t))\mathrm{d}t, \tag{4.1}$$

where $\Lambda_n^\varepsilon(t)$ is a switching process with transition rate on the set S,

$$\mathbb{P}(\Lambda_n^\varepsilon(t+\Delta) = j \mid \Lambda_n^\varepsilon(t) = i, X_n^\varepsilon(t) = x) = \frac{1}{\varepsilon} q_{ij}(x)\Delta + o(\Delta), \quad \text{if } j \neq i, \tag{4.2}$$

for small $\Delta > 0$, $i, j \in S$, $x \in M$, and $\varepsilon > 0$ is a small parameter. $U_n^\varepsilon(\cdot)$ is a unique element such that $X_n^\varepsilon(t) = \mathbf{p}U_n^\varepsilon(t)$, where $\mathbf{p} : O(M) \to M$ is the canonical projection map from the orthonormal frame bundle on $O(M)$ to M. Precise details and conditions of this system will be specified later. Obviously, (4.1) and (4.2) together is a slow-fast system.

It is not too difficult to see that under some conditions, the effective behavior of the slow process (4.1) can be accurately described by the averaged system as $\varepsilon \to 0$ and $n \to \infty$, utilizing the averaging principle. To be more specific, if $X_n^\varepsilon(t) \approx x$, if the jump coefficient $x \mapsto q_{ij}(x)$ is continuous and the jump-matrix is uniformly ergodic, one expects that the fast process $\Lambda_n^\varepsilon(t)$ equilibrates in the stationary measure corresponding to the jump kernel.

This observation implies that, as $\varepsilon \to 0$ and $n \to \infty$, the slow process converges to an averaged process defined as follows

$$\begin{cases} \mathrm{d}\overline{X}(t) = \bar{b}(\overline{X}(t))\mathrm{d}t, \\ \overline{X}(0) = x_0, \end{cases} \tag{4.3}$$

where $\bar{b}(x) = \sum_{i \in S} b(x,i)\pi_i^x(t)$ and $\pi^x(t) = (\pi_i^x(t))_{i \in S}$ is the unique invariant probability measure of the fast process with the slow variable being "frozen" at a deterministic point $x \in M$. The application of this averaging principle provides an effective method to reduce computational complexity. It can be viewed as a variant of the law of large numbers.

In contrast to the averaging principle, the large deviation principle (LDP) excels in providing a more precise description of the dynamic behavior, it specifically addresses the characterization of the exponential decay rate associated with probabilities of rare events. Informally, LDP is the estimate of the form

$$\mathbb{P}(X_n(t) \approx \gamma(t)) \sim e^{-nI(\gamma)}, \qquad \text{as } n \to \infty,$$

for $\gamma : [0, \infty) \to M$. I takes the form

$$I(\gamma) = \begin{cases} I_0(\gamma(0)) + \int_0^\infty \mathcal{L}(\gamma(s), \dot{\gamma}(s))\,\mathrm{d}s, & \text{if } \gamma \in \mathcal{AC}(M), \\ \infty, & \text{otherwise}, \end{cases}$$

where $\mathcal{AC}$ denotes the set of absolutely continuous trajectories. I_0 quantifies the large deviations for $X_n(0)$ alone, and the map $\mathcal{L} : TM \to [0, \infty]$ is called the Lagrangian.

The large deviation principle indeed quantifies the decay of probabilities for trajectories away from the solution of the averaging principle (4.3), as the solution of (4.3) is the unique trajectory for which $I(\overline{X}) = 0$.

The main purpose of this paper is to prove an LDP around such an averaged process on M. The theory of LDP is one of the classical topics in probability theory, see [DZ98, dH08, FK06], which has widespread applications in different areas such as information theory, thermodynamics, statistics, and engineering.

Let us mention some works related to our purposes. Huang, Mandjes, and Spreij [HMS16], studied large deviations for Markov-modulated diffusion processes with rapid switching. In [PS24], Peletier and Schlottke proved pathwise LDP of switching Markov processes by exploiting the connection between Hamilton-Jacobi (HJ) equations and Hamilton-Jacobi-Bellman (HJB) equations. In [KS20], Kraaij and Schlottke studied the LDP for the slow-fast system under regular conditions, where the fast process is a switching process. For the proof, they used the Bootstrapping procedure, which is a technology for comparison principle of the HJB equation. Later, Della Corte and Kraaij [DCK24] continued to explore LDP in the context of molecular motors modeled by a diffusion process driven by the gradient of a weakly periodic potential that depends on an internal degree of freedom. The switch of the internal state, which can freely be interpreted as a molecular switch, is modeled as a Markov jump process that depends on the location of the motor. Subsequently, Hu, Kraaij, and Xi [HKX23] considered the Cox-Ingersoll-Ross processes with state-dependent fast switching in the case of the degenerate diffusion coefficient.

Although there are extensive results on LDPs for slow-fast systems in Euclidean space, there is not much work in the context of Riemannian manifolds. Röckner and Zhang [RZ04] studied sample path large deviations for diffusion processes on configuration spaces over a Riemannian manifold. Kraaij, Redig and Versendaal [KRV19] generalized classical large deviation theorems on complete, smooth Riemannian manifolds, and also considered Riemannian Brownian motion in the single time-scale context. Furthermore, Versendaal [Ver21] studied large deviations for $g(t)$-Brownian motion in a complete, evolving Riemannian manifold with respect to a collection $\{g(t)\}_{t\in\mathbb{R}}$ of Riemannian metrics, smoothly depending on t again in the single time-scale context.

Motivated by the aforementioned papers about LDP for slow-fast processes on Euclidean space and simple LDP on Riemannian manifold, it is a natural question to ask how to generalize the above large deviation results for slow-fast processes to Riemannian manifolds. In this paper, we address this question. That is we prove LDPs with an action-integral rate function for the slow process by nonlinear semigroup methods together with the HJB equation techniques. Note that our drift coefficient of slow process only satisfies locally one-sided Lipschitz continuity, which is weaker than the bounded condition. Moreover, the rate functions are related to the Hamiltonian $\mathcal{H} : T^*M \to [0, \infty]$ obtained by taking the Legendre transform of Lagrangian

$\mathcal{L} : TM \to [0, \infty]$. One formally defines that

$$\mathcal{H}(x, \mathrm{d}f(x)) = \sup_{\pi \in \mathcal{P}(S)} \left\{ \int B_{x,\mathrm{d}f(x)}(z)\pi(\mathrm{d}z) - \mathcal{I}(x, \pi) \right\},$$

where

$$B_{x,\mathrm{d}f(x)}(z) = b(x, z)\mathrm{d}f(x) + \frac{1}{2} |\mathrm{d}f(x)|^2$$

coming from the slow process $X_n(t)$ and Donsker-Varadhan function

$$\mathcal{I}(x, \pi) = - \inf_{g>0} \int \frac{R_x g(z)}{g(z)} \pi(\mathrm{d}z),$$

where R_x is the generator corresponding to the fast process $\Lambda_n(t)$ defined by

$$R_x g(z) = \sum_{j \in S} q_{zj}(x) \left(g(j) - g(z)\right).$$

Although following the proof ideas from Feng and Kurtz's book [FK06], considering the comparison principle and the existence of solutions of HJB equations, we need to put forward some new ideas to show those owing to the special properties of the Riemannian manifold.

We first discover special properties on M, which have caused difficulties but also is the key innovation in our proof:

(i) The first one, to ensure the exponential tightness, we find a good containment function:
$$\Upsilon(x) = \frac{1}{2} \log(1 + f^2(x)),$$
where the smooth function $f(x)$ approximates $d(x_0, x)$ for some $x_0 \in M$ and satisfying formally $\sup_z \mathcal{H}(z, \mathrm{d}\Upsilon(z)) < C < \infty$ which plays the role of a relaxed Lyapunov function.

(ii) The second one, the distance function $d(x, y)$, x, $y \in M$ is not always smooth. More specifically, $d(x, y)$ is not smooth on the cut-locus of x or y. This happens because the shortest path (geodesic) between two points may not be unique, for example, a spherical surface. Compared with $d(x, y)$, $d^2(x, y)$ is smooth when x closed to y. We use $d^2(x, y)$ in the proof of comparison principle.

(iii) The third one, we need to prove the global existence of solutions for an HJB equation on M to obtain an action-integral rate function. To establish existence we need to solve an appropriate control problem. A key obstacle is the construction from local solutions to global solutions.

4.2 Riemannian manifolds

In this section, we introduce some of the definitions, properties, and symbols that were mentioned earlier. This can be found in any textbook on Riemannian manifold, for example, [Lee03, Wan14].

Throughout the paper, (M, g) is a d-dimensional connected complete Riemannian manifold. We start with the definition of *chart*, which is used to prove Condition 4.32. A coordinate chart (or just a chart) on M is a pair $(\mathcal{O}, \varphi)$, where $\mathcal{O}$ is a homeomorphism from $\mathcal{O}$ to an open subset $\tilde{\mathcal{O}} = \varphi(\mathcal{O}) \subset \mathbb{R}^d$.

The tangent space of M at $x \in M$ is denoted by T_xM. We denote by $\langle\cdot,\cdot\rangle_x = g(\cdot,\cdot)$ the scalar product on T_xM with the associated norm $|\cdot|_x$, where the subscript x is sometimes omitted. The tangent bundle of M is denoted by $TM := \sqcup_{x\in M} T_xM$, which is naturally a manifold. Let $T_x^*M = (T_xM)^*$ be the *cotangent space* at $x \in M$, namely the dual space of the tangent space T_xM (the space of linear functions on T_xM). Let $T^*M = \sqcup_{x\in M} T_x^*M$, which is called the *cotangent bundle* on M.

Given a piecewise smooth curve $\gamma : [a, b] \to M$ joining x to y, i.e. $\gamma(a) = x$ and $\gamma(b) = y$, we can define the length of γ by $l(\gamma) = \int_a^b |\dot{\gamma}(t)|\mathrm{d}t$. Then the Riemannian distance $d(x, y)$, which induces the original topology on M, is defined by minimizing this length over the set of all such curves joining x to y.

Let ∇ be the Levi-Civita connection associated with the Riemannian metric. Let γ be a smooth curve in M. A vector field X is said to be parallel along γ if and only if $\nabla_{\dot{\gamma}_t} X = 0$. If $\dot{\gamma}$ itself is parallel along γ, we say that γ is a geodesic, and in this case $|\dot{\gamma}|$ is constant. When $|\dot{\gamma}| = 1$, γ is said to be normalized. A geodesic joining x to y in M is said to be minimal if its length equals $d(x, y)$.

A Riemannian manifold is complete if for any $x \in M$ all geodesics emanating from x are defined for all $-\infty < t < \infty$. By the Hopf-Rinow Theorem [Lee97, Theorem 6.13], we know that if M is complete then any pair of points in M can be joined by a minimal geodesic. Moreover, (M, d) is a complete metric space and bounded closed subsets are compact.

Given a (piecewise) smooth curve $\gamma : [a, b] \to M$, we denote parallel transport along γ from $\gamma(t_0)$ to $\gamma(t_1)$ by τ_{γ,t_0t_1}, or simply $\tau_{t_0t_1}$ whenever the meant curve is clear. If points x, $y \in M$ can be connected by a unique geodesic of minimal length, we will also write τ_{xy} meaning parallel transport from x to y along this specific geodesic.

The exponential map $\exp_x : T_xM \to M$ at x is defined by $\exp_x v = \gamma_v(1, x)$ for each $v \in T_xM$, where $\gamma(\cdot) = \gamma_v(\cdot, x)$ is the geodesic starting at x with velocity v. Then $\exp_x(tv) = \gamma_v(t, x)$ for each real number t. Note that the mapping $\exp_x$ is differentiable on T_xM for any $x \in M$.

In many cases, the minimal geodesic is not unique. For instance, for the unit sphere $\mathbb{S}^d$, each half circle linking the highest and the lowest points is a minimal geodesic.

This fact leads to the notion of cut-locus.

Definition 4.1. Let $x \in M$. For any $X \in \mathbb{S}_x := \{X \in T_x M : |X| = 1\}$, let

$$r(X) := \sup\{t > 0 : d(x, \exp_x(tX)) = t\}.$$

If $r(X) < \infty$ then we call $\exp_x(r(X)X)$ a cut-point of x. The set

$$\mathrm{cut}(x) := \{\exp_x(r(X)X) : X \in \mathbb{S}_x, r(X) < \infty\}$$

is called the *cut-locus* of the point of x. Moreover, the quantity

$$i_x := \inf\{r(X) : X \in \mathbb{S}_x\}$$

is called the *injectivity radius* of x. For any set $A \subseteq M$ we write $i(A) := \inf_{x \in A} i_x$ the injectivity radius of A.

Lemma 4.2 ([Kli82])**.** *The injectivity radius i_x depends continuously on x. In particular, if $K \subseteq M$ is compact we have $i(K) > 0$.*

Note that $i(K) > 0$ is used to find a smooth distance on M.

Definition 4.3. Let $\mathcal{T}(M)$ be the space of smooth vector fields on M and let ∇ be any connection on M. The formula

$$\mathcal{R}(X,Y)Z := \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z + \nabla_{[X,Y]} Z,$$

for X, Y, $Z \in \mathcal{T}(M)$, defines a function $\mathcal{R} : \mathcal{T}(M) \times \mathcal{T}(M) \times \mathcal{T}(M)$ called the Riemannian curvature of M, where $[X,Y] = XY - YX$ is the commutator of X and Y.

By taking the trace of the curvature tensor with respect to the first and the last entry, we obtain a 2-tensor which we will call the Ricci tensor of the (co)-metric g, denoted by $\mathcal{R}_g$.

To prove the existence solutions of HJB equations on M, we need the definitions of push-forward and pullback.

Definition 4.4 (Push-forward)**.** If M and N are smooth manifolds and $\varphi : M \to N$ is a smooth map, for each $p \in M$ we define a map

$$\varphi_{*p} : T_p M \to T_{\varphi(p)} N, \tag{4.4}$$

called the *push-forward* associated with φ, by

$$(\varphi_{*p}(v))(f) = v(f \circ \varphi), \quad v \in T_p M,\ f \in C^\infty(M).$$

Definition 4.5 (Pullback)**.** If M and N are smooth manifolds and $\varphi : M \to N$ be an invertible smooth map, for each $p \in M$ we define a map

$$\varphi_p^* : T^*_{\varphi(p)} N \to T^*_p M$$

by *pullback* associated with φ

$$(\varphi_p^* \xi)(v) = \xi(\varphi_{*p}(v)), \quad \xi \in T^*_{\varphi(p)} N,\ v \in T_p M. \tag{4.5}$$

We can put all φ_{*p} and φ_p^* together to obtain $\varphi_* : TM \to TN$ and $\varphi^* : T^*N \to T^*M$, respectively.

The next lemma shows that tangent vectors to curves behave well under composition with smooth maps.

Lemma 4.6 (Proposition 3.11 in [Lee03])**.** *Let $\varphi : M \to N$ be a smooth map, and let $\gamma : J \to M$ be a smooth curve, where $J \in \mathbb{R}$ is an interval. For any $t \in J$, the tangent vector to the composite curve $\varphi \circ \gamma$ at $t = t_0$ is given by*

$$(\varphi \dot{\circ} \gamma)(t_0) = (\varphi \circ \gamma)_* \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t_0} = \varphi_* \dot{\gamma}(t_0).$$

The chain rule for total derivatives is important in Riemannian manifolds because it allows us to compute the derivative of a composite function.

Lemma 4.7 (The chain rule for total derivatives, Proposition A.24 in [Lee03])**.** *Suppose V, W, X are finite-dimensional vector spaces, $U \subset V$ and $\tilde{U} \subset W$ are open sets, and $F : U \to \tilde{U}$ and $G : \tilde{U} \to X$ are maps. If F is differentiable at $a \in U$ and G is differentiable at $F(a) \in U$, then $G \circ F$ is differentiable at a, and*

$$D(G \circ F)(a) = DG(F(a)) \circ DF(a). \tag{4.6}$$

4.3 Constructing a diffusion process with fast switching on Riemannian manifolds

In the above section, we only gave the basic knowledge about the large deviation principle. We next state the definition of the orthonormal frame bundle and horizontal lift to construct a diffusion process with switching on $M \times S$, for which we want to study the large deviation behaviour,

$$\mathrm{d}X_n^\varepsilon(t) = \frac{1}{\sqrt{n}} U_n^\varepsilon(t) \circ \mathrm{d}W(t) + b(X_n^\varepsilon(t), \Lambda_n^\varepsilon(t))\mathrm{d}t, \tag{4.7}$$

where $\Lambda_n^\varepsilon(t)$ is a switching process with transition rate on the set S,

$$\mathbb{P}(\Lambda_n^\varepsilon(t+\triangle) = j \mid \Lambda_n^\varepsilon(t) = i, X_n^\varepsilon(t) = x) = \begin{cases} \frac{1}{\varepsilon} q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j \neq i, \\ 1 + \frac{1}{\varepsilon} q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j = i, \end{cases} \tag{4.8}$$

for small $\Delta > 0$, $i, j \in S$, $x \in M$, and $\varepsilon > 0$ is a small parameter.

We start by establishing that the above process exists. As the switch is taking place on the finite set S, the key issue to be resolved is the non-explosiveness of the diffusion process (4.7). In the context without switching, non-explosiveness is implied by a lower bound on the curvature and gradient of the drift. We will also assume this for our result.

Assumption 4.1. *For each $i \in S$, $b(\cdot, i)$ in (4.7) is a C^1-smooth vector field on M. There is a constant $\rho(n)$ such that the $CD(\rho(n), \infty)$ curvature condition*

$$\inf_{i \in S} \mathcal{R}_g - \nabla b(\cdot, i) \geq \rho(n) g$$

holds where $\mathcal{R}_g$ is the Ricci tensor of the (co)-metric g.
Theorem 4.8. *Under Assumption 4.1, the system, (4.7) and (4.8), has a unique non-explosive strong solution $(X_n(t), \Lambda_n(t))$ with initial value $(X_n(0), \Lambda_n(0)) = (x_0, k_0)$.*

The proof follows the same method as Proposition 3.1. We extend it to the context of Riemannian manifolds.

We next turn to present the definition of the orthonormal frame bundle and the horizontal lift.

Let $O_x(M)$ be the space of all orthonormal bases of $T_x M$. Denote $O(M) := \sqcup_{x \in M} O_x(M)$, which is called the *orthonormal frame bundle* over M. Obviously, $O_x(M)$ is isometric to $O(d)$, the group of orthogonal $(d \times d)$-matrices.

Let $\mathbf{p} : O(M) \to M$ with $\mathbf{p}u := x$ if $u \in O_x(M)$, which is called the *canonical projection* from $O(M)$ onto M. Now, given $e \in \mathbb{R}^d$, our goal is to define the corresponding horizontal vector field on $O(M)$. On the one hand, for any $u \in O(M)$ we have $ue \in T_{\mathbf{p}u} M$. Let u_s be the parallel transportation of u along the geodesic $\exp_{\mathbf{p}u}(sue)$, $s \geq 0$. We obtain a vector

$$H_e(u) := \frac{\mathrm{d}}{\mathrm{d}s} u_s|_{s=0} \in T_u O(M).$$

Thus, we have defined a vector filed H_e on $O(M)$ which is indeed C^∞-smooth. In particular, let $\{e_i\}_{i=1}^d$ be an orthonormal basis on $\mathbb{R}^d$, define

$$\Delta_{O(M)} := \sum_{i=1}^{d} H_{e_i}^2.$$

This operator is independent of the choice of the basis $\{e_i\}$. We call $\Delta_{O(M)}$ *the horizontal Laplace operator.* On the other hand, for any vector field Z on M, we define its *horizontal lift* by $\mathbf{H}_Z(u) := H_{u^{-1}Z}(u)$, $u \in O(M)$, where $u^{-1}Z$ is the unique vector $e \in \mathbb{R}^d$ such that $Z_{\mathbf{p}u} = ue$.

Let Δ_M be the Laplace-Beltrami operator,

$$\Delta_M f = \frac{1}{\sqrt{G}} \frac{\partial}{\partial x^i} \left(\sqrt{G} g^{ij} \frac{\partial f}{\partial x^j} \right), \quad f \in C^2(M). \tag{4.9}$$

We have the conclusion below, the horizontal Laplacian $\Delta_{O(M)}$ is the lift of the Laplace-Beltrami operator Δ_M to the orthonormal frame bundle $O(M)$.

Lemma 4.9 (Proposition 3.1.2 of [Hsu02]). *Let $f \in C^\infty(M)$, and $\tilde{f} = f \circ \mathbf{p}$ its lift to $O(M)$. Then for any $u \in O(M)$,*

$$\Delta_M f(x) = \Delta_{O(M)} \tilde{f}(u),$$

where $x = \mathbf{p}u$.

Having the preparations of orthonormal frame bundle and horizontal lift, we can establish a diffusion process and (4.7) with switching (4.8) in detail.

Proof of Theorem 4.8. We divide the proof into two steps.

Step 1: A SDE with a fixed switching state.

Let $b : \mathbb{R}^d \to TM$ be a C^1-smooth vector field on M. According to the idea of [Wan14, Section 2.1], we study a diffusion process generated by $A_n^M := \frac{1}{2n}\Delta_M + b$, where Δ_M is a Laplace-Beltrami operator in (4.9).

To this end, we first construct the corresponding *Horizontal diffusion process* generator by $A_n^{O(M)} := \frac{1}{n}\Delta_{O(M)} + \mathrm{H}_b$ on $O(M)$ by solving the Stratonovich stochastic differential equation

$$\mathrm{d}U_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{d} H_{e_j}(U_n(t)) \circ \mathrm{d}W^j(t) + H_b(U_n(t))\mathrm{d}t, \quad U_n(0) = u_0 \in O(M),$$

where $W(t) := (W^1(t), \ldots, W^d(t))$ is the d-dimensional Brownian motion on a complete filtered probability space $(\Omega,\ \mathcal{F},\ \{\mathcal{F}_t\}_{t\geq 0},\ \mathbb{P})$. Since H_b is C^1, it is well known that (see e.g. [Elw82, Chapter IV, Section 6]) the equation has a unique solution up to the lifetime $\zeta := \lim_{j\to\infty} \zeta_j$, where

$$\zeta_j := \inf\{t \geq 0 : \mathrm{d}(\mathbf{p}U,\ \mathbf{p}U_n(t)) \geq j\}, \quad j \geq 1.$$

Using Assumption 4.1, we further get that

$$\mathbb{P}(\zeta = \infty) = 1,$$

which means that ξ is the infinite lifetime, see [Hsu02, Section 4.2].

Let $X_n(t) = \mathbf{p}U_n(t)$. Then $X_n(t)$ solves the equation

$$\mathrm{d}X_n(t) = \frac{1}{\sqrt{n}} U_n(t) \circ \mathrm{d}W(t) + b(X_n(t))\mathrm{d}t, \quad X_n(0) = x_0 := \mathbf{p}u_0 \tag{4.10}$$

up to the infinite lifetime ζ. By the Itô formula, for any $f(\cdot) \in C_0^2(M)$,

$$f(X_n(t)) - f(x_0) - \int_0^t A_n^M f(X_n(s))\mathrm{d}s = \frac{1}{\sqrt{n}} \int_0^t \langle (U_n(s))^{-1}\mathrm{grad} f(X_n(s)), \mathrm{d}W(s)\rangle$$

is a martingale up to the infinite lifetime ζ; that is $X_n(t)$ is the diffusion process generated by A_n^M, and we call it the A_n^M-diffusion process. When $b=0$, then $X_n(t)$ is generated by $\frac{1}{2n}\Delta_M$ and is called the Brownian motion on M.

Step 2: the SDE with switching for any states. Here, we are going to introduce SDE with switching in (4.10). To achieve this, for $S=\{1,2\dots,N\}$ with $N<\infty$, we let the drift coefficient of the slow process depend on $i\in S$, where i represents the state of the switching process.

We construct the joint process as follows. Initialize the process from (x_0,k_0) and run the diffusion process $X_n(t)$ with $b=b(\cdot,k_0)$ in (4.10) as in Step 1. As this process has an infinite lifetime, we can wait until the first switch as indicated by the jump kernel

$$\mathbb{P}(\Lambda_n^\varepsilon(t+\triangle)=j\mid\Lambda_n^\varepsilon(t)=i,X_n^\varepsilon(t)=x)=\begin{cases}\frac{1}{\varepsilon}q_{ij}(x)\triangle+\circ(\triangle), & \text{if } j\neq i,\\ 1+\frac{1}{\varepsilon}q_{ij}(x)\triangle+\circ(\triangle), & \text{if } j=i,\end{cases}\tag{4.11}$$

for small $\Delta>0$, $i,j\in S$, $x\in M$, and $\varepsilon>0$ is a parameter.

We then run (4.10) with the state to which the jump kernel points us to jump. As S is finite, we can repeat this process and obtain our desired switching process with an infinite lifetime. □

4.3.1 The main results

In this paper, we consider the slow-fast systems (4.7) and (4.8). We first collect all the assumptions that are needed before giving the main results.
Assumption 4.2. *Let $\varepsilon=\frac{1}{n}$, this shows that small disturbance and fast switching have the same rate.*

This assumption means that the slow-fast system (4.7) and (4.8) becomes

$$\mathrm{d}X_n(t)=\frac{1}{\sqrt{n}}U_n(t)\circ\mathrm{d}W(t)+b(X_n(t),\Lambda_n(t))\mathrm{d}t,\tag{4.12}$$

and

$$\mathbb{P}(\Lambda_n(t+\triangle)=j\mid\Lambda_n(t)=i,\ X_n(t)=x)=\begin{cases}nq_{ij}(x)\triangle+\circ(\triangle), & \text{if } j\neq i,\\ 1+nq_{ij}(x)\triangle+\circ(\triangle), & \text{if } j=i.\end{cases}\tag{4.13}$$

In the following, we will focus on (4.12) and (4.13).
Assumption 4.3. *Fix $x_0\in M$ and define $r(x)=d(x,x_0)$. We say that b is linear growth if there exists a constant $C>0$ such that, for all $x\in M$,*

$$|b(x,i)|\le C(1+r(x)),\quad\forall\ i\in S.$$

Assumption 4.4. *We say that b is a locally one-sided Lipschitz function if, for any compact set $K\subseteq M$, there exists a constant $C_K>0$ such that, for all x, $y\in K$, it* *holds that*

$$\mathrm{d}_x\left(\frac{1}{2}d^2(\cdot,y)\right)(x)b(x,i)-\mathrm{d}_y\left(-\frac{1}{2}d^2(x,\cdot)\right)(y)b(y,i)\leq C_K d^2(x,y),\quad \forall\ i\in S,$$

where $d(x,y)<i(K)$ and $i(K)$ is the injectivity radius of K defined in Section 4.2.
Assumption 4.5. *For any $x\in M$, $(q_{ij}(x))_{i,j\in S}$ is a conservative, irreducible transition rate matrix, and $\sup_{i\in S}\sum_{j\in S,j\neq i}q_{ij}(x)<\infty$.*
Assumption 4.6. *For any compact sets $K\subseteq M$, there exists a constant $C_K>0$ such that*

$$|q_{ij}(x)-q_{ij}(y)|\leq C_K d(x,y),\quad x,\ y\in K,\ i,\ j\in S.$$

Then, we give some remarks on these assumptions.

- Assumption 4.3 controls the rate at which the process may deviate to prove exponential tightness.
- Assumption 4.4 is set for proving the comparison principle.
- Assumptions 4.5 and 4.6 of a fast switching process for any given x ensures the existence of an invariant probability measure that satisfies the averaging principle.

In the following, we give the main result.
Theorem 4.10 (Large deviations for slow processes)**.** *Let $(X_n(t),\Lambda_n(t))$ be the Markov processes on $M\times S$. Consider the setting of Assumptions 4.2, 4.3, 4.4, 4.5 and 4.6. Suppose that the large deviation principle holds for $X_n(0)$ on M with speed n and a good rate function I_0.*

Then, the large deviation principle is satisfied with speed n for the processes $X_n(t)$ with a good rate function I having action-integral representation,

$$I(\gamma)=\begin{cases}I_0(\gamma(0))+\int_0^\infty\mathcal{L}\left(\gamma(s),\dot\gamma(s)\right)\mathrm{d}s, & \text{if } \gamma\in\mathcal{AC}(M),\\ \infty, & \text{otherwise.}\end{cases}$$

where $\mathcal{L}:TM\to[0,\infty]$ is the Legendre transform of $\mathcal{H}$ given by $\mathcal{L}(x,v)=\sup_{p\in T_x^*M}\{\langle v,p\rangle-\mathcal{H}(x,p)\}$, and

$$\mathcal{H}(x,\mathrm{d}f(x))=\sup_{\pi\in\mathcal{P}(S)}\left\{\int B_{x,\mathrm{d}f(x)}(z)\pi(dz)-\mathcal{I}(x,\pi)\right\}\tag{4.14}$$

where

$$B_{x,\mathrm{d}f(x)}(z)=b(x,z)\mathrm{d}f(x)+\frac{1}{2}\left|\mathrm{d}f(x)\right|^2$$

coming from the slow process $X_n(t)$ and Donsker-Varadhan function

$$\mathcal{I}(x,\pi)=-\inf_{g>0}\int\frac{R_xg(z)}{g(z)}\pi(dz),$$

where R_x is the generator corresponding to the fast process $\Lambda_n(t)$ defined by

$$R_x g(z) = \sum_{j \in S} q_{zj}(x) \left(g(j) - g(z) \right).$$

The proof of Theorem 4.10 follows the strategy used in Section 2.2, where the most crucial part is proving the assumptions in Proposition 2.14. However, in this chapter, our process involves fast-slow systems on a Riemannian manifold, so Proposition 2.14 has been modified.

Proposition 4.11 (Adaptation of Proposition 2.14 to our context)**.** *Let $(X_n(t), \Lambda_n(t))$ be Markov processes on $M \times S$. Suppose that*

(a) *$X_n(0)$ satisfies large deviation principle;*

(b) *there exists an operator $H \subset ex - \lim_{n\to\infty} H_n$ in the sense Definition 2.13 on Riemannian manifold;*

(c) *we have exponential compact containment of the process $(X_n(t), \Lambda_n(t))$;*

(d) *for all $\lambda > 0$ and $h \in C_b(M)$, the comparison principle holds for $f - \lambda H f = h$.*

Then the following hold:

(i) (Limit of nonlinear semigroup) There exists a unique operator semigroup $V(t)$ such that

$$\lim_{n\to\infty} \|V_n(t) f_n - V(t) f\| = 0 \tag{4.15}$$

and there exists a unique $R(\lambda)f$ such that

$$\lim_{m\to\infty} \|R(t/m)^m f - V(t) f\| = 0, \tag{4.16}$$

whenever $f \in \overline{\mathcal{D}(H)}$, $f_n \in C_b(M \times S)$, and $\|f_n - f\| \to 0$.

(ii) (Large deviation principle) $X_n(t)$ satisfies the large deviation principle with good rate function I given by

$$I(x) = I_0(x(t_0)) + \sup_{k\in\mathbb{N}} \sup_{0=t_0<t_1<\cdots<t_k<\infty} \sum_{i=0}^{k} I^V_{t_{i+1}-t_i}(x(t_{i+1}) \mid x(t_i)), \tag{4.17}$$

where for $\Delta t = t_{i+1} - t_i > 0$ and $x(t_{i+1})$, $x(t_i) \in M$, the conditional rate functions $I^V_{\Delta t}(x(t_{i+1}) \mid x(t_i))$ are

$$I^V_{\Delta t}(x(t_{i+1}) \mid x(t_i)) = \sup_{f \in C_b(M)} [f(x(t_{i+1})) - V(\Delta t) f(x(t_i))].$$

Suppose in addition that

(e) $V(t) = \mathbf{V}(t)$ *with* $\mathbf{V}$ *as in* (2.10).

Then the rate function (4.17) *can be represented in the following action-integral form:*

$$I(\gamma) = \begin{cases} I_0(\gamma(0)) + \int_0^\infty \mathcal{L}(\gamma(s), \dot{\gamma}(s))ds, & \text{if } \gamma \in \mathcal{AC}(M), \\ \infty, & \text{otherwise.} \end{cases} \tag{4.18}$$

The proof of Theorem 4.10 is thus immediate upon checking Proposition 4.11 (a) to (e) for our switching diffusion. We will verify (a) to (d) in Section 4.4 and (e) in Section 4.5.

4.4 The proof of Proposition 4.11 (a) to (d)

In this section, we will establish Proposition 4.11 (a) to (d) for our switching diffusion:

Using the discussion in the previous section, we can prove items (i) and (ii) of Theorem 4.10 once the following four facts are established:

- Item (b): we obtaining a limiting multi-valued Hamiltonian $H \subseteq \text{ex} - \lim H_n$ in Section 4.4.1;
- Part of item (d): We identify a single valued Hamiltonian $\mathbf{H}$ via a suitable eigenvalue problem in Section 4.4.1;
- Item (c): we obtain the compact containment condition in Section 4.4.2;
- Part of item (d): we prove the comparison principle for H and $\mathbf{H}$ in Section 4.4.3.

4.4.1 Identification of a multi-valued Hamiltonian

Our first goal is to obtain a multi-valued Hamiltonian $H \subseteq \text{ex} - \lim H_n$. We consider the solution $(X_n(t), \Lambda_n(t))$ of the system (4.12) and (4.13) with the generator A_n^M:

$$A_n^M f(x,i) = \frac{1}{2n}\Delta_M f(x,i) + b(x,i)\mathrm{d}f(x,i) + n\sum_{i\in S} q_{ij}(x)(f(x,j) - f(x,i)). \tag{4.19}$$

We give a multi-valued limit Hamiltonian by the generator A_n^M. Denote by $C_c^2(M)$ the set of twice continuously differentiable functions that are constant outside a compact set.

Proposition 4.12 (Multi-valued limit Hamiltonian)**.** *Let* $(X_n(t), \Lambda_n(t))$ *be a Markov process on* $M \times S$ *with generator* A_n^M *in* (4.19)*. Set* $H_n = \frac{1}{n}e^{nf}A_n^M e^{nf}$ *and*

$$H := \Big\{(f, H_{f,\phi}) \mid f \in C_b^2(M), H_{f,\phi} \in C_b(M \times S) \text{ and } \phi \in C_b^2(M \times S)\Big\}, \tag{4.20}$$

where

$$H_{f,\phi}(x,i) = b(x,i)\mathrm{d}f(x) + \frac{1}{2}|\mathrm{d}f(x)|^2 + \sum_{j\in S} q_{ij}(x)[e^{\phi(x,j)-\phi(x,i)} - 1]. \tag{4.21}$$

Then, $H \subset ex - \lim_{n\to\infty} H_n$.

Proof. By the generator A_n^M in (4.19), for $\mathrm{e}^{nf} \in \mathcal{D}(A_n^M)$ we get a nonlinear generator

$$\begin{aligned} H_n f(x,i) &= \frac{1}{n} e^{-nf} A_n^M e^{nf}(x,i) \\ &= b(x,i)\mathrm{d}f(x,i) + \frac{1}{2}|\mathrm{d}f(x,i)|^2 + \frac{1}{2n}\Delta_M f(x,i) \\ &\quad + \sum_{j\in S} q_{ij}(x)[e^{n(f(x,j)-f(x,i))} - 1]. \end{aligned} \tag{4.22}$$

When $n \to \infty$, (4.22) is not convergent due to the divergence of the fourth term. To proceed, instead of using f in (4.22), we take a sequence

$$f_n(x,i) = f(x) + \frac{1}{n}\phi(x,i), \quad \forall\, f \in C_b^2(M) \text{ and } \phi \in C_b^2(M \times S).$$

As $\mathrm{d}f_n(x,i) = \mathrm{d}f(x) + \frac{1}{n}\mathrm{d}\phi(x,i)$, (4.22) implies

$$\begin{aligned} H_n f_n(x,i) = b(x,i) \left(\mathrm{d}f(x) + \frac{1}{n}\mathrm{d}\phi(x,i)\right) + \frac{1}{2}\left|\mathrm{d}f(x) + \frac{1}{n}\mathrm{d}\phi(x,i)\right|^2 \\ + \frac{1}{2n}\Delta_M \left(f(x) + \frac{1}{n}\phi(x,i)\right) + \sum_{j\in S} q_{ij}(x)[e^{\phi(x,j)-\phi(x,i)} - 1]. \end{aligned}$$

Taking $n \to \infty$ gives the following uniform limit:

$$H_{f,\phi}(x,i) = b(x,i)\mathrm{d}f(x) + \frac{1}{2}\left|\mathrm{d}f(x)\right|^2 + \sum_{j\in S} q_{ij}(x)[e^{\phi(x,j)-\phi(x,i)} - 1],$$

establishing the claim. □

A single valued Hamiltonian via the eigenvalue problem

In the multi-valued operator H, we seek a single-valued operator that we will use to establish the comparison principle in Section 4.4.3 below. In particular, we aim to find for any $f \in \mathcal{D}(H)$ a unique g such that $(f,g) \in H$ and g do not depend on $i \in S$. This unique g will then be the basis to define $\mathbf{H}f$.

Consider (4.21) of Proposition 4.12:

$$H_{f,\phi}(x,i) = b(x,i)\mathrm{d}f(x) + \frac{1}{2}|\mathrm{d}f(x)|^2 + \sum_{j\in S} q_{ij}(x)[e^{\phi(x,j)-\phi(x,i)} - 1]. \tag{4.23}$$

As the eigenvalue problem is one in terms of the fast process, we decompose (4.23) into a function depending on i

$$B_{x,\mathrm{d}f(x)}(i) = b(x,i)\mathrm{d}f(x) + \frac{1}{2}|\mathrm{d}f(x)|^2 \tag{4.24}$$

and the jump operator acting on the state i:

$$R_x \mathrm{e}^{\phi(x,i)} = \sum_{j\in S} q_{ij}(x)[e^{\phi(x,i)} - e^{\phi(x,j)}]$$

We thus seek a $\overline{\phi}$ such that there is a constant $\mathcal{H}(x,\mathrm{d}f(x))$ such that

$$\mathcal{H}(x,\mathrm{d}f(x)) := B_{x,\mathrm{d}f(x)}(i) + e^{-\overline{\phi}(x,i)} R_x e^{\overline{\phi}(x,i)} \tag{4.25}$$

is independent of i. Rewriting this equation in terms of $\overline{g} = e^{\overline{\phi}}$, we thus aim to find $\overline{g}$ and $\mathcal{H}(x,\mathrm{d}f(x))$ such that

$$\left(R_x + B_{x,\mathrm{d}f(x)}\right)\overline{g}(i) = \mathcal{H}(x,\mathrm{d}f(x))\overline{g}(i).$$

In other words, we aim to find the principal eigenfunction and eigenvalue for the operator $R_x + B_{x,\mathrm{d}f(x)}$ in terms of i, which can be carried out using the Perron-Frobenius theorem and leads to the representation (4.14).

Proposition 4.13 (Principal-eigenvalue problem)**.** *Let Assumption 4.5 be satisfied. For each $(x,\mathrm{d}f(x))$, there exist $\overline{g} > 0$ and a unique eigenvalue $\mathcal{H}(x,\mathrm{d}f(x)) \in \mathbb{R}$ such that*

$$\left(R_x + B_{x,\mathrm{d}f(x)}\right)\overline{g} = \mathcal{H}(x,\mathrm{d}f(x))\overline{g}, \tag{4.26}$$

with $\mathcal{H}(x,\mathrm{d}f(x))$ given by

$$\begin{aligned}\mathcal{H}(x,\mathrm{d}f(x)) &= \sup_{\pi\in\mathcal{P}(S)} \inf_{g>0} \int \frac{\left(R_x + B_{x,\mathrm{d}f(x)}\right)g(i)}{g(i)}\pi(\mathrm{d}i) \\ &= \sup_{\pi\in\mathcal{P}(S)} \left\{\int B_{x,\mathrm{d}f(x)}(i)\pi(\mathrm{d}i) - \mathcal{I}(x,\pi)\right\}\end{aligned} \tag{4.27}$$

where

$$\mathcal{I}(x,\pi) = -\inf_{g>0}\int \frac{R_x g(i)}{g(i)}\pi(\mathrm{d}i). \tag{4.28}$$

Proof. Using Assumption 4.5, from the Perron-Frobenius theorem in [DV75], we can obtain there exists a unique eigenvalue with associated eigenfunction which has the representation (4.27). $\square$

We next aim to define a new operator in terms of $\mathcal{H}$. We first note the following result that can be obtained as in [DCK24, Propositions 4.7 and 4.8].

Lemma 4.14. *The map $\mathcal{H}$ in* (4.27) *is continuous in* (x,p) *and convex in* p *for fixed* x*.*

As a direct consequence, we are able the introduce our single valued operator $\mathbf{H}$. Recall that $C_c^2(M)$ is the set of twice continuously differentiable functions that are constant outside of a compact set.

Definition 4.15. Set $\mathbf{H} \subseteq C_b(M) \times C_b(M)$ with domain $\mathcal{D}(\mathbf{H}) = C_c^2(M)$ by

$$\mathbf{H}f(x) := \mathcal{H}(x, \mathrm{d}f(x)) \tag{4.29}$$

with $\mathcal{H}$ as in (4.27).

4.4.2 Exponential compact containment

In this section, the key step, in obtaining exponential tightness on a Riemannian manifold, is to find a good containment function that can limit our analysis to a compact set.

Definition 4.16 (Good containment function)**.** We say that $\Upsilon : M \to \mathbb{R}$ is a good containment function (for H) if

(Υa) $\Upsilon \geq 0$ and there exists a point x_0 such that $\Upsilon(x_0) = 0$,

(Υb) Υ is twice continuously differentiable,

(Υc) for every $c \geq 0$, the set $\{x \in M \mid \Upsilon(x) \leq c\}$ is compact,

(Υd) we have $\sup_x \mathcal{H}(x, \mathrm{d}\Upsilon(x)) < \infty$.

Let us denote by d the Riemannian distance function associated to the metric g. Fix $x_0 \in M$ and consider the radial function $r(x) = d(x, x_0)$. Since r is not everywhere smooth, it is not suitable for constructing a good containment function as in Definition 4.16. However, since r is 1-Lipschitz (with respect to the metric g), we can find a smooth function f with $f(x_0) = r(x_0) = 0$ and such that $\|f - r\| \leq 1$ and $|\mathrm{d}f| \leq 2$. Using this, we define Υ by

$$\Upsilon(x) = \frac{1}{2}\log(1 + f^2(x)). \tag{4.30}$$

We now show that Υ can be used as a good containment function. The following is an adaptation of [Ver21, Proposition 4.11].

Lemma 4.17. *Let M be a complete Riemannian manifold. Under Assumption 4.3, Υ defined in* (4.30) *is a good containment function for the Hamiltonian $\mathcal{H}$ in* (4.29)*.*

Proof. This proof is inspired by [KRV19, Ver21], and is therefore only different in checking property d. We prefer to spell out the proof of (Υa)-(Υc) as it will be used to prove (Υd).

Clearly $\Upsilon \geq 0$ and $\Upsilon(x_0) = 0$, and $\Upsilon \in C^\infty(M)$.

Furthermore, since r is smooth, it follows that Υ is smooth. Now, for $c > 0$, the continuity of Υ implies that $\{x \in M \mid \Upsilon(x) \le c\}$ is closed. Furthermore, $\Upsilon(x) \le c$ implies that $f(x) \le \sqrt{e^{2c} - 1}$. It follows that $d(x, x_0) \le 1 + f(x) \le 1 + \sqrt{e^{2c} - 1}$. Hence, $\{x \in M \mid \Upsilon(x) \le c\}$ is bounded. Since M is complete, we conclude that $\{x \in M \mid \Upsilon(x) \le c\}$ is compact.

Note that for all $x \in M$,

$$\mathrm{d}\Upsilon(x) = \frac{f(x)}{1 + f^2(x)} \mathrm{d}f(x). \tag{4.31}$$

This, together with Assumption 4.3 and $|\mathrm{d}f| \le 2$, for $z \in S$, we first estimate that

$$\begin{aligned} b(x,z)\mathrm{d}\Upsilon(x) &= b(x,z)\mathrm{d}f(x)\frac{f(x)}{1+f^2(x)} \\ &\le |b(x,z)| \cdot |\mathrm{d}f(x)| \cdot \frac{f(x)}{1+f^2(x)} \\ &\le C(2 + f(x))\frac{f(x)}{1+f^2(x)}. \end{aligned} \tag{4.32}$$

Hence, $\sup_{x,z} b(x,z)\mathrm{d}\Upsilon(x) < \infty$. Now recall the Hamiltonian $\mathcal{H}$ in (4.29), from (4.32), we obtain

$$\begin{aligned} \mathcal{H}(x, \mathrm{d}\Upsilon(x)) &= \sup_{\pi \in \mathcal{P}(S)} \left\{ \int B_{x,\mathrm{d}\Upsilon(x)}(z)\pi(\mathrm{d}z) - \mathcal{I}(x,\pi) \right\} \\ &\le \int B_{x,\mathrm{d}\Upsilon(x)}(z)\pi(\mathrm{d}z) \\ &= \int \left(b(x,z)\mathrm{d}\Upsilon(x) + \frac{1}{2}|\mathrm{d}\Upsilon(x)|^2 \right) \pi(\mathrm{d}z) \\ &\le C \int \left(\frac{f(x)}{1+f^2(x)} + \frac{f^2(x)}{1+f^2(x)} + \frac{f^2(x)}{(1+f^2(x))^2} \right) \pi(\mathrm{d}z), \end{aligned}$$

where the first inequality uses the definition of supremum and $\mathcal{I}(x,\pi)$ is nonnegative. We conclude that $\sup_x \mathcal{H}(x, \mathrm{d}\Upsilon(x)) < \infty$, which implies that Υ is a good containment function. $\square$

Applying the good containment function (4.30), we proceed to consider the exponential compact containment of the system $(X_n(t), \Lambda_n(t))$.

Proposition 4.18 (Exponential compact containment condition)**.** *Let $(X_n(t), \Lambda_n(t))$ be a Markov process corresponding to A_n^M. Then the exponential compact containment condition as in Definition 2.11.*

The result follows using martingale control techniques as in Proposition 3.9 using Υ from Lemma 5.4.

4.4.3 Comparison principle

One of the key steps in the modern doubling of variables procedure in the comparison principle proofs is the estimate

$$\mathcal{H}\left(x_\alpha, \mathrm{d}_x \frac{\alpha}{2} d^2(\cdot, y_\alpha)\right)(x_\alpha) - \mathcal{H}\left(y_\alpha, -\mathrm{d}_y \frac{\alpha}{2} d^2(x_\alpha, \cdot)\right)(y_\alpha) \leq \alpha C d^2(x_\alpha, y_\alpha), \quad (4.33)$$

for suitable x_α, y_α satisfying $\alpha d^2(x_\alpha, y_\alpha) \to 0$ as $\alpha \to \infty$.

In our case, the Hamiltonian is that

$$\mathcal{H}(x, \mathrm{d}f(x)) = \sup_{\pi \in \mathcal{P}(S)} \left\{ \int B_{x,\mathrm{d}f(x)}(z)\pi(\mathrm{d}z) - \mathcal{I}(x, \pi) \right\},$$

where

$$B_{x,\mathrm{d}f(x)}(z) = b(x,z)\mathrm{d}f(x) + \frac{1}{2}\left|\mathrm{d}f(x)\right|^2. \quad (4.34)$$

In the proof of Lemma 3.16 below, we will pick an optimizer π^* for $\mathcal{H}\left(y, -\mathrm{d}_y \frac{\alpha}{2} d^2(x_\alpha, \cdot)\right)(y_\alpha)$, so that the estimate (4.33) reduces to

- the use of Assumption 4.4 to control the difference of the two terms that include the drift b;
- properties of the Riemannian metric d to threat the quadratic part $|df|^2$, see Lemma 4.19 below;
- estimates on objects of the type $|\mathcal{I}(x_\alpha, \pi^*) - \mathcal{I}(y_\alpha, \pi^*)|$, see Lemma 4.30.

A final issue arises from the fact that the metric d^2 is non-differentiable on the cut-locus, which we will treat by using that as $\alpha d^2(x_\alpha, y_\alpha) \to 0$, we will always work outside of the cut-locus.

Smooth distance functions

We first present the differential property of the distance function to deal with the quadratic part in (4.34), the proof is shown in [KRV19, Appendix C.1].

Lemma 4.19. *Let x, $y \in M$ and assume that $x \notin cut(y)$ (or equivalently, $y \notin cut(x)$), where $cut(\cdot)$ is a cut-locus. Then for all $V \in T_y M$ we have*

$$\mathrm{d}_y(d^2(x, \cdot))(y)(V) = 2\langle \dot{\gamma}(1),\ V\rangle_{g(y)},$$

where $\gamma : [0,1] \to M$ is the unique geodesic of minimal length connecting x and y. Consequently, we obtain

$$\tau_{x,y}\mathrm{d}_x(d^2(\cdot, y))(x) = -\mathrm{d}_y(d^2(x, \cdot))(y). \quad (4.35)$$

Remark 4.20. Note that (4.35) implies that if $x \notin \mathrm{cut}(y)$ (or equivalently, $y \notin \mathrm{cut}(x)$), we have

$$|\mathrm{d}_x(d^2(\cdot, y))(x)|^2_{g(x)} = |\mathrm{d}_y(d^2(x, \cdot))(y)|^2_{g(y)}$$

useful for estimating the quadratic part in the estimate (4.33).

Our approach to proving the comparison principle is to double variables, as in the classical setting of viscosity solutions in Euclidean spaces, using the distance function as a penalizing function.

Lemma 4.21 (Lemma A.10 in [CK17])**.** *Let u be bounded and upper semicontinuous, let v be bounded and lower semicontinuous, and let Υ be a good containment function as defined in* (4.30)*.*

Fix $\delta > 0$. For every $m > 0$ there exist points $x_{\delta,m}$, $y_{\delta,m} \in M$, such that

$$\begin{aligned} &\frac{u(x_{\delta,m})}{1-\delta} - \frac{v(y_{\delta,m})}{1+\delta} - \frac{m}{2} d^2(x_{\delta,m}, y_{\delta,m}) - \frac{\delta}{1-\delta}\Upsilon(x_{\delta,m}) - \frac{\delta}{1+\delta}\Upsilon(y_{\delta,m}) \\ &= \sup_{x,y\in M} \left\{ \frac{u(x)}{1-\delta} - \frac{v(y)}{1+\delta} - \frac{m}{2} d^2(x,y) - \frac{\delta}{1-\delta}\Upsilon(x) - \frac{\delta}{1+\delta}\Upsilon(y) \right\}. \end{aligned} \tag{4.36}$$

Additionally, for every $\delta > 0$ we have that

(a) *The set $\{x_{\delta,m},\ y_{\delta,m} \mid m > 0\}$ is relatively compact in M.*

(b) *All limit points of $\{(x_{\delta,m},\ y_{\delta,m})\}_{m>0}$ are of the form $(z,\ z)$ and for these limit points we have*
$$u(z) - v(z) = \sup_{x\in M} u(x) - v(x).$$

(c) *We have*
$$\lim_{m\to\infty} m d^2(x_{\delta,m},\ y_{\delta,m}) = 0.$$

In the proof of the comparison principle, Lemma 4.29, below, we will have to work with smooth test functions that are derived from the optimization procedure in (4.36) above. Due to the presence of the cut-locus, smoothness of $\frac{m}{2}d^2$ is, however, not guaranteed. For any fixed δ, we see that the injectivity radius $i(K) = \inf_{x\in K} i_x$ is bounded away from 0 on the compact K obtained in (a). Thus by (c) our optimizing values must lie in the complement of the cut-locus for large m. The next lemma allows us to replace d^2 by a smooth function behaving similarly outside the cut-locus.

Lemma 4.22. *For any compact set $K \subseteq M$, there is smooth function $\Psi_K : M^2 \to [0,\infty)$ satisfying*

$$\begin{aligned} \Psi_K(x,y) &= \frac{1}{2} d^2(x,y) && \text{if } d(x,y) \le \frac{i(K)}{2}, \\ \Psi_K(x,y) &> \frac{1}{8} i(K)^2 && \text{if } d(x,y) > \frac{i(K)}{2}. \end{aligned}$$

The proof is similar to Lemma 7.7 of [KRV19].

The necessary operators for proving comparison principle
To prove the comparison principle for the Hamilton-Jacobi equation in terms of H and relate it to the variational Hamiltonian $\mathbf{H}$ of Definition 4.15, we introduce two new pairs of Hamiltonians (H_1, H_2) and $(H_\dagger, H_\ddagger)$ that serve as natural upper and lower bounds for H and $\mathbf{H}$ respectively. These new Hamiltonians are both defined in terms of the containment function Υ of (4.30), which introduces unboundedness in our test functions, allowing us to work with optimizing points in the definition of viscosity sub and supersolutions.

Denote by $C_l^\infty(M)$ the set of smooth functions on M that has a lower bound and by $C_u^\infty(M)$ the set of smooth functions on M that has an upper bound. Denote $C_\Upsilon := \sup_{(x,i)\in M\times S} B_{x,\mathrm{d}\Upsilon(x)}(i) < \infty$.
Definition 4.23 (Multi-valued operators)**.** Recall the definition of $H_{f,\phi}$ in (4.21).

- For $f \in C_l^\infty(M)$, $\delta \in (0,1)$ and $\phi \in C_b^2(M \times S)$. Set

$$f_1^\delta(x) := (1-\delta)f(x) + \delta\Upsilon(x),$$

$$H_{1,f,\phi}^\delta(x,i) := (1-\delta)H_{f,\phi}(x,i) + \delta C_\Upsilon,$$

  and set

$$H_1 := \left\{ \left(f_1^\delta, H_{1,f,\phi}^\delta\right) \,\middle|\, f \in C_l^\infty(M), \delta \in (0,1),\ \phi \in C_b^2(M\times S)\right\}.$$

- For $f \in C_u^\infty(M)$, $\delta \in (0,1)$ and $\phi \in C_b^2(M \times S)$. Set

$$f_2^\delta(x) := (1+\delta)f(x) - \delta\Upsilon(x),$$

$$H_{2,f,\phi}^\delta(x,i) := (1+\delta)H_{f,\phi}(x,i) - \delta C_\Upsilon,$$

  and set

$$H_2 := \left\{ \left(f_2^\delta, H_{2,f,\phi}^\delta\right) \,\middle|\, f \in C_u^\infty(M),\ \delta \in (0,1),\ \phi \in C_b^2(M\times S)\right\}.$$

We use the single valued Hamiltonian $\mathbf{H}$ to define two new single valued operators.
Definition 4.24 (Single valued operators)**.** Recall the definition of $\mathcal{H}(x, \mathrm{d}f(x))$ of (4.27).

- For $f \in C_l^\infty(M)$ and $\delta \in (0,1)$ set

$$f_\dagger^\delta(x) := (1-\delta)f(x) + \delta\Upsilon(x),$$

$$H_{\dagger,f}^\delta(x) := (1-\delta)\mathcal{H}(x,\mathrm{d}f(x)) + \delta C_\Upsilon,$$

  and set

$$H_\dagger := \left\{ \left(f_\dagger^\delta, H_{\dagger,f}^\delta\right) \,\middle|\, f \in C_l^\infty(M),\ \delta \in (0,1)\right\}.$$

- For $f \in C_u^\infty(M)$ and $\delta \in (0,1)$ set
$$f_\ddagger^\delta(x) := (1+\delta)f(x) - \delta\Upsilon(x),$$
$$H_{\ddagger,f}^\delta(x) := (1+\delta)\mathcal{H}(x, \mathrm{d}f(x)) - \delta C_\Upsilon,$$
and set
$$H_\ddagger := \left\{ \left(f_\ddagger^\delta,\ H_{\ddagger,f}^\delta\right) \,\middle|\, f \in C_u^\infty(M),\ \delta \in (0,1)\right\}.$$

We collect H, $\mathbf{H}$, $H_\dagger$, $H_\ddagger$, H_1 and H_2 in Figure 3.1, which intuitively provides the proof strategy for the comparison principle in the following subsection. Note that to obtain the comparison principle for H only the left-hand side of the figure is necessary. We aim to establish a variational expression for the rate function, however, by showing that $V(t) = \mathbf{V}(t)$. This we will carry out in Section 4.5 on which we will show that the variational resolvent will give viscosity solutions in terms of the Hamilton-Jacobi equation in terms of $\mathbf{H}$. The right-hand side of the figure will show that all viscosity solutions under consideration must be the same.

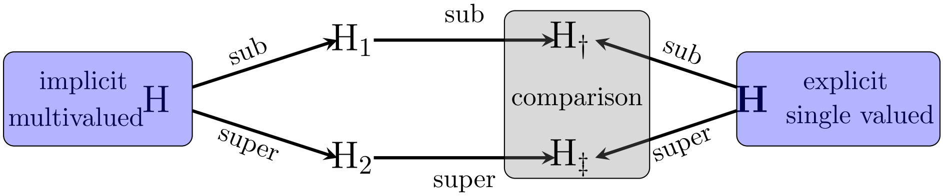

Figure 4.1: An arrow connecting an operator A with operator B with subscript 'sub' means that viscosity subsolutions of $f - \lambda A f = h$ are also viscosity subsolutions of $f - \lambda B f = h$. Similarly, we get the description for arrows with a subscript 'super'. The middle gray box around the operators $H_\dagger$ and $H_\ddagger$ indicates that the comparison principle holds for subsolutions of $f - \lambda H_\dagger f = h$ and supersolutions of $f - \lambda H_\ddagger f = h$. The left blue box indicates that H is an implicit and multi-valued operator. The right blue box indicates $\mathbf{H}$ is an explicit single valued operator.

Main propositions: comparison principle

Based on the above preparations, we are ready to state the proposition of this subsection.

Proposition 4.25 (Comparison principle)**.** *Let Assumptions 4.3, 4.4, 4.5 and 4.6 be satisfied. Let h_1, $h_2 \in C_b(M)$ and $\lambda > 0$. Let u be any subsolution to $f - \lambda H f = h_1$ and let v be any supersolution to $f - \lambda H f = h_2$. Then we have that*
$$\sup_{x \in M} u(x) - v(x) \leq \sup_x h_1(x) - h_2(x).$$

Proof. The result is immediate from Lemmas 4.26, 4.28, and 3.16 below. ☐

The proofs of the next three lemmas are analogous to those in [HKX23].
Lemma 4.26. *Let Assumption 4.5 be satisfied. Fix $\lambda > 0$ and $h \in C_b(M)$.*

(a) *Every subsolution to $f - \lambda H_1 f = h$ is also a subsolution to $f - \lambda H_\dagger f = h$.*

(b) *Every supersolution to $f - \lambda H_1 f = h$ is also a supersolution to $f - \lambda H_\ddagger f = h$.*

Lemma 4.27. *Let Assumption 4.3 be satisfied. Fix $\lambda > 0$ and $h \in C_b(M)$.*

(a) *Every subsolution to $f - \lambda \mathbf{H} f = h$ is also a subsolution to $f - \lambda H_\dagger f = h$.*

(b) *Every supersolution to $f - \lambda \mathbf{H} f = h$ is also a supersolution to $f - \lambda H_\ddagger f = h$.*

Lemma 4.28. *Let Assumption 4.3 be satisfied. Fix $\lambda > 0$ and $h \in C_b(M)$.*

(a) *Every subsolution to $f - \lambda H f = h$ is also a subsolution to $f - \lambda H_1 f = h$.*

(b) *Every supersolution to $f - \lambda H f = h$ is also a supersolution to $f - \lambda H_2 f = h$.*

In addition to the lemmas above, we still need to verify the comparison principle for $f - \lambda H_\dagger f = h_1$ and $f - \lambda H_\ddagger f = h_2$ on M from Figure 3.1.
Lemma 4.29. *Suppose Assumptions 4.3, 4.4, 4.5 and 4.6 hold. Let $h_1, h_2 \in C_b(M)$ and $\lambda > 0$. Let u be any subsolution to $f - \lambda H_\dagger f = h_1$ and let v be any supersolution to $f - \lambda H_\ddagger f = h_2$. Then we have*

$$\sup_{x \in M} u(x) - v(x) \leq \sup_{x \in M} h_1(x) - h_2(x). \tag{4.37}$$

Proof. For a sub and supersolution u and v, $\delta \in (0,1)$ and $m \geq 1$, we follow (4.36) and set

$$\Phi_{\delta,m}(x,y) := \frac{u(x)}{1-\delta} - \frac{v(y)}{1+\delta} - \frac{m}{2} d^2(x,y) - \frac{\delta}{1-\delta}\Upsilon(x) - \frac{\delta}{1+\delta}\Upsilon(y), \tag{4.38}$$

By Lemma 4.21, we find a compact set K and $(x_{\delta,m}, y_{\delta,m}) \in K \times K$ satisfying

$$\Phi_{\delta,m}(x_{\delta,m},\ y_{\delta,m}) = \sup_{(x,y) \in M \times M} \Phi_{\delta,m}(x,\ y). \tag{4.39}$$

By Lemma 4.22, we can replace $\frac{m}{2} d^2$ by Ψ_K and consider

$$\widehat{\Phi}_{\delta,m}(x,y) := \frac{u(x)}{1-\delta} - \frac{v(y)}{1+\delta} - m\Psi_K(x,y) - \frac{\delta}{1-\delta}\Upsilon(x) - \frac{\delta}{1+\delta}\Upsilon(y). \tag{4.40}$$

It follows from 4.39 that for large m we have

$$\widehat{\Phi}_{\delta,m}(x_{\delta,m},\ y_{\delta,m}) = \sup_{(x,y) \in M \times M} \widehat{\Phi}_{\delta,m}(x,\ y). \tag{4.41}$$

In view of (4.41), it follows that $x_{\delta,m}$ is the unique maximizing point for

$$\sup_{x\in M} u(x) - \varphi_1^{\delta,m}(x) = u(x_{\delta,m}) - \varphi_1^{\delta,m}(x_{\delta,m})$$

where $\varphi_1^{\delta,m}$ is constructed by taking the appropriate remaining terms of (4.38), with an additional penalization $(1-\delta)d^2(x, x_{\delta,m})$ to turn $x_{\delta,m}$ into the unique optimizer:

$$\begin{aligned}
\varphi_1^{\delta,m}(x) :&= -(1-\delta)\Phi_{\delta,m}(x, y_{\delta,m}) + u(x) + (1-\delta)d^2(x, x_{\delta,m}) \\
&= (1-\delta)\left(-\frac{u(x)}{1-\delta} + \frac{v(y_{\delta,m})}{1+\delta} + m\Psi_K(x, y_{\delta,m}) + \frac{\delta}{1-\delta}\Upsilon(x) + \frac{\delta}{1+\delta}\Upsilon(y_{\delta,m})\right) \\
&\quad + u(x) + (1-\delta)d^2(x, x_{\delta,m}) \\
&= (1-\delta)\left(\frac{v(y_{\delta,m})}{1+\delta} + m\Psi_K(x, y_{\delta,m}) + \frac{\delta}{1-\delta}\Upsilon(x) + \frac{\delta}{1+\delta}\Upsilon(y_{\delta,m})\right) \\
&\quad + (1-\delta)d^2(x, x_{\delta,m}) \\
&= (1-\delta)\left(m\Psi_K(x, y_{\delta,m}) + d^2(x, x_{\delta,m}) + \frac{\delta}{1+\delta}\Upsilon(y_{\delta,m}) + \frac{v(y_{\delta,m})}{1+\delta}\right) + \delta\Upsilon(x).
\end{aligned}$$

Since u is a viscosity subsolution of $f - \lambda H_\dagger f = h_1$ we conclude that

$$u(x_{\delta,m}) - \lambda\left[(1-\delta)\mathcal{H}(x_{\delta,m}, p^1_{\delta,m}) + \delta C_\Upsilon\right] \leq h_1(x_{\delta,m}), \tag{4.42}$$

where for large m

$$p^1_{\delta,m} := m d_x \Psi_K(\cdot, y_{\delta,m})(x_{\delta,m}) = m \, \mathrm{d}_x\left(\frac{1}{2}d^2(\cdot, y_{\delta,m})\right)(x_{\delta,m}). \tag{4.43}$$

Similarly, we obtain that $y_{\delta,m}$ it the unique optimizer for

$$\inf_{y\in M} v(x) - \varphi_2^{\delta,m}(y) = v(y_{\delta,m}) - \varphi_2^{\delta,m}(y_{\delta,m}),$$

where

$$\varphi_2^{\delta,m}(y) := (1+\delta)\left(m\Psi_K(x_{\delta,m}, y) - d^2(y, y_{\delta,m}) - \frac{\delta}{1-\delta}\Upsilon(x_{\delta,m}) + \frac{u(x_{\delta,m})}{1-\delta}\right) - \delta\Upsilon(y).$$

As v is a viscosity supersolution of $f - \lambda H_\ddagger f = h_2$, we then know that

$$v(x_{\delta,m}) - \lambda\left[(1+\delta)\mathcal{H}(y_{\delta,m}, p^2_{\delta,m}) - \delta C_\Upsilon\right] \geq h_2(y_{\delta,m}), \tag{4.44}$$

where for large m

$$p^2_{\delta,m} := -m \, \mathrm{d}_y\left(\frac{1}{2}d^2(x_{\delta,m}, \cdot)\right)(y_{\delta,m}). \tag{4.45}$$

By item (c) of Lemma 4.21, we have

$$\lim_{m\to\infty} m d^2(x_{\delta,m}, y_{\delta,m}) = 0. \tag{4.46}$$

Taking (4.42), (4.44) and (4.46) into account, we obtain that

$$\begin{aligned}
\sup_{x\in M} u(x) - v(x) &\le \liminf_{\delta\to 0}\liminf_{m\to\infty}\left(\frac{u(x_{\delta,m})}{1-\delta} - \frac{v(y_{\delta,m})}{1+\delta}\right) \\
&\le \liminf_{\delta\to 0}\liminf_{m\to\infty}\bigg\{\frac{h_1(x_{\delta,m})}{1-\delta} - \frac{h_2(y_{\delta,m})}{1+\delta} && (4.47)\\
&\quad + \frac{\delta}{1-\delta}C_\Upsilon + \frac{\delta}{1+\delta}C_\Upsilon && (4.48)\\
&\quad + \lambda\big(\mathcal{H}(x_{\delta,m},p^1_{\delta,m}) - \mathcal{H}(y_{\delta,m},p^2_{\delta,m})\big)\bigg\}, && (4.49)
\end{aligned}$$

where in the first inequality we use (4.39) and drop the nonnegative functions $d^2(\cdot,\cdot)$ and $\Upsilon(\cdot)$.

The term (4.48) vanishes as $\delta \to 0$. For the term (4.47), the sequence $(x_{\delta,m}, y_{\delta,m})$ takes its values in a compact set and, hence, admits converging subsequences as $m \to \infty$. By (b) of Lemma 3.18, these subsequences converge to points of the form (x, x). Hence, by the above analysis, we get

$$\begin{aligned}
\sup_{x\in M} u(x) - v(x) &\le \lambda \liminf_{\delta\to 0}\liminf_{m\to\infty}\big(\mathcal{H}(x_{\delta,m},p^1_{\delta,m}) - \mathcal{H}(y_{\delta,m},p^2_{\delta,m})\big)\\
&\quad + \sup_{x\in M} h_1(x) - h_2(x).
\end{aligned}$$

It follows that the comparison principle holds for $f - \lambda H_\dagger f = h_1$ and $f - \lambda H_\ddagger f = h_2$ whenever for any $\delta > 0$

$$\liminf_{m\to\infty}\big(\mathcal{H}(x_{\delta,m},p^1_{\delta,m}) - \mathcal{H}(y_{\delta,m},p^2_{\delta,m})\big) \le 0. \tag{4.50}$$

To that end, recall $\mathcal{H}(x, \mathrm{d}f(x))$ in (4.29):

$$\mathcal{H}(x,\mathrm{d}f(x)) = \sup_{\pi\in\mathcal{P}(S)}\left\{\int B_{x,\mathrm{d}f(x)}(z)\pi(\mathrm{d}z) - \mathcal{I}(x,\pi)\right\},$$

where $\pi \mapsto \int B_{x,\mathrm{d}f(x)}(z)\pi(\mathrm{d}z)$ is bounded and continuous, and $\mathcal{I}(x,\cdot)$ has compact sub-level sets in $\mathcal{P}(S)$. Thus, there exists an optimizer $\pi_{\delta,m} \in \mathcal{P}(S)$ such that

$$\mathcal{H}(x_{\delta,m},p^1_{\delta,m}) = \int B_{x_{\delta,m},p^1_{\delta,m}}(z)\pi_{\delta,m}(\mathrm{d}z) - \mathcal{I}(x_{\delta,m},\pi_{\delta,m}) \tag{4.51}$$

and

$$\mathcal{H}(y_{\delta,m},p^2_{\delta,m}) \ge \int B_{y_{\delta,m},p^2_{\delta,m}}(z)\pi_{\delta,m}(\mathrm{d}z) - \mathcal{I}(y_{\delta,m},\pi_{\delta,m}). \tag{4.52}$$

Combining (4.51) and (4.52), we obtain

$$\mathcal{H}(x_{\delta,m},p^1_{\delta,m}) - \mathcal{H}(y_{\delta,m},p^2_{\delta,m})$$

$$\leq \int \left(B_{x_{\delta,m},p^1_{\delta,m}}(z) - B_{y_{\delta,m},p^2_{\delta,m}}(z)\right)\pi_{\delta,m}(\mathrm{d}z) \tag{4.53}$$

$$+ \mathcal{I}(y_{\delta,m},\pi_{\delta,m}) - \mathcal{I}(x_{\delta,m},\pi_{\delta,m}). \tag{4.54}$$

It is enough to prove that (4.53) and (4.54) go to 0 as $m \to \infty$. For (4.53), by calculating the difference of integrand $B_{x,p}$ in detail, for any $z \in S$, from (4.24), (4.43), (4.45) and Remark 4.20, one has

$$\begin{aligned}
&B_{x_{\delta,m},p^1_{\delta,m}}(z) - B_{y_{\delta,m},p^2_{\delta,m}}(z)\\
&= m\mathrm{d}_x\left(\frac{1}{2}d^2(\cdot,y_{\delta,m})\right)(x_{\delta,m})b(x_{\delta,m},z) + \frac{1}{2}\left|m\ \mathrm{d}_x\left(\frac{1}{2}d^2(\cdot,y_{\delta,m})\right)(x_{\delta,m})\right|^2\\
&- \left[-m\ \mathrm{d}_y\left(\frac{1}{2}d^2(x_{\delta,m},\cdot)\right)(y_{\delta,m})b(y_{\delta,m},z) + \frac{1}{2}\left|-m\ \mathrm{d}_y\left(\frac{1}{2}d^2(x_{\delta,m},\cdot)\right)(y_{\delta,m})\right|^2\right]\\
&= m\mathrm{d}_x\left(\frac{1}{2}d^2(\cdot,y_{\delta,m})\right)(x_{\delta,m})b(x_{\delta,m},z) + m\ \mathrm{d}_y\left(\frac{1}{2}d^2(x_{\delta,m},\cdot)\right)(y_{\delta,m})b(y_{\delta,m},z)\\
&+ \frac{m^2}{2}\left(\left|\mathrm{d}_x\left(\frac{1}{2}d^2(\cdot,y_{\delta,m})\right)(x_{\delta,m})\right|^2 - \left|-\mathrm{d}_y\left(\frac{1}{2}d^2(x_{\delta,m},\cdot)\right)(y_{\delta,m})\right|^2\right) \qquad (4.55)\\
&= m\mathrm{d}_x\left(\frac{1}{2}d^2(\cdot,y_{\delta,m})\right)(x_{\delta,m})b(x_{\delta,m},z) + m\ \mathrm{d}_y\left(\frac{1}{2}d^2(x_{\delta,m},\cdot)\right)(y_{\delta,m})b(y_{\delta,m},z)\\
&\leq Cd^2(x_{\delta,m},y_{\delta,m}),
\end{aligned}$$

where in the last inequality, we use Assumption 4.4. Noting that the last term in line 5 vanishes. This is happened because, fix $\delta > 0$, there is a compact $K^\delta \subseteq M$ such that $\{x_{m,\delta},\ y_{m,\delta} \mid m > 0\}$ is contained in K^δ by item (a) of Lemma 4.21. By the continuity of the injectivity radius and the compactness of K^δ, we can find a $\Delta > 0$ such that $i(K^\delta) \geq \Delta > 0$. Then there exists a unique geodesic of minimal length connecting $x_{m,\delta}$ and $y_{m,\delta}$. Furthermore, by Lemma 4.19 we have

$$\mathrm{d}_x d^2(\cdot,y_{m,\delta})(x_{m,\delta}) = -\tau_{x_{m,\delta},\ y_{m,\delta}}\mathrm{d}_y d^2(x_{m,\delta},\ \cdot)(y_{m,\delta}), \tag{4.56}$$

where $\tau_{x_{m,\delta},y_{m,\delta}}$ denotes parallel transport along the unique geodesic of minimal length connecting $x_{m,\delta}$ and $y_{m,\delta}$. As parallel transport is an isometry, we find as in Remark 4.20 that

$$\left|\mathrm{d}_x\left(\frac{1}{2}d^2(\cdot,y_{\delta,m})\right)(x_{\delta,m})\right|^2_{g(x_{\delta,m})} = \left|-\mathrm{d}_y\left(\frac{1}{2}d^2(x_{\delta,m},\cdot)\right)(y_{\delta,m})\right|^2_{g(y_{\delta,m})}$$

Hence, (4.53) is sufficiently small, as $m \to \infty$, using (4.46) and (4.55). To obtain that (4.54) is sufficiently small, we utilize the equi-continuity of $\mathcal{I}(\cdot,\pi)$ established in Lemma 4.30 below for the spatial variable. This finishes the proof of (4.50) and the comparison principle for $H_\dagger$ and $H_\ddagger$. □

Here, we state the equi-continuity of $\mathcal{I}(\cdot,\pi)$ to finish the proof of the comparison principle of $H_\dagger$ and $H_\ddagger$ in Lemma 3.16. The proof is analogous to Lemma 3.20.

Lemma 4.30. *Let Assumption 4.6 be satisfied. Recall* (4.28)*:*

$$\mathcal{I}(x,\pi) = -\inf_{g>0}\int \frac{R_x g(z)}{g(z)}\pi(\mathrm{d}z).$$

For any compact set $K \subseteq M$ and for all $\pi \in \mathcal{P}(S)$, then $\{x \mapsto J(x,\pi)\}_{x\in K,\pi\in\mathcal{P}(S)}$ is equi-continuous.

4.5 The proof of Proposition 4.11 (e)

In this final chapter, we will establish (e) of Proposition 4.11, which is the key statement to obtain the variational representation of the rate function in Theorem 4.10.

The proof is based on the analysis of variational semigroups and resolvents of Chapter 8 in [FK06] and are based on their main Conditions 8.9, 8.10, and 8.11 of [FK06], which we adapt to the Riemannian context below as Conditions 4.31 and 4.32.

We will then carry out two main steps.

- We will show in Section 4.5.1 which key results of [FK06, Chapter 8] are used to obtain $V(t) = \mathbf{V}(t)$, and how this relates to our set-up in Section 4.4.
- We verify in Sections 4.5.2 and 4.5.3 Conditions 4.31 and 4.32 respectively in our context.

Condition 4.31. Suppose that the map $T^*M \ni (x,p) \to \mathcal{H}(x,p) \in \mathbb{R}$ is continuous, and is convex in the second variable p. Define $\mathcal{L}$ as its Legendre transform. Suppose that there is a good containment function Υ for $\mathcal{H}$. Then

(a) the function $\mathcal{L} : TM \to [0,\infty]$ is lower semi-continuous and for each compact set $K \subseteq M$ and $c \in \mathbb{R}$ the set

$$\{(x,v) \in TM \mid x \in K,\ \mathcal{L}(x,v) \le c\}$$

is compact in TM.

(b) for each compact $K \subseteq M$, any finite time $T > 0$ and finite bound $C \ge 0$, there exists a compact set $\hat{K} = \hat{K}(K,T,C) \subseteq M$ such that $x \in \mathcal{AC}(M)$ and $x(0) \in K$, if

$$\int_0^T \mathcal{L}(x(s),\dot{x}(s))\mathrm{d}s \le C,$$

then $x(t) \in \hat{K}$ for all $0 \le t \le T$.

(c) for each compact set $K \subseteq M$ and $c \in \mathbb{R}$, there exists a right-continuous non-decreasing function $\psi_{K,c} : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\lim_{r\to\infty} r^{-1}\psi_{K,c}(r) = 0. \tag{4.57}$$

and

$$|\mathrm{d}f(x)v| \leq \psi_{K,c}(\mathcal{L}(x,v)), \qquad \forall\,(x,v) \in TM,\ x \in K,$$

where $f \in C_{K,c}$ and

$$C_{K,c} := \left\{ f \in C_c^2(M) \,\middle|\, \forall\, x \in K,\ |df(x)| \leq c \right\}. \tag{4.58}$$

Condition 4.32. For any initial point $x(0) \in M$, $T > 0$ and $f \in \mathcal{D}(\mathbf{H})$, there exists an absolutely continuous curve $x : [0,T] \to M$ such that for all $0 < t \leq T$

$$\int_0^t \mathcal{H}(x(s), \mathrm{d}f(x(s)))\mathrm{d}s + \int_0^t \mathcal{L}(x(s), \dot{x}(s))\mathrm{d}s = \int_0^t \mathrm{d}f(x(s))\dot{x}(s)\mathrm{d}s. \tag{4.59}$$

4.5.1 Connecting Conditions 4.31 and 4.32 to Section 4.4

In this section, we state two results of [FK06] and show how these can be used to obtain $V(t) = \mathbf{V}(t)$. For readability, we repeat the definitions of $\mathbf{V}$ and $\mathbf{R}$:

$$\mathbf{V}(t)f(x) := \sup_{\substack{\gamma \in \mathcal{AC} \\ \gamma(0)=x}} \left\{ f(\gamma(t)) - \int_0^t \mathcal{L}(\gamma(s), \dot{\gamma}(s))\mathrm{d}s \right\} \tag{4.60}$$

and

$$\mathbf{R}(\lambda)h(x) := \sup_{\substack{\gamma \in \mathcal{AC} \\ \gamma(0)=x}} \left\{ \int_0^\infty \lambda^{-1} e^{-\lambda^{-1} t} \left(h(\gamma(t)) - \int_0^t \mathcal{L}(\gamma(r), \dot{\gamma}(r))\mathrm{d}r \right) \mathrm{d}s \right\}. \tag{4.61}$$

Proposition 4.33 (Lemma 8.18 of [FK06]). *Let Conditions 4.31 and 4.32 be satisfied. For any $f \in C_b(M)$, $t \geq 0$ and $x \in M$, we have*

$$\lim_{m \to \infty} |\mathbf{R}(t/m)^m f(x) - \mathbf{V}(t)f(x)| = 0.$$

The next result of [FK06], obtainable from the proof of Theorem 8.27 in [FK06], establishes that the variational resolvent gives viscosity solutions for the operator $\mathbf{H}$.

Proposition 4.34. *Let Conditions 4.31 and 4.32 be satisfied. Then we have*

- $\mathbf{R}(\lambda)h$ *is a visocosity subsolution to* $f - \lambda \mathbf{H} f = h$,
- *the lower semi-continuous regularization* $(\mathbf{R}(\lambda)h)_*$ *of* $\mathbf{R}(\lambda)h$ *is a visocosity supersolution to* $f - \lambda \mathbf{H} f = h$.

Combining these two statements with Proposition 4.11 (i), it follows that Proposition 4.11 (e), namely that $V(t) = \mathbf{V}(t)$, is satisfied if $R(\lambda) = \mathbf{R}(\lambda)$. Using the results of Section 4.4, we thus obtain the following result:

Proposition 4.35. *Let Conditions 4.31 and 4.32 be satisfied. Then $V(t)f = \mathbf{V}(t)$.*

Proof. By Proposition 4.34, Lemma 4.27 and Lemma 3.16, $\mathbf{R}(\lambda)h$ equals the unique viscosity solution to the pair of equations

$$f - \lambda H_\dagger f = h, \qquad f - \lambda H_\ddagger f = h,$$

and thus equals $R(\lambda)h$ from Proposition 4.11. By Propositions 4.33 and Proposition 4.11 (i), it follows that $V(t) = \mathbf{V}(t)$ establishing the claim. □

We are left to prove Conditions 4.31 and 4.32 in the following two sections.

4.5.2 Verification of Condition 4.31

In this section, we verify Condition 4.31.
Proposition 4.36. *Let Assumption 4.3 be satisfied. Then Condition 4.31 holds.*

Proof. To obtain Item (a), observe that $\mathcal{L} \geq 0$ follows from $\mathcal{H}(x, 0) = 0$. The Lagrangian $\mathcal{L}$ is convex, and lower semicontinuous as it is the Legendre transform of $\mathcal{H}$. For $C \geq 0$, we prove that the set $\{(x, v) \in TM : x \in K, \mathcal{L}(x, v) \leq C\}$ is bounded, and hence is relatively compact. For any $p \in T_x^*M$ and $v \in T_xM$, we have

$$pv \leq \mathcal{L}(x, v) + \mathcal{H}(x, p) \quad x \in K.$$

Thereby, if $\mathcal{L}(x, v) \leq C$, then

$$|v| = \sup_{|p|=1} pv \leq \sup_{|p|=1} [\mathcal{L}(x, v) + \mathcal{H}(x, p)] \leq C + C_1,$$

where C_1 exists due to continuity of $\mathcal{H}$ obtained in Lemma 4.14 and $x \in K$. Then for $R := C + C_1$,

$$\{(x, v) \in TM : \mathcal{L}(x, v) \leq C\} \subseteq \{v : |v| \leq R\},$$

thus $\{\mathcal{L} \leq C\}$ is a bounded subset in TM.

For item (b), recalling that by Assumption 4.3 and Lemma 4.17 the level sets of Υ are compact and we control the growth of Υ. For $K \subseteq M$, $T > 0$, $C \geq 0$ and $x \in \mathcal{AC}(M)$ as above, this follows by noting that

$$\begin{aligned}
\Upsilon(x(t)) &= \Upsilon(x(0)) + \int_0^t \mathrm{d}\Upsilon(x(s))\dot{x}(s)\mathrm{d}s \\
&\leq \Upsilon(x(0)) + \int_0^t [\mathcal{L}(x(s), \dot{x}(s)) + \mathcal{H}(x(s), \mathrm{d}\Upsilon(x(s)))]\, \mathrm{d}s \\
&\leq \sup_{y \in K} \Upsilon(y) + C_1 + T \sup_{z \in M} \mathcal{H}(z, \mathrm{d}\Upsilon(z)) = C < \infty,
\end{aligned}$$

for any $0 \leq t \leq T$, so that the compact set $\hat{K} = \{z \in M : \Upsilon(x) \leq C\}$ satisfies the condition.

Proof of (c) is inspired by that of Lemma 10.21 of [FK06]. We first prove that $\mathcal{L}(x,v)$ is superlinear. Recall that by Lemma 4.14 $\mathcal{H}$ is continuous, which implies

$$\overline{H}_K(c) := \sup_{x\in K} \sup_{p\in T_x^*M, |p|\leq c} \mathcal{H}(x,p) < \infty.$$

Using the definition of $\mathcal{L}$, it thus follows for any $(x,v)\in TM$, $x\in K$ with $|v|>0$ that

$$\frac{\mathcal{L}(x,v)}{|v|} \geq \sup_{p\in T_x^*M,\, |p|\leq c} \frac{pv}{|v|} - \frac{\overline{H}_K(c)}{|v|} = c - \frac{\overline{H}_K(c)}{|v|}$$

It follows that

$$\lim_{N\uparrow\infty} \inf_{x\in K} \inf_{v\in T_xM:|v|=N} \frac{\mathcal{L}(x,v)}{|v|} = \infty.$$

Secondly, for $s\geq 0$, define the map $\vartheta(s)$ by

$$\vartheta(s) := s \inf_{x\in K} \inf_{v\in T_xM:|v|\geq s} \frac{\mathcal{L}(x,v)}{|v|}. \tag{4.62}$$

It thus follows that ϑ is a strictly increasing function satisfying

$$\lim_{s\uparrow\infty} \frac{\vartheta(s)}{s} = \infty. \tag{4.63}$$

Next, define $\Psi_{K,c}(r) =: C_{K,c}\vartheta^{-1}(r)$ with $\vartheta^{-1}(r) = \inf\{\omega : \vartheta(\omega)\geq r\}$. By monotonicity of ϑ, we have for any $x\in K$ that

$$\vartheta(C_{K,c}^{-1}|\mathrm{d}f(x)v|) \overset{(4.58)}{\leq} \vartheta(|v|) \overset{(4.62)}{\leq} \mathcal{L}(x,v).$$

Hence by monotonicity of $\Psi_{K,c}$, we find $|\mathrm{d}f(x)v| \leq \Psi_{K,c}(\mathcal{L}(x,v))$ for any $f\in C_{K,c}$, and $(x,v)\in TM$ with $x\in K$. Finally (4.57) follows by (4.63) and the definition of $\vartheta^{-1}(r)$. □

4.5.3 Verification of Condition 4.32

In this section, we verify Condition 4.32: the construction of curve with arbitrary lifetime, starting point and $f\in\mathcal{D}(\mathbf{H}) = C_c^2(M)$ satisfying

$$\int_0^t \mathcal{H}(x(s),\mathrm{d}f(x(s)))\mathrm{d}s + \int_0^t \mathcal{L}(x(s),\dot{x}(s))\mathrm{d}s = \int_0^t \mathrm{d}f(x(s))\dot{x}(s)\mathrm{d}s. \tag{4.64}$$

Proposition 4.37. *Let Assumption 4.3 be satisfied. Then Condition 4.32 holds.*

The key insight in Lemma 4.38 below is that, when working on local coordinate charts, the problem can be transferred to Euclidean space. Solutions can then be found via convex analysis and differential inclusion theory, see e.g. [Dei92, Lemma

5.1]. Transferring back the solution to the manifold leads to locally defined solutions of (4.64). We perform this analysis in Section 4.5.3 below.

As usual, the problem thus resides into patching these curves together to form a curve of arbitrary length. For this, we need to control the time of existence for our local solutions. We do so in multiple steps. Fix some time interval $[0,T]$ for which we aim to construct our curve.

- In Lemma 4.41, we will show that for any $T>0$ and any curve satisfying (4.64) the curve remains in a compact set $\widehat{K}$ up to time T_0. We can thus construct curves locally on sets that have a radius that is lower bounded by the injectivity radius $i(\widehat{K}) = \inf_{x\in\widehat{K}} i_x > 0$.
- Given any such locally constructed curve, we control the Lagrangian linearly in time in Lemma 4.42
- Using this linear Lagrangian growth, we show in Lemma 4.43 that the squared distance to the starting point of the curve grows at most linearly. As the size of the ball is controlled by the injectivity radius on $\widehat{K}$ it follows that there is a lower bound on the interval of existence of the locally constructed curve.

Based on these three steps, we conclude that we can construct solutions to (4.64) on arbitrarily sized intervals $[0,T]$.

Local construction of solutions

In the first result, we show how the various quantities in (4.64) transfer from M to a local coordinate chart. This result is essentially a write-up of basic Riemannian coordinate transformations acting on $\mathcal{L}$ and $\mathcal{H}$. We write it down for an arbitrary smooth invertible map from a subset of a manifold M to a subset of a manifold N.

Lemma 4.38. *Let M be a Riemannian manifold. For an invertible smooth map $\varphi : \mathcal{O} \subseteq M \to \varphi(\mathcal{O}) := N$, via push-forward and pullback in Section 4.2 define*

$$\mathcal{H}_\varphi := \mathcal{H} \circ \varphi^* : T^*N \to \mathbb{R}$$

and

$$\mathcal{L}_\varphi := \mathcal{L} \circ \varphi_*^{-1} : TN \to \mathbb{R},$$

where $\mathcal{H} : T^*M \to \mathbb{R}$ and $\mathcal{L} : TM \to \mathbb{R}$. Define $f_\varphi = f \circ \varphi^{-1}$. Let $x : [0,T] \to \mathcal{O}$, suppose that $y(s) = \varphi(x(s)) : [0,T] \to \varphi(\mathcal{O})$, then we have that

(a) $\mathrm{d}f_\varphi(y(s))\dot{y}(s) = \mathrm{d}f(x(s))\dot{x}(s)$,

(b) $\mathcal{H}_\varphi(y(s), \mathrm{d}f_\varphi(y(s))) = \mathcal{H}(x(s), \mathrm{d}f(x(s)))$,

(c) $\mathcal{L}_\varphi(y(s), \dot{y}(s)) = \mathcal{L}(x(s), \dot{x}(s))$.

(d) *$\mathcal{L}_\varphi$ is the Legendre transform of $\mathcal{H}_\varphi$, i.e., $\mathcal{H}_\varphi(\eta,\xi) = \sup_{w\in T_\eta N} \{\xi(w) - \mathcal{L}_\varphi(\eta, w)\}$, for any $(\eta,\xi) \in T^*N$.*

Proof. We start to prove item (a). By Lemma 4.6, there exists a curve $x(s)$ on $\mathcal{O}$ such that

$$\begin{aligned}\mathrm{d}f_\varphi(y(s))\dot{y}(s) &= \mathrm{d}f_\varphi(\varphi(x(s)))\varphi_*(\dot{x}(s)) \\ &= \mathrm{d}f(x(s))\dot{x}(s),\end{aligned}$$

where in the last show we used the chain rule (4.6) such that

$$\begin{aligned}\mathrm{d}f_\varphi(\varphi(x(s))) &= \mathrm{d}(f\circ\varphi^{-1})(\varphi(x(s))) \\ &= \mathrm{d}f(\varphi^{-1}(\varphi(x(s))))\mathrm{d}(\varphi^{-1}(\varphi(x(s))))\phi_*(\dot{x}(s)) \\ &= \mathrm{d}f(x(s))\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=s}\varphi^{-1}(\varphi(x(t))) \\ &= \mathrm{d}f(x(s))\dot{x}(s).\end{aligned} \tag{4.65}$$

We then prove item (b) based on the ideas when we obtain item (a). By calculating, we have

$$\begin{aligned}\mathcal{H}_\varphi(y(s),\mathrm{d}f_\varphi((y(s)))) &= \mathcal{H}\circ\varphi^*(\varphi(x(s)),\mathrm{d}f_\varphi(\varphi(x(s)))) \\ &= \mathcal{H}(x(s),\varphi^*(\mathrm{d}f_\varphi(\varphi(x(s)))) \\ &= \mathcal{H}(x(s),\mathrm{d}f(x(s))),\end{aligned}$$

where in the last equality we use (4.65). Therefore, item (b) is obtained. We continue to prove item (c) by simple calculating, and get

$$\begin{aligned}\mathcal{L}_\varphi(y(s),\dot{y}(s)) &= \mathcal{L}\circ\varphi_*^{-1}(\varphi(x(s)),\varphi_*(\dot{x}(s))) \\ &= \mathcal{L}(x(s),\dot{x}(s)).\end{aligned}$$

To prove item (d), for any $(\eta,\xi)\in T^*N$, we have

$$\begin{aligned}\mathcal{H}_\varphi(\eta,\xi) &= \mathcal{H}(\varphi^{-1}(\eta),\varphi^*(\xi)) \\ &= \sup_{v\in T_\eta M}\left\{\varphi^*(\xi)(v)-\mathcal{L}(\varphi^{-1}(\eta),v)\right\} \\ &\overset{(4.5)}{=} \sup_{\varphi_*(v)\in T_\eta N}\left\{\xi(\varphi_*(v))-\mathcal{L}_\varphi(\eta,\varphi_*(v))\right\} \\ &= \sup_{w\in T_\eta N}\left\{\xi(w)-\mathcal{L}_\varphi(\eta,w)\right\},\end{aligned}$$

where the second equality is the fact that $\mathcal{L}$ is the Legendre transform of $\mathcal{H}$. The proof is completed. □

Using a transfer of M to a coordinate chart, we can work on Euclidian space. We will construct local solutions using convex analysis and differential inclusion theory. Below, we will use the notion of a subdifferential.

Definition 4.39. For a general convex functional $p \mapsto \Phi(p)$ we denote the subdifferential at $p_0 \in \mathbb{R}^d$ as the set

$$\partial_p \Phi(p_0) := \{\xi \in \mathbb{R}^d : \Phi(p) \geq \Phi(p_0) + \xi(p - p_0), \forall p \in \mathbb{R}^d\}.$$

In the next result, we obtain a local solution to (4.64) by transferring to a chart. We follow the notation of Lemma 4.38

Lemma 4.40. *Let M be a Riemannian manifold and let $x_0 \in M$. Let $\varphi : \mathcal{O} \subseteq M \to \varphi(\mathcal{O}) \subseteq \mathbb{R}^d$ be a coordinate chart. Consider the open ball $\mathcal{O} := B_R(x_0)$ around x_0 with the radius $R > 0$ strictly smaller than the injectivity radius i_{x_0} at x_0. Fix $f \in C^1(M)$. Then the following content holds.*

(a) *There exists a solution $y(t) : [0, T_0(x)) \to \varphi(\mathcal{O}) \subseteq \mathbb{R}^d$ to the differential inclusion*

$$\begin{cases} \dot{y}(t) \in \partial_p \mathcal{H}_\varphi(y(t), \mathrm{d}f_\varphi(y(t)), \\ y(0) = 0 = \varphi(x_0) \end{cases} \tag{4.66}$$

with

$$T_0(x) = \inf \left\{ t > 0 \,\middle|\, y(t) \notin \varphi(B_{R/2}(x_0)) \right\}. \tag{4.67}$$

(b) *Set $x(t) = \varphi^{-1}(y(t))$. Then the curve $x : [0, T_0(x)) \to B_{R/2}(x_0) \subseteq M$ satisfies $x(0) = x_0$ and*

$$\int_0^t \mathcal{H}(x(s), \mathrm{d}f(x(s)))\mathrm{d}s + \int_0^t \mathcal{L}(x(s), \dot{x}(s))\mathrm{d}s = \int_0^t \mathrm{d}f(x(s))\dot{x}(s)\mathrm{d}s. \tag{4.68}$$

for any $t < T_0(x)$.

Proof. We first prove the existence of a solution to the differential inclusion (4.66). By taking $\mathcal{O} = B_R(x_0)$ in Lemma 4.38 and define $T_0(x) = \inf \left\{ t > 0 \,\middle|\, y(t) \notin \varphi(B_{R/2}(x_0)) \right\}$, the subdifferential $\partial_p \mathcal{H}_\varphi(y(t), \mathrm{d}f_\varphi(y(t))$ satisfies all the conditions of Lemma 5.1 of [Dei92]. Note that for this statement, we use that the convexity of $\mathcal{H}$ in p obtained in Lemma 4.14 transfers to $\mathcal{H}_\varphi$. Hence, there exists a solution $y(t)$ such that (4.66) holds.

Next, we turn to prove that there exists a solution such that (4.68) holds by local construction. To do it, for the initial point $x_0 \in M$, there exists a ball $B_R(x_0)$ with R strictly smaller than i_{x_0}. We claim that

$$\int_0^t \mathcal{H}_\varphi(y(s), \mathrm{d}f_\varphi(y(s)))\mathrm{d}s + \int_0^t \mathcal{L}_\varphi(y(s), \dot{y}(s))\mathrm{d}s = \int_0^t \mathrm{d}f_\varphi(y(s))\dot{y}(s)\mathrm{d}s. \tag{4.69}$$

on $\varphi(\mathcal{O})$. Then (4.68) follows from (4.69) and Lemma 4.38.

We are left to prove that (4.69) holds. On the one hand, we have that

$$\mathcal{H}_\varphi(y(s), \mathrm{d}f_\varphi(y(s))) \geq \mathrm{d}f_\varphi(y(s))\dot{y}(s) - \mathcal{L}_\varphi(y(s), \dot{y}(s)),$$

for all $y(s) \in \varphi(\mathcal{O})$, via convex duality. Then, integrating the above inequality gives one inequality in (4.69).

Regarding the other inequality, via (4.66) we obtain for all $p \in \varphi(\mathcal{O})$,

$$\mathcal{H}_\varphi(y(s), p) \geq \mathcal{H}_\varphi(y(s), \mathrm{d}f_\varphi(y(s))) + \dot{y}(s)\left(p - \mathrm{d}f_\varphi(y(s))\right),$$

and as a consequence

$$\mathcal{H}_\varphi(y(s), \mathrm{d}f_\varphi(y(s))) \leq \mathrm{d}f_\varphi(y(s))\dot{y}(s) - \mathcal{L}_\varphi(y(s), \dot{y}(s)),$$

and integrating gives the other inequality. □

Lower bounding the time of existence of local solutions

The first step in lower bounding the time of existence of local solutions is a priori control of any curve satisfying (4.64). The next result follows as a by-product of Condition 4.31.

Lemma 4.41. *Let Assumption 4.3 be satisfied. Let $K_0 \subseteq M$ a compact set and $T > 0$. For any $f \in \mathcal{D}(\mathbf{H})$, then there is a compact set $\hat{K} \subseteq M$ such that any curve $x : [0, T_0) \to M$ with $T_0 \leq T$ satisfying $x(0) \in K_0$ and for all $t < T_0$*

$$\int_0^t \mathcal{H}(x(s), \mathrm{d}f(x(s)))\mathrm{d}s + \int_0^t \mathcal{L}(x(s), \dot{x}(s))\mathrm{d}s = \int_0^t \mathrm{d}f(x(s))\dot{x}(s)\mathrm{d}s \tag{4.70}$$

it holds that $x(t) \in \hat{K}$ for any $t < T_0$.

Recall that by Definition 4.15 of $\mathbf{H}$ we have $\mathcal{D}(\mathbf{H}) = C_c^2(M)$.

Proof. First of all, note that

$$c_f := \sup_z \{-\mathbf{H}f(z)\} < \infty$$

as $\mathcal{H}$ is continuous by Lemma 4.14 and $f \in C_c^2(M)$. Furthermore, write $\|f\| = \sup_x |f(x)|$. Note that by (4.70), we have for any curve

$$\begin{aligned}\int_0^t \mathcal{L}(x(s), \dot{x}(s))\mathrm{d}s &= f(x(t)) - f(x(0)) - \int_0^t \mathcal{H}(x(s), \mathrm{d}f(x(s)))\mathrm{d}s \\ &\leq 2\|f\| + tc_f \leq 2\|f\| + Tc_f\end{aligned}$$

the result thus follows by Condition 4.31 (b). □

For the next three lemmas and the proof of Condition 4.32, we first provide a short sketch of the approach before giving a rigorous proof. The approach involves proof by contradiction. Specifically, we assume that there does not exist a global curve on $[0, T]$ satisfying (4.59). We first find the maximum time interval $[0, T_{\max}]$, $T_{\max} < T$, in which the curve satisfies (4.59). However, by extending the existing curve through patching in a chart at a new point and a lower bound on the time length of the extension, we obtain a new curve that operates over a longer time interval, which leads to a contradiction.

Next, we show the curve as in Lemma 4.40 has a Lagrangian cost that grows linearly in time uniformly in their starting point in a compact set.

Lemma 4.42. *Let M be a Riemannian manifold and $K \subseteq M$ a compact set. Fix $R \in [i(K)/2, i(K))$. Then there is a constant C such that for any curve $x(t) : [0, T_0(x)) \to B_{R/2}(x_0)$ with $x(0) = x_0 \in K$ as in Lemma 4.40, we have*

$$\int_0^t \mathcal{L}(x(s), \dot{x}(s)) \mathrm{d}s \leq Ct$$

for any $t < T_0(x)$.

Proof. First of all, denote by $\hat{K}$ the compact set obtained by covering K by balls of radius $R/2$. No considered curve can leave $\hat{K}$ by construction.

Denote $c_{f,\hat{K}} = \sup_{z \in \hat{K}} \{-\mathbf{H}f(z)\}$. As x satisfies (4.68), by Condition 4.31 (c), there exists a function $\psi_{\hat{K},R}$, R is independent of x, such that for $t < T_0(x)$

$$\begin{aligned} \int_0^t \mathcal{L}(x(s), \dot{x}(s)) \mathrm{d}s &= \int_0^t \mathrm{d}f(x(s))\dot{x}(s) \mathrm{d}s - \int_0^t \mathbf{H}f(x(s)) \mathrm{d}s \\ &\leq \int_0^t \psi_{\hat{K},R} \left(\mathcal{L}(x(s), \dot{x}(s)) \right) \mathrm{d}s + tc_{f,\hat{K}}. \end{aligned}$$

Furthermore, as $\psi_{\hat{K},R}$ is non-decreasing and the fact that $\frac{\psi_{\hat{K},R}(r)}{r}$ converges to 0 for $r \to \infty$, there exist $0 < m < 1$ and $r^* \geq 1$ such that $\frac{\Psi_{\hat{K},R}(r)}{r} \leq m$ for $r \geq r^*$. Proceeding our estimate, by splitting the integral into regions $[0, t] = I_1 \cup I_2$ with

$$\begin{aligned} I_1 &:= \{ s \in [0, t] \mid \mathcal{L}(x(s), \dot{x}(s)) \geq r^* \}, \\ I_2 &:= \{ s \in [0, t] \mid \mathcal{L}(x(s), \dot{x}(s)) < r^* \}, \end{aligned}$$

we get

$$\begin{aligned} \int_0^t \mathcal{L}(x(s), \dot{x}(s)) \mathrm{d}s &\leq \int_{I_1} \frac{\psi_{\hat{K},R}(\mathcal{L}(x(s), \dot{x}(s)))}{\mathcal{L}(x(s), \dot{x}(s))} \mathcal{L}(x(s), \dot{x}(s)) \mathrm{d}s \\ &\quad + \int_{I_2} \psi_{\hat{K},R}(\mathcal{L}(x(s), \dot{x}(s))) \mathrm{d}s + tc_{f,\hat{K}} \\ &\leq m \int_0^t \mathcal{L}(x(s), \dot{x}(s)) \mathrm{d}s + t \left(\psi_{\hat{K},R}(r^*) + c_{f,\hat{K}} \right). \end{aligned}$$

Rearranging terms leads to

$$\int_0^t \mathcal{L}(x(s), \dot{x}(s)) \mathrm{d}s \leq t \frac{\psi_{\hat{K},R}(r^*) + c_{f,\hat{K}}}{1 - m}$$

establishing the claim with $C = \frac{\psi_{\hat{K},R}(r^*)+c_{f,\hat{K}}}{1-m}$. $\square$

Next, we control the speed at which curves as in Lemma 4.40 move away from their starting point.

Lemma 4.43. *Let M be a Riemannian manifold and $K \subseteq M$ a compact set. Fix $R \in [i(K)/2, i(K))$.*

Then there is a $C > 0$ such that for any $x_0 \in K$ and any curve $x(t) : [0, T_0(x)) \to B_{R/2}(x_0)$ with $x(0) = x_0$ as in Lemma 4.40, we have

$$\frac{1}{2} d^2(x(t), x_0) \leq tC$$

for any $t < T_0(x)$. In particular $T_0(x) \geq \frac{R^2}{8C}$.

The proof of Lemma 4.43 relies on Lemma 4.40 and the following preliminary lemma. We first state the preliminary lemma before proceeding to prove Lemma 4.43.

Lemma 4.44. *Let $K \subseteq M$ be a compact set in M. For any $x_0 \in K$ and radius $R < i_{x_0}$, set*

$$g_{x_0,R}(x) = \theta_R\left(\frac{1}{2}d^2(x, x_0)\right)$$

where $\theta_R : [0, \infty) \to [0, \frac{3}{4}R]$ is a smooth non-decreasing function, satisfying $\theta_R'(r) \leq 1$ where $\theta_R(r) = r$ for $r \leq R/2$ and $\theta_R(r)$ is constant for $r \geq \frac{3}{4}R$.

For any such R, we have $g_{x_0,R} \in C_{K,R}$ where $C_{K,R}$ was defined in Condition 4.31 (c) *equation* (4.58)*. Moreover, $g_{x_0,R} \in \mathcal{D}(\mathbf{H})$.*

Proof. By construction, we have

$$\mathrm{d}g_{x_0,R}(x) = \theta_R'\left(\frac{1}{2}d^2(x, x_0)\right) d(x, x_0)$$

which by the properties of θ_R satisfies

$$|\mathrm{d}g_{x_0,R}(x)| \leq d(x, x_0) \leq R$$

for any $x \in B(x_0, R)$. In particular, we have $g_{x_0,R} \in C_{K,R}$. Moreover, since $g_{x_0,R}$ is twice continuously differentiable and constant outside of a compact set, we conclude that $g_{x_0,R} \in \mathcal{D}(\mathbf{H})$. $\square$

Proof of Lemma 4.43. Fix $x_0 \in K$ and any curve $x(t) : [0, T_0(x)) \to B_{R/2}(x_0)$ with $x(0) = x_0$ as in Lemma 4.40. Let $g_{x_0,R} \in \mathcal{D}(\mathbf{H})$ be any smooth bounded function as in Lemma 4.44 and $\hat{K}$ be the compact set obtained by covering K by balls of radius $R/2$.

It thus follows by the proof strategy of Lemma 4.42 that for any $t < T_0(x)$, we have

$$\frac{1}{2}d^2(x(t), x_0) \leq t\left(mC_1 + \psi_{\hat{K},R}(r^*)\right).$$

The result thus follows for $C = mC_1 + \psi_{\hat{K},R}(r^*)$. □

We are ready to verify Condition 4.32.

Proof of Condition 4.32 . We argue by contradiction. Fix $x_0 \in M$ and $T > 0$. Suppose there does not exist an absolutely continuous curve $x(t)$, $t \in [0, T]$ started at $x_0 \in M$ such that (4.59) holds.

In other words,

$$T_{\max} = \sup\left\{T_0(x) \,|\, \exists\, x : [0, T_0(x)) \to M \text{ satisfying } (4.59), x(0) = x_0\right\} < T. \quad (4.71)$$

By Lemma 4.41 there is a compact set $K \subseteq M$ such that any curve considered in (4.71) stays in K. Fix $\varepsilon < \frac{R^2}{8C} \leq T_0(x)$ as in Lemma 4.43.

Fix the curve x satisfying (4.59) with $x(0) = x_0$, $T_0(x) > T_{\max} - \varepsilon$. Patching the curve $\tilde{x} : [0, T_0(\tilde{x}))$ started from $x(T_0(x) - \varepsilon)$ obtained from Lemma 4.40 to the curve x at time $T_0(x) - \varepsilon$, we obtain from Lemma 4.43 that this curve, is a solution to (4.59) on the time interval $[0, T_0(x) - \varepsilon + T_0(\tilde{x}))$, which contradicts (4.71).

This establishes the claim. □

Part II

Parameter estimations

Chapter 5

Parameter estimations for singular SDEs

From this chapter, we turn to study parameter estimations. In this chapter, by taking Zvonkin's transformation, we investigate parameter estimation for a class of multi-dimensional stochastic differential equations with small perturbation parameters in diffusion coefficients, where the drift coefficients not only have an unknown parameter θ but also are Hölder continuous. These processes may enhance the applicability of our results to considerable practical models. Due to the irregular drift, the primary challenge is dealing with the mean square error between an accurate and numerical solution. Under these settings, we demonstrate the consistency and asymptotic normality of error concerning the least squares estimator in probability when stepsize $\delta \to 0$ and small parameter $\varepsilon \to 0$ simultaneously. Moreover, we extend the results to the case of stochastic functional differential equations.

This chapter is based on

[HXZ24] Yanyan Hu, Fubao Xi, and Min Zhu. Asymptotic properties for the parameter estimation instochastic (functional) differential equations with Hölder drift. *Stochastics*, 96(1):766–798, 2024

5.1 Introduction

The theory of parameter estimations for stochastic differential equations (SDEs) is one of the active research fields. The background, motivation, applications, and fundamental results of the theory are well established; many methods have been put forward to solve the parameter estimation problem of diffusion processes, such as maximum likelihood estimation, Bayes estimation, and least squares estimation (LSE) based on continuous or discrete observations.

To the best of our knowledge, there are numerous results on parameter estimation for SDEs with regular drift under various settings. [WS16] studied maximum likelihood estimation for drift parameters in diffusion processes. As for more complex processes, we can refer to [WWMX16] and [Lon09], who investigated the maximum likelihood estimation of McKean-Vlasov SDEs and studied the parameter estimation problem for one-dimensional Ornstein-Uhlenbeck processes with small Lévy noise, respectively. In particular, for considering a high-frequency sample of discrete observations of the diffusion processes at time points, the related works are crucial for parameter estimation; see, e.g., [FZ89, Yos92, Kes97, Lon09, DGCL18, AHPP23] and references therein.

However, the parameter estimation problem with irregular drift has not been well studied yet. This is one of our motivations. Moreover, a lot of authors have paid attention to many types of SDEs with irregular drifts, such as Hölder continuity, Hölder-Dini continuity, and even only integrability (e.g., [GM01, KR05, Zha05, Zha11, Zha16] and references within) over the past few years. Specifically, [Wan16] showed gradient estimations and applications for SDEs in Hilbert space with multiplicative noise and Dini continuous drift via Zvonkin's transformation. In addition, the problem of convergence rate can also be solved by this transformation. For instance, [BHY19] showed the convergence rate of the numerical solution and accurate solution for SDEs with Hölder-Dini continuous drift. Subsequently, this method has raised considerable attention, and more realistic models are emerging. For example, [Hua19] applied this transformation to derive exponential convergence for functional SDEs with Hölder continuous drift, [HY21a] further extended to the case of α-stable process.

At the same time, there has been some development in the setting of diffusion coefficients for parameter estimation problems. Long, Shimizu, and Sun [LSS13] studied the problem of parameter estimation for discretely observed stochastic processes driven by additive small Lévy noises, namely $\sigma = 1$. Although they gave a brief remark that their methodology can be easily extended to the more general case of semi-martingale noises. However, the diffusion coefficients discussed in that paper still are additive, which restricts the applicability of their models. In a similar framework, Long, Ma, and Shimizu [LMS17] further considered parameter estimation for discretely observed SDEs driven by small Lévy noises, where σ is a linear multiplicative and they did not impose Lipschitz condition on σ. After that, Ren and Wu [RW19b, RW21] studied the parameter estimation for Mckean-Vlasov SDEs, where the diffusion coefficient is nonlinear.

Motivated by the previous literature about SDEs with irregular drift coefficients and more general diffusion coefficients, in the present work, we would like to explore the parameter estimate of multidimensional SDEs via the least square method. We allow that the diffusion σ is nonlinear and the drift b is bounded and Hölder continuous. The least square method involves solving SDEs and constructing contrast functions, and often it is difficult to deal with them. For this reason, we approximate the solutions of SDEs considered by selecting an appropriate numerical scheme and designing the corresponding contrast functions based on the approximated equations. In this work, we present the least square method for SDEs/SFDEs, which is one of the well-known methods in the theory of parameter estimations for SDEs/SFDEs that is used to prove the consistency and asymptotic normality of the estimator. In the case of singular drifts, the techniques used in existing literature [Lon09] can not be applied, because it uses the analysis of standard SDEs under Lipschitz conditions. The techniques applied in this work use in particular the regularity of the non-degenerate Kolmogorov equation and some basic convergence principles from functional analysis. An important result is that the appropriate numerical approximations converge in the mean square to the accurate solutions of SDEs/SFDEs considered (see Theorems 5.16 and 5.21). Moreover, since the approximation method also involves SFDEs, by using the ideas from [BS18] and adapting them for the considered SFDEs we establish the approximated equation of (5.19) below and its contrast function, and deduce the square error between accurate solutions and numerical solutions, which gives us a necessary condition for the consistency and asymptotic normality.

5.2 Preliminaries and main results

Throughout this chapter, the following notation will be used. For $d, m, p \in \mathbb{N}$, the set of all positive integers, let $(\mathbb{R}^d, \langle\cdot,\cdot\rangle, |\cdot|)$ be the d-dimensional Euclidean space with the inner product $\langle\cdot,\cdot\rangle$ inducing the norm $|\cdot|$. Let $\mathbb{R}^d \otimes \mathbb{R}^m$ the collection of all $d \times m$ matrixes with real entries, which is endowed with the Hilbert-Schmidt norm $\|\cdot\|_{\mathrm{HS}}$. For $A \in \mathbb{R}^d \otimes \mathbb{R}^m$, A^* denotes the transpose of A. Concerning a square matrix A, A^{-1} means the inverse of A provided that $\det A \neq 0$. For $p \in \mathbb{N}$, let Θ be an open bounded convex subset of $\mathbb{R}^p$, and $\overline{\Theta}$ is the closure of Θ. For $r > 0$ and $x \in \mathbb{R}^p$, $B_r(x)$ means the closed ball centered at x with the radius r. Let $\mathscr{B}_b(\mathbb{R}^d)$ be the collection of all bounded measurable functions $f : \mathbb{R}^d \to \mathbb{R}$, endowed with the uniform norm $\|f\|_\infty := \sup_{x\in\mathbb{R}^d} |f(x)|$. For a real number $a > 0$, $\lfloor a \rfloor$ stands for the integer part of a. Let $\tau > 0$ be a fixed number and $\mathscr{C} = C([-\tau, 0]; \mathbb{R}^d)$, which is endowed with the uniform norm $\|f\|_{\mathscr{C}} = \sup_{-\tau\leq\gamma\leq 0} |f(\gamma)|$, $\gamma \in [-\tau, 0]$. For $f \in C([-\tau, \infty); \mathbb{R}^d)$ and $t \geq 0$, let $f_t \in \mathscr{C}$ be defined by $f_t(\gamma) = f(t + \gamma)$, $\gamma \in [-\tau, 0]$. In terminology, $(f_t)_{t>0}$ is called the segment (or window) process corresponding to $(f(t))_{t\geq-\tau}$. For $A =: (A_1, A_2, \ldots, A_p) \in \mathbb{R}^p \otimes \mathbb{R}^{pd}$ with $A_k \in \mathbb{R}^p \otimes \mathbb{R}^d$, $k = 1, \ldots, p$, and $B \in \mathbb{R}^d$, let us define $A \circ B \in \mathbb{R}^p \otimes \mathbb{R}^p$ by

$$A \circ B = (A_1 B, A_2 B, \cdots, A_p B), \tag{5.1}$$

and more details about this symbol can be found in the appendix.

5.2.1 LSE for SDEs

In this subsection, we fix the time horizon $T > 0$. For the scale parameter $\varepsilon \in (0, 1)$, we are interested in the following SDE

$$\mathrm{d}X^{\varepsilon}(t) = b(X^{\varepsilon}(t), \theta)\mathrm{d}t + \varepsilon\, \sigma(X^{\varepsilon}(t))\mathrm{d}W(t), \quad t \in [0, T], \qquad X^{\varepsilon}(0) = x_0 \in \mathbb{R}^d, \quad (5.2)$$

where $b : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$, $\sigma : \mathbb{R}^d \to \mathbb{R}^d \otimes \mathbb{R}^d$ and $(W(t))_{t\geq 0}$ is a d-dimensional Brownian motion defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with the filtration $(\mathscr{F}_t)_{t\geq 0}$ satisfying the usual condition (i.e., $\mathscr{F}_0$ contains all $\mathbb{P}$-null sets and $\mathscr{F}_t = \mathscr{F}_{t+} := \bigcap_{s>t} \mathscr{F}_s$). In what follows, denote the initial value of (5.2) by $X^{\varepsilon}(0)$ when we emphasize the dependence on ε, and ε is a small perturbation. In (5.2), we assume that the drift b and the diffusion σ are known apart from the parameter $\theta \in \Theta$.

Intuitively, we derive the underlying deterministic ordinary differential equation under the true value θ_0 of the drift parameter corresponding to $\varepsilon = 0$ in (5.2),

$$\mathrm{d}X^0(t) = b(X^0(t), \theta_0)\mathrm{d}t, \quad t \in [0, T], \quad X^0(0) = x_0 \in \mathbb{R}^d. \qquad (5.3)$$

Herein, it is worth pointing out that (5.2) and (5.3) share the same initial value.
Definition 5.1 (Strong solution)**.** For any $T \geq 0$, a continuous adapted process $(X(t))_{t\in[0,T]}$ on $\mathbb{R}^d$ is called a strong solution of (5.2), if

$$\int_0^t \mathbb{E}(|b(X(s), \theta)| + \|\sigma(X(s))\|_{\mathrm{HS}}^2)\mathrm{d}s < \infty, \qquad t \in [0, T],$$

and $\mathbb{P}$-a.s.

$$X(t) = X(0) + \int_0^t b(X(s), \theta)\mathrm{d}s + \int_0^t \sigma(X(s))\mathrm{d}W(s), \qquad t \in [0, T].$$

Throughout the chapter, for any $x, y \in \mathbb{R}^d$ and $\theta \in \overline{\Theta}$, we assume the following conditions
Assumption 5.1. *b is bounded and there exists a constant $C > 0$ and $\alpha \in (0, 1]$ such that*

$$\sup_{\theta\in\overline{\Theta}} |b(x, \theta) - b(y, \theta)| \leq C|x - y|^{\alpha}.$$

Assumption 5.2. *σ is invertible, and*

$$\|\sigma\|_{\mathrm{HS}} + \|\nabla\sigma\|_{\mathrm{HS}} + \|\sigma^{-1}\|_{\mathrm{HS}} + \|\nabla\sigma^{-1}\|_{\mathrm{HS}} < \infty.$$

Assumption 5.3. *For $j = 1, 2$, there exists a constant $C > 0$ such that*

$$\sup_{\theta\in\overline{\Theta}} \|(\nabla_{\theta}^{j} b)(x, \theta) - (\nabla_{\theta}^{j} b)(y, \theta)\|_{\mathrm{HS}} \leq C|x - y|^{\alpha}.$$

Remark 5.2. Under Assumption 5.1 and $\|\sigma\|_{\mathrm{HS}} < \infty$ in Assumption 5.2, (5.2) enjoys a unique strong solution $(X^\varepsilon(t))_{t\in[0,T]}$; see, e.g., [BH22, Theorem 1.4]. Because we want to study the consistency and asymptotic normality of the least squares estimator, we add Assumption 5.3 and $\|\nabla\sigma\|_{\mathrm{HS}} + \|\sigma^{-1}\|_{\mathrm{HS}} + \|\nabla\sigma^{-1}\|_{\mathrm{HS}} < \infty$ in Assumption 5.2. Likewise, under Assumption 5.1, (5.3) admits a unique solution $(X^0(t))_{t\in[0,T]}$. In addition, we can also generalize in this thesis that b is locally bounded rather than uniformly bounded. However, when b satisfies the unbounded condition, this is a challenge that we will investigate in future papers.

Without loss of generality, one may assume that there exists a sufficiently large integer $n > 0$ such that the stepsize

$$\delta := \frac{T}{n} \in (0, 1).$$

Now, for $k = 1, 2, \ldots, n$, we introduce the following Euler–Maruyama (EM) scheme

$$Y^\varepsilon((k+1)\delta) = Y^\varepsilon(k\delta) + b(Y^\varepsilon(k\delta), \theta)\delta + \varepsilon\sigma(Y^\varepsilon(k\delta))\Delta W_k, \tag{5.4}$$

where $\Delta W_k := W((k+1)\delta) - W(k\delta)$. The continuous EM scheme reads

$$\mathrm{d}Y^\varepsilon(t) = b(Y^\varepsilon(t_\delta), \theta)\mathrm{d}t + \varepsilon\sigma(Y^\varepsilon(t_\delta))\mathrm{d}W(t), \qquad t \in [0, T], \tag{5.5}$$

with the same initial value $Y^\varepsilon(0) = X^\varepsilon(0) = x_0 \in \mathbb{R}^d$, where $t_\delta = \lfloor t/\delta \rfloor \delta$. Then, by (5.5) and the theory of least squares method, we design the following contrast function

$$\Psi_{n,\varepsilon}(\theta) = \varepsilon^{-2}\delta^{-1} \sum_{k=1}^{n} F_k^*(\theta)\hat{\sigma}(Y^\varepsilon((k-1)\delta)F_k(\theta). \tag{5.6}$$

Herein, for $k = 1, 2, \ldots, n$,

$$F_k(\theta) := Y^\varepsilon(k\delta) - Y^\varepsilon((k-1)\delta) - b(Y^\varepsilon((k-1)\delta), \theta)\delta, \tag{5.7}$$

$$\hat{\sigma}(Y^\varepsilon(k\delta)) := (\sigma\sigma^*)^{-1}(Y^\varepsilon(k\delta)). \tag{5.8}$$

To achieve the LSE of $\theta \in \overline{\Theta}$, it suffices to choose an element $\hat{\theta}_{n,\varepsilon} \in \overline{\Theta}$ such that

$$\Psi_{n,\varepsilon}(\hat{\theta}_{n,\varepsilon}) = \min_{\theta\in\Theta} \Psi_{n,\varepsilon}(\theta). \tag{5.9}$$

Whence,

$$\hat{\theta}_{n,\varepsilon} = \arg\min_{\theta\in\Theta} \Psi_{n,\varepsilon}(\theta).$$

Set

$$\Phi_{n,\varepsilon}(\theta) := \varepsilon^2(\Psi_{n,\varepsilon}(\theta) - \Psi_{n,\varepsilon}(\theta_0)). \tag{5.10}$$

It follows from (5.9) that

$$\Phi_{n,\varepsilon}(\hat{\theta}_{n,\varepsilon}) = \min_{\theta\in\Theta} \Phi_{n,\varepsilon}(\theta). \tag{5.11}$$

Likewise, we reformulate $\hat{\theta}_{n,\varepsilon} \in \Theta$ ensuring (5.11) to hold true as

$$\hat{\theta}_{n,\varepsilon} = \arg\min_{\theta\in\Theta} \Phi_{n,\varepsilon}(\theta). \tag{5.12}$$

In this work, $\hat{\theta}_{n,\varepsilon}$ satisfing (5.12) is named as the LSE of $\theta \in \Theta$.

Remark 5.3. If $b(\cdot,\theta)$ is explicit concerning the parameter θ, then the least squares estimator $\hat{\theta}_{n,\varepsilon}$ can indeed be obtained by Fermat's theorem; see Example 5.7 below for more details.

For the sake of notation brevity, for any $x \in \mathbb{R}^d$ and $\theta \in \Theta$, let

$$G(x,\theta,\theta_0) = b(x,\theta_0) - b(x,\theta). \tag{5.13}$$

One of the first main results is concerned with the consistency of the LSE of $\theta \in \Theta$.

Theorem 5.4 (Consistency)**.** *Let Assumptions 5.1 and 5.2 hold and assume further $H(\theta) > 0$ for any $\theta \neq \theta_0 \in \overline{\Theta}$. Then we have*

$$\hat{\theta}_{n,\varepsilon} \to \theta_0 \qquad \text{in probability as } \varepsilon \to 0 \quad \text{and} \quad n \to \infty.$$

Set, for any $x \in \mathbb{R}^d$ and $\theta \in \overline{\Theta}$,

$$I(\theta) := \int_0^T (\nabla_\theta b)^*(X^0(t),\theta)\hat{\sigma}(X^0(t))(\nabla_\theta b)(X^0(t),\theta)\mathrm{d}t, \tag{5.14}$$

$$\Upsilon(x,\theta_0) := (\nabla_\theta b)^*(x,\theta_0)\hat{\sigma}(x)\sigma(x) \tag{5.15}$$

and

$$K(\theta) =: -2\int_0^T \{(\nabla_\theta^{(2)} b^*)(X^0(t),\theta) \circ (\hat{\sigma}(X^0(t))G(X^0(t),\theta,\theta_0))\}\mathrm{d}t, \tag{5.16}$$

where the notation $(\nabla_\theta^{(2)} b^*) := (\nabla_\theta(\nabla_\theta b^*))$ and "$\circ$" is defined in (5.1).

The second result is presented below, revealing the asymptotic normality property of $\hat{\theta}_{n,\varepsilon}$.

Theorem 5.5 (Asymptotic normality)**.** *Let the assumptions of Theorem 5.4 hold, suppose further that Assumption 5.3 hold. If $I(\cdot)$ and $K(\cdot)$ are continuous. Then we have*

$$\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0) \to I^{-1}(\theta_0)\int_0^T \Upsilon(X^0(t),\theta_0)\mathrm{d}W(t) \qquad \text{in probability}$$

as $\varepsilon \to 0$ and $n \to \infty$.

Remark 5.6. Although we handle the least square estimator on the Hölder drift condition, there is no loss of efficiency to the consistency and asymptotic normality of the estimator.

Here, we provide an example of real value SDE, which can further intuitively explain the results of Theorems 5.4 and 5.5.

Example 5.7. Let $\theta \in (c_1, c_2)$ for some constants c_1 and c_2 with $c_1 < c_2$. Assume that

$$\mathbb{P}(|X^\varepsilon(t)| < \infty) = 1. \tag{5.17}$$

For any $\varepsilon \in (0,1)$, we consider the following real value SDE,

$$\mathrm{d}X^{\varepsilon}(t) = [(1+\theta)|X^{\varepsilon}(t)|]^{1/2}\mathrm{d}t + \varepsilon(1+|X^{\varepsilon}(t)|)\mathrm{d}W(t), \quad t \in (0,T] \tag{5.18}$$

with initial datum $X^{\varepsilon}(0) = x_0$, where θ is an unknown parameter.

Obviously, for any $x \in \mathbb{R}$, form (5.18) we have

$$b(\theta,x) = ((1+\theta)|x|)^{\frac{1}{2}}, \quad \sigma(x) = 1+|x|.$$

Namely, (5.18) can be reformulated as (5.2).

Next, we aim to examine whether all the assumptions imposed in Theorems 5.4 and 5.5 hold. Indeed, b is bounded via (5.17) and $\theta \in (c_1, c_2)$, and by a direct calculation, for any x, $y \in \mathbb{R}$, there exists a constant $C > 0$ such that

$$\begin{aligned}|b(\theta,x) - b(\theta,y)| &= (1+\theta)^{\frac{1}{2}}(|x|^{\frac{1}{2}} - |y|^{\frac{1}{2}}) \\ &\leq (1+\theta)^{\frac{1}{2}}|x-y|^{\frac{1}{2}} \leq C|x-y|^{\frac{1}{2}}.\end{aligned}$$

The Assumption 5.1 is checked. After that, the real diffusion coefficient σ is invertible and there exists a constant $C > 0$ such that $|\sigma| \leq C$ by (5.17), and we have

$$|\sigma^{-1}| \leq 1, \quad |\nabla\sigma| \leq 1 \ \text{ and } \ |\nabla\sigma^{-1}| \leq 1.$$

Hence, the Assumption 5.2 holds.

We are left to prove that the Assumption 5.3 holds. To do it, since

$$\nabla_{\theta}b(\theta,x) = \frac{1}{2}((1+\theta)|x|)^{-\frac{1}{2}}|x| = \frac{1}{2\sqrt{1+\theta}}|x|^{\frac{1}{2}},$$

one has

$$\begin{aligned}&|\nabla_{\theta}b(\theta,x) - \nabla_{\theta}b(\theta,y)| \\ &\quad = \frac{1}{2}(1+\theta)^{-\frac{1}{2}}(|x|^{\frac{1}{2}} - |y|^{\frac{1}{2}}) \leq \frac{1}{2}(1+\theta)^{-\frac{1}{2}}|x-y|^{\frac{1}{2}} \leq \frac{1}{2}|x-y|^{\frac{1}{2}}.\end{aligned}$$

Similarly,

$$\nabla_{\theta}(\nabla_{\theta}b(\theta,x)) = -\frac{1}{4}(1+\theta)^{-\frac{3}{2}}|x|^{\frac{1}{2}},$$

then

$$|\nabla_{\theta}(\nabla_{\theta}b(\theta,x)) - \nabla_{\theta}(\nabla_{\theta}b(\theta,y))| = \left|-\frac{1}{4}(1+\theta)^{-\frac{3}{2}}(|x|^{\frac{1}{2}} - |y|^{\frac{1}{2}})\right| \leq |x-y|^{\frac{1}{2}}.$$

Thus, Assumption 5.3 holds.

Then, we continue to provide the contrast function below, which plays an important role in obtaining the consistency and asymptotic normality from Theorems 5.13 and 5.4. Based on (5.6) with

$$F_k(\theta) = Y^{\varepsilon}(k\delta) - Y^{\varepsilon}((k-1)\delta)) - ((1+\theta)|Y^{\varepsilon}((k-1)\delta))|)^{\frac{1}{2}}\delta$$

and

$$\hat{\sigma}(Y^{\varepsilon}(k\delta)) = \frac{1}{(1+|Y^{\varepsilon}((k-1)\delta))|)^2},$$

the contrast function admits the form

$$\Phi_{n,\varepsilon}(\theta) = \varepsilon^{-2}\delta^{-1}\sum_{k=1}^{n}\frac{\left|Y^{\varepsilon}(k\delta) - Y^{\varepsilon}((k-1)\delta)) - ((1+\theta)|Y^{\varepsilon}((k-1)\delta))|)^{\frac{1}{2}}\delta\right|^2}{(1+|Y^{\varepsilon}((k-1)\delta))|)^2}.$$

Noting that the derivative of $\Phi_{n,\varepsilon}(\theta)$ is

$$\frac{\mathrm{d}\Phi_{n,\varepsilon}(\theta)}{\mathrm{d}\theta} = -\frac{1}{\varepsilon^2\sqrt{1+\theta}}\sum_{k=1}^{n}\frac{\left|Y^{\varepsilon}(k\delta) - Y^{\varepsilon}((k-1)\delta)) - ((1+\theta)|Y^{\varepsilon}((k-1)\delta)|)^{\frac{1}{2}}\delta\right|}{(1+|Y^{\varepsilon}((k-1)\delta))|)^2|Y^{\varepsilon}((k-1)\delta))|^{-\frac{1}{2}}},$$

and when

$$\frac{\mathrm{d}\Phi_{n,\varepsilon}(\theta)}{\mathrm{d}\theta} = 0,$$

one obtains

$$\hat{\theta}_{n,\varepsilon} = \frac{(|Y^{\varepsilon}(k\delta)| - |Y^{\varepsilon}((k-1)\delta))|)^2}{\delta^2|Y^{\varepsilon}((k-1)\delta))|} - 1.$$

of the unknown parameter θ. In terms of Theorem 5.4, $\hat{\theta}_{n,\varepsilon} \to \theta$ in probability as $\varepsilon \to 0$ and $n \to \infty$. Next, from (5.14) and (5.15), it follows that

$$I(\theta_0) = \frac{1}{4(1+\theta_0)}\int_0^T \frac{|X^0(t)|}{1+|X^0(t)|}\mathrm{d}t$$

and

$$\int_0^T \Upsilon(X^0(t),\theta_0)\mathrm{d}W(t) = \frac{1}{4(1+\theta_0)}\int_0^T \frac{|X^0(t)|}{1+|X^0(t)|}\mathrm{d}W(t).$$

At last, according to Theorem 5.5, we conclude that

$$\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0) \to I^{-1}(\theta_0)\int_0^T \Upsilon(X^0(t),\theta_0)\mathrm{d}W(t) \qquad \text{in probability}$$

as $\varepsilon \to 0$ and $n \to \infty$.

5.2.2 LSE for SFDEs

In this subsection, we discuss the extension of Theorems 5.4 and 5.5 to the general model when the drift contains delay term. In many applications, the future state of stochastic systems is not only dependent on the present but also on history. So we are interested in more realistic processes called SFDEs, namely,

$$\mathrm{d}X^{\varepsilon}(t) = b(X^{\varepsilon}(t),\theta)\mathrm{d}t + Z(X_t^{\varepsilon})\mathrm{d}t + \varepsilon\sigma(X^{\varepsilon}(t))\mathrm{d}W(t), \quad t\in[0,T], \quad X_0^{\varepsilon} = \xi \in \mathscr{C}, \tag{5.19}$$

where $Z : \mathscr{C} \to \mathbb{R}^d$. (5.19) can be regarded as the perturbation of the SDE (5.2). The perturbation function Z is dependent on the history state and describes a delayed feedback loop that has a weak impact on the unperturbed dynamics.

In order to prove the existence and uniqueness of a solution to (5.19), we shall add the following assumption about the delay drift.

Assumption 5.4. *Z is bounded and there exists a constant $C > 0$ such that*

$$|Z(\xi) - Z(\eta)| \leq C\|\xi - \eta\|_{\mathscr{C}}, \quad \xi,\ \eta \in \mathscr{C}.$$

Remark 5.8. The singularity of our drift is only reflected in a part of the coefficients, namely b, which depend on the present state. This happens because the regularity of the non-degenerate Kolmogorov equation related to Zvonkin's transformation depends on the present state.

Remark 5.9. Since Hölder continuity is stronger than Dini continuity, according to [HZ19, Theorem 2.1] for $\mathbb{H} = \mathbb{R}^d$, under Assumptions 5.1, 5.2 and 5.4, the SFDE (5.19) has a unique non-explosive solution denoted by X_t^ε with $X_0^\varepsilon = \xi$.

When $\varepsilon = 0$, it is obvious that (5.19) can be rewritten as

$$\mathrm{d}X^0(t) = b(X^0(t), \theta)\mathrm{d}t + Z(X_t^0)\mathrm{d}t.$$

For $k = 1, 2, \ldots, n$, the discrete-time EM scheme of (5.19) is defined as

$$Y^\varepsilon((k+1)\delta) = Y^\varepsilon(k\delta) + b(Y^\varepsilon(k\delta), \theta)\delta + Z(\hat{Y}_{k\delta}^\varepsilon)\delta + \varepsilon\sigma(Y^\varepsilon(k\delta))\Delta W_k \tag{5.20}$$

with same value $Y^\varepsilon(0) = X_0^\varepsilon = \xi$, and $\hat{Y}_t^\varepsilon \in \mathscr{C}$ is defined in the way

$$\hat{Y}_t^\varepsilon(\gamma) := Y^\varepsilon((t+\gamma) \wedge t_\delta), \quad \gamma \in [-\tau, 0].$$

The truncated EM scheme associated (5.19) is described as

$$\mathrm{d}Y^\varepsilon(t) = b(Y^\varepsilon(t_\delta), \theta)\mathrm{d}t + Z(\hat{Y}_t^\varepsilon)\mathrm{d}t + \varepsilon\sigma(Y^\varepsilon(t_\delta))\mathrm{d}W(t), \qquad t \in [0, T], \tag{5.21}$$

with the same initial value $Y^\varepsilon(\gamma) = X^\varepsilon(\gamma) = \xi(\gamma)$, $\gamma \in [-\tau, 0]$. It is obvious that one obtains the relations

$$\|Y_t^\varepsilon\|_{\mathscr{C}} = \sup_{-\tau \leq \gamma \leq 0} |Y^\varepsilon(t+\gamma)| = \sup_{(t-\tau)^+ \leq s \leq t} |Y^\varepsilon(s)| \tag{5.22}$$

and

$$\|\hat{Y}_t^\varepsilon\|_{\mathscr{C}} = \sup_{-\tau \leq \gamma \leq 0} |Y^\varepsilon((t+\gamma) \wedge t_\delta)| \leq \|Y_t^\varepsilon\|_{\mathscr{C}}. \tag{5.23}$$

Then, on the basis of (5.20), we design the following contrast function

$$\Psi_{n,\varepsilon}^Z(\theta) = \varepsilon^{-2}\delta^{-1}\sum_{k=1}^n (F_k^Z)^*(\theta)\hat{\sigma}(Y_{(k-1)\delta}^\varepsilon)F_k^Z(\theta). \tag{5.24}$$

Herein, for $k = 1, 2, \dots, n$,

$$F_k^Z(\theta) := Y^\varepsilon(k\delta) - Y^\varepsilon((k-1)\delta) - b(Y^\varepsilon((k-1)\delta), \theta)\delta - Z(\hat{Y}^\varepsilon_{(k-1)\delta})\delta$$

and $\hat{\sigma}(\cdot)$ is defined in (5.8). Thus, we can get the least squares estimator $\hat{\theta}^Z_{n,\varepsilon}$ by (5.24) in the same way as in Subsection 5.2.1.

Remark 5.10. The truncated EM method has recently aroused widespread concern; see, e.g., [NBK+20, BS18]. For SFDEs, the simple discrete-time observations are insufficient to build the contrast function because the SFDEs involved are path-dependent. Linear interpolation is a traditional method of approximating the functional solution, and in the thesis, we use truncated EM. Compared with linear interpolation, the advantage of truncation is that no additional continuity conditions need to be imposed on the initial value.

The main result of LSEs about SFDEs is stated below.

Theorem 5.11. *Under Assumptions 5.1, 5.2, 5.3 and 5.4, for any $\theta \neq \theta_0 \in \overline{\Theta}$, it yields that*

$$\hat{\theta}^Z_{n,\varepsilon} \to \theta_0 \qquad \text{in probability}$$

as $\varepsilon \to 0$ and $n \to \infty$. Moreover, for $I(\cdot)$ in (5.14)*, $K(\cdot)$ in* (5.16) *and $\Upsilon(\cdot, \theta_0)$ in* (5.15)*, if $I(\cdot)$ and $K(\cdot)$ are continuous, we have that*

$$\varepsilon^{-1}(\hat{\theta}^Z_{n,\varepsilon} - \theta_0) \to I^{-1} \int_0^T \Upsilon(X^0(t), \theta_0)\mathrm{d}W(t) \qquad \text{in probability}$$

as $\varepsilon \to 0$ and $n \to \infty$.

Remark 5.12. Last but not least, from the results of Theorem 5.11, the delay part is reflected in the least square estimator $\hat{\theta}^Z_{n,\varepsilon}$. That is,

$$\hat{\theta}^Z_{n,\varepsilon} = \arg\min_{\theta \in \Theta} \Psi^Z_{n,\varepsilon}(\theta),$$

which means that

$$\Psi^Z_{n,\varepsilon}(\hat{\theta}^Z_{n,\varepsilon}) = \min_{\theta \in \Theta} \Psi^Z_{n,\varepsilon}(\theta),$$

where $\Psi^Z_{n,\varepsilon}(\theta)$ is the contrast function defined in (5.24). Therefore, we see that the delay term Z essentially originates from the construction of the contrast function.

5.3 Proof of main results

In this section, we complete the proof of the main theorems in the above section. First, we prove Theorems 5.4 and 5.5, which are the main results in Subsection 5.2.1 about SDEs.

5.3.1 Proof of Theorems 5.4 and 5.5

Before presenting the proof of Theorems 5.4 and 5.5, it is necessary that we prove Lemmas 5.13 - 5.19 below. We first give Lemma 5.13, which shows finite pth moment of the solution $Y^\varepsilon(\cdot)$.

Lemma 5.13. *Under Assumptions 5.1 and 5.2 for $p \geq 2$, there exists a constant $C_T > 0$ such that*

$$\sup_{0<t\leq T} \mathbb{E}|Y^\varepsilon(t)|^p \leq C_T(1+|x_0|^p). \tag{5.25}$$

Proof. we only need to apply the fundamental inequality, the Hölder inequality for integrals for time, and the Burkhold-Davis-Gundy (BDG) inequality for the term involving martingale. For $C > 0$, it holds that

$$\begin{aligned}
1+&\mathbb{E}|Y^\varepsilon(t)|^p \\
&\leq 1 + C|x_0|^p + Ct^{p-1}\int_0^t \mathbb{E}|b(Y^\varepsilon(s_\delta),\theta)|^p \mathrm{d}s + C\mathbb{E}\Big(\int_0^t (\varepsilon\|\sigma(Y^\varepsilon(s_\delta))\|_{\mathrm{HS}})^p \mathrm{d}s\Big)^{\frac{p}{2}} \\
&\leq 1 + C|x_0|^p + C(t^{p-1}+t^{\frac{p-2}{2}})\int_0^t \{\mathbb{E}|b(Y^\varepsilon(s_\delta),\theta)|^p + \mathbb{E}\|\sigma(Y^\varepsilon(s_\delta))\|_{\mathrm{HS}}^p\}\mathrm{d}s \\
&\leq 1 + C|x_0|^p + C(t^{p-1}+t^{\frac{p-2}{2}})\int_0^t \{1+\mathbb{E}|Y^\varepsilon(s_\delta)|^p\}\mathrm{d}s,
\end{aligned}$$

where the third inequality applies Assumptions 5.1 and 5.2. Since

$$\sup_{0\leq s\leq t}|Y^\varepsilon(s_\delta)| \leq \sup_{0\leq s\leq t}|Y^\varepsilon(s)|,$$

we derive that

$$1+\mathbb{E}|Y^\varepsilon(t)|^p \leq 1 + C|x_0|^p + C(t^{p-1}+t^{\frac{p-2}{2}})\int_0^t (1+\mathbb{E}|Y^\varepsilon(s)|^p)\mathrm{d}s$$

for $C > 0$. Consequently, the Gronwall inequality yields that there exists a constant $C_T > 0$ such that

$$\sup_{0\leq t\leq T} \mathbb{E}|Y^\varepsilon(t)|^p \leq C_T(1+|x_0|^p).$$

The proof is completed. $\square$

Lemma 5.14. *Under Assumptions 5.1 and 5.2, for any $\delta \in (0,1)$ and $\varepsilon > 0$, it holds that*

$$\mathbb{E}|Y^\varepsilon(t) - Y^\varepsilon(t_\delta)|^2 \leq C\delta(\delta+\varepsilon^2). \tag{5.26}$$

Proof. Applying the elementary inequality, the Hölder inequality, and the Itô isometry, we deduce that

$$\begin{aligned}\mathbb{E}|Y^\varepsilon(t)-Y^\varepsilon(t_\delta)|^2 &\le 2\mathbb{E}\left|\int_{t_\delta}^t b(Y^\varepsilon(s_\delta),\theta)\mathrm{d}s\right|^2+2\mathbb{E}\left|\varepsilon\int_{t_\delta}^t \sigma(Y^\varepsilon(s_\delta))\mathrm{d}W(s)\right|^2\\ &\le 2\delta\mathbb{E}\int_{t_\delta}^t |b(Y^\varepsilon(s_\delta),\theta)|^2\mathrm{d}s+2\varepsilon^2\mathbb{E}\int_{t_\delta}^t \|\sigma(Y^\varepsilon(s_\delta))\|_{\mathrm{HS}}^2\mathrm{d}s\\ &\le C\delta(\delta+\varepsilon^2),\end{aligned}$$

where the third inequality utilizes the fact that b and σ are bounded from Assumptions 5.1 and 5.2 respectively. $\square$

Since b is singular, we need to construct a regular transform to remove b. For any $\lambda>0$, consider the following partial differential equation (PDE),

$$Lu^\lambda+b+\nabla_b u^\lambda=\lambda u^\lambda, \tag{5.27}$$

where

$$L=\frac{1}{2}\varepsilon^2\mathrm{Tr}\Big(\sigma\sigma^*\nabla^2 u^\lambda\Big),$$

and $\nabla_b u^\lambda:=\langle\nabla u^\lambda,b\rangle$ is the directional derivative of u^λ along b.

Motivated by the result, Lemma 2.1 in [BHY19], by the same line, we introduce the following lemma, the regularity of the non-degenerate Kolmogorov equation.

Lemma 5.15. *Under Assumptions 5.1 and 5.2, for a sufficiently large $\lambda>0$, (5.27) admits a unique solution $u^\lambda\in C([0,T];C_b^2(\mathbb{R}^d;\mathbb{R}^d)$ with*

$$\|\nabla u^\lambda\|_\infty\le\frac{1}{2},\quad \|\nabla^2 u^\lambda\|_\infty\le\frac{1}{2}. \tag{5.28}$$

We can construct a diffeomorphism on $\mathbb{R}^d$, i.e,

$$V^\lambda(x):=x+u^\lambda(x),\quad x\in\mathbb{R}^d, \tag{5.29}$$

and $V^\lambda(x)$ inherits the smoothness of $u^\lambda(x)$.

The following two lemmas are essential in consistency and asymptotic distribution properties. One of the difficulties in the proof is dealing with $\mathbb{E}|Y^\varepsilon(t)-X^\varepsilon(t)|^2$ and $\mathbb{E}|X^\varepsilon(t)-X^0(t)|^2$ owing to the α-continuous drift. So, motivated by [BHY19], we will adopt Zvonkin's transformation.

Lemma 5.16. *Assume that Assumptions 5.1 and 5.2 hold. Then, for any $\delta\in(0,1)$ and $\varepsilon>0$, there exists a constant $C_T>0$ such that*

$$\begin{aligned}&\sup_{0\le t\le T}\mathbb{E}|Y^\varepsilon(t_\delta)-X^0(t)|^2\\ &\le C_T\left[\left(\delta^\alpha(\delta^\alpha+\varepsilon^{2\alpha})^\alpha+\delta(\varepsilon^2+\varepsilon^4)(\delta+\varepsilon^2)\right)\mathrm{e}^{1+\varepsilon^2}+(\varepsilon^2+\varepsilon^4)+\delta(\delta+\varepsilon^2)\right].\end{aligned} \tag{5.30}$$

Proof. For any $t \in [0,T]$, by the triangle inequality, it holds that

$$\begin{aligned} &\mathbb{E}|Y^\varepsilon(t_\delta) - X^0(t)|^2 \\ &\quad \le 3(\mathbb{E}|Y^\varepsilon(t_\delta) - Y^\varepsilon(t)|^2 + \mathbb{E}|Y^\varepsilon(t) - X^\varepsilon(t)|^2 + \mathbb{E}|X^\varepsilon(t) - X^0(t)|^2). \end{aligned} \tag{5.31}$$

Let us first estimate $\mathbb{E}|Y(t)^\varepsilon - X^\varepsilon(t)|^2$ using Zvonkin's transformation. Applying the Itô formula to $V^\lambda(Y^\varepsilon(t))$ and $V^\lambda(X^\varepsilon(t))$, one has

$$\begin{aligned} \mathrm{d}V^\lambda(Y^\varepsilon(t)) &= \nabla V^\lambda(Y^\varepsilon(t)) b(Y^\varepsilon(t_\delta),\theta)\mathrm{d}t + \varepsilon \nabla V^\lambda(Y^\varepsilon(t))\sigma(Y^\varepsilon(t_\delta))\mathrm{d}W(t) \\ &\quad + \frac{1}{2}\varepsilon^2 \mathrm{Tr}\Big(\sigma(Y^\varepsilon(t_\delta))\sigma(Y^\varepsilon(t_\delta))^* \nabla^2 u^\lambda(Y^\varepsilon(t))\Big)\mathrm{d}t, \end{aligned} \tag{5.32}$$

$$\begin{aligned} \mathrm{d}V^\lambda(X^\varepsilon(t)) &= \nabla V^\lambda(X^\varepsilon(t)) b(X^\varepsilon(t),\theta)\mathrm{d}t + \varepsilon \nabla V^\lambda(X^\varepsilon(t))\sigma(X^\varepsilon(t))\mathrm{d}W(t) \\ &\quad + \frac{1}{2}\varepsilon^2 \mathrm{Tr}\left(\sigma\sigma^* \nabla^2 u^\lambda\right)(X^\varepsilon(t))\mathrm{d}t, \end{aligned} \tag{5.33}$$

respectively. Then we use (5.27) to deal with the drift, and we have

$$\begin{aligned} \mathrm{d}V^\lambda(Y^\varepsilon(t)) &= \lambda u^\lambda(Y^\varepsilon(t))\mathrm{d}t + \nabla V^\lambda(Y^\varepsilon(t))(b(Y^\varepsilon(t_\delta),\theta) - b(Y^\varepsilon(t),\theta))\mathrm{d}t \\ &\quad + \varepsilon \nabla V^\lambda(Y^\varepsilon(t))\sigma(Y^\varepsilon(t_\delta))\mathrm{d}W(t) \\ &\quad + \frac{1}{2}\varepsilon^2 \mathrm{Tr}\Big([\sigma(Y^\varepsilon(t_\delta))\sigma(Y^\varepsilon(t_\delta))^* - \sigma(Y^\varepsilon(t))\sigma(Y^\varepsilon(t))^*]\nabla^2 u^\lambda(Y^\varepsilon(t))\Big)\mathrm{d}t, \end{aligned}$$

$$\mathrm{d}V^\lambda(X^\varepsilon(t)) = \lambda u^\lambda(X^\varepsilon(t))\mathrm{d}t + \varepsilon \nabla V^\lambda(X^\varepsilon(t))\sigma(X^\varepsilon(t))\mathrm{d}W(t), \tag{5.34}$$

respectively. Hence, the above equations, together with the elementary inequality, the Hölder inequality and the Itô isometry, yield that

$$\begin{aligned} &\mathbb{E}|V^\lambda(Y^\varepsilon(t)) - V^\lambda(X^\varepsilon(t))|^2 \\ &\quad \le CT\int_0^t \mathbb{E}|u^\lambda(Y^\varepsilon(s)) - u^\lambda(X^\varepsilon(s))|^2\mathrm{d}s \\ &\qquad + CT\int_0^t \mathbb{E}|\nabla V^\lambda(Y^\varepsilon(s))(b(Y^\varepsilon(s_\delta),\theta) - b(Y^\varepsilon(s),\theta))|^2\mathrm{d}s \\ &\qquad + C\varepsilon^2\int_0^t \mathbb{E}\|\nabla V^\lambda(Y^\varepsilon(s))\sigma(Y^\varepsilon(s_\delta)) - \nabla V^\lambda(X^\varepsilon(s))\sigma(X^\varepsilon(s))\|_{\mathrm{HS}}^2\mathrm{d}s \\ &\qquad + CT\varepsilon^4\int_0^t \mathbb{E}\Big|\mathrm{Tr}\Big([\sigma(Y^\varepsilon(s_\delta))\sigma(Y^\varepsilon(s_\delta))^* - \sigma(Y^\varepsilon(s))\sigma(Y^\varepsilon(s))^*]\nabla^2 u^\lambda(Y^\varepsilon(s))\Big)\Big|^2\mathrm{d}s \\ &\quad =: \sum_{i=1}^4 G_i(t). \end{aligned} \tag{5.35}$$

Subsequently, we estimate them one by one.

For the $G_1(t)$, we use the fact from the Taylor expansion, there exists $\xi^\varepsilon(s)$ between $Y^\varepsilon(s)$ and $X^\varepsilon(s)$ such that

$$|u^\lambda(Y^\varepsilon(s)) - u^\lambda(X^\varepsilon(s))| \le \|\nabla u^\lambda(\xi^\varepsilon(s))\|_\infty |Y^\varepsilon(s) - X^\varepsilon(s)|. \tag{5.36}$$

This, together with (5.28), we infer that

$$G_1(t) \le CT \int_0^t \mathbb{E}|Y^\varepsilon(s) - X^\varepsilon(s)|^2 \mathrm{d}s. \tag{5.37}$$

For the $G_2(t)$, (5.28) and Assumption 5.1 lead to

$$\begin{aligned} G_2(t) &\le CT \int_0^t \mathbb{E}\Big(\|\nabla V^\lambda\|_\infty^2 |b(Y^\varepsilon(s_\delta),\theta) - b(Y^\varepsilon(s),\theta)|^2\Big)\mathrm{d}s \\ &\le CT \int_0^t \mathbb{E}|b(Y^\varepsilon(s_\delta),\theta) - b(Y^\varepsilon(s),\theta)|^2 \mathrm{d}s \\ &\le CT \int_0^t \mathbb{E}|Y^\varepsilon(s_\delta) - Y^\varepsilon(s)|^{2\alpha} \mathrm{d}s \\ &\le CT \int_0^t \Big(\mathbb{E}|Y^\varepsilon(s_\delta) - Y^\varepsilon(s)|^2\Big)^\alpha \mathrm{d}s, \end{aligned} \tag{5.38}$$

where in the last step we have taken advantage of the Jensen inequality for the concave function x^α, $\alpha \in (0,1]$.

For the $G_3(t)$, by Assumption 5.1, Assumption 5.2, (5.28) and the Taylor expansion, one has

$$\begin{aligned} G_3(t) &\le C\varepsilon^2 \int_0^t \mathbb{E}\Big\|\Big(\nabla V^\lambda(Y^\varepsilon(s)) - \nabla V^\lambda(X^\varepsilon(s))\Big)\sigma(Y^\varepsilon(s_\delta))\Big\|_{\mathrm{HS}}^2 \mathrm{d}s \\ &\quad + C\varepsilon^2 \int_0^t \mathbb{E}\Big\|\nabla V^\lambda(X^\varepsilon(s))\Big(\sigma(Y^\varepsilon(s_\delta)) - \sigma(X^\varepsilon(s))\Big)\Big\|_{\mathrm{HS}}^2 \mathrm{d}s \\ &\le C\varepsilon^2 \int_0^t (\mathbb{E}|Y^\varepsilon(s) - X^\varepsilon(s)|^2 + \mathbb{E}|Y^\varepsilon(s_\delta) - Y^\varepsilon(s)|^2)\mathrm{d}s. \end{aligned} \tag{5.39}$$

For the $G_4(t)$, one uses a similar operation when we deal with $G_3(t)$, and obtains that

$$\begin{aligned} G_4(t) &\le CT\varepsilon^4 \int_0^t \mathbb{E}\Big(\|\nabla^2 u^\lambda\|_\infty^2 \|\sigma(Y^\varepsilon(s_\delta))\sigma(Y^\varepsilon(s_\delta))^* - \sigma(Y^\varepsilon(s))\sigma(Y^\varepsilon(s))^*\|_{\mathrm{HS}}^2\Big)\mathrm{d}s \\ &\le CT\varepsilon^4 \int_0^t \mathbb{E}\|\sigma(Y^\varepsilon(s_\delta))\sigma(Y^\varepsilon(s_\delta))^* - \sigma(Y^\varepsilon(s))\sigma(Y^\varepsilon(s))^*\|_{\mathrm{HS}}^2 \mathrm{d}s \\ &\le CT\varepsilon^4 \int_0^t \Big(\mathbb{E}\|(\sigma(Y^\varepsilon(s_\delta)) - \sigma(Y^\varepsilon(s)))\sigma(Y^\varepsilon(s_\delta))^*\|_{\mathrm{HS}}^2 \\ &\quad + \mathbb{E}\|\sigma(Y^\varepsilon(s))(\sigma(Y^\varepsilon(s_\delta)) - \sigma(Y^\varepsilon(s)))^*\|_{\mathrm{HS}}^2\Big)\mathrm{d}s \end{aligned} \tag{5.40}$$

$$\leq CT\varepsilon^4 \int_0^t \mathbb{E}|Y^\varepsilon(s_\delta) - Y^\varepsilon(s)|^2 \mathrm{d}s.$$

Consequently, taking (5.37)-(5.40) into consideration, we deduce from (5.35) and (5.26) that

$$\begin{aligned}
&\mathbb{E}|V^\lambda(Y^\varepsilon(t)) - V^\lambda(X^\varepsilon(t))|^2 \\
&\leq C\int_0^t \Big((T+\varepsilon^2)\mathbb{E}|Y^\varepsilon(s_\delta) - Y^\varepsilon(s)|^2 + T(\mathbb{E}|Y^\varepsilon(s_\delta) - Y^\varepsilon(s)|^2)^\alpha \\
&\quad + (\varepsilon^2 + T\varepsilon^4)\mathbb{E}|Y^\varepsilon(s) - X^\varepsilon(s)|^2\Big)\mathrm{d}s \\
&\leq C_T\int_0^t \Big((1+\varepsilon^2)\mathbb{E}|Y^\varepsilon(s_\delta) - Y^\varepsilon(s)|^2 + (\mathbb{E}|Y^\varepsilon(s_\delta) - Y^\varepsilon(s)|^2)^\alpha \\
&\quad + (\varepsilon^2 + \varepsilon^4)\mathbb{E}|Y^\varepsilon(s) - X^\varepsilon(s)|^2\Big)\mathrm{d}s \\
&\leq C_T\int_0^t \big((1+\varepsilon^2)\mathbb{E}|Y^\varepsilon(s) - X^\varepsilon(s)|^2 + \delta^\alpha(\delta^\alpha + \varepsilon^{2\alpha})^\alpha + \delta(\varepsilon^2+\varepsilon^4)(\delta+\varepsilon^2)\big)\,\mathrm{d}s.
\end{aligned} \tag{5.41}$$

Note that we work within a fixed time horizon, $[0,T]$, so T is absorbed into C in the above second inequality.

To proceed, we first need to find the relation between $|Y^\varepsilon(t) - X^\varepsilon(t)|$ and $|V^\lambda(Y^\varepsilon(t)) - V^\lambda(X^\varepsilon(t))|$ to handle the left side of (5.41). Taking advantage of the elementary inequality $(a+b)^2 \leq 2(a^2+b^2)$, for any $a, b > 0$, it follows from (5.28) that

$$\begin{aligned}
&|Y^\varepsilon(t) - X^\varepsilon(t)|^2 \\
&\quad = |Y^\varepsilon(t) + u^\lambda(Y^\varepsilon(t)) - (X^\varepsilon(t) + u^\lambda(X^\varepsilon(t))) + u^\lambda(X^\varepsilon(t)) - u^\lambda(Y^\varepsilon(t))|^2 \\
&\quad = |V^\lambda(Y^\varepsilon(t)) - V^\lambda(X^\varepsilon(t)) + u^\lambda(X^\varepsilon(t)) - u^\lambda(Y^\varepsilon(t))|^2 \\
&\quad \leq 2\Big(|V^\lambda(Y^\varepsilon(t)) - V^\lambda(X^\varepsilon(t))|^2 + |u^\lambda(X^\varepsilon(t)) - u^\lambda(Y^\varepsilon(t))|^2\Big) \\
&\quad \leq 2|V^\lambda(Y^\varepsilon(t)) - V^\lambda(X^\varepsilon(t))|^2 + \frac{1}{2}|Y^\varepsilon(t) - X^\varepsilon(t)|^2.
\end{aligned}$$

Hence, we have

$$|Y^\varepsilon(t) - X^\varepsilon(t)|^2 \leq 4|V^\lambda(Y^\varepsilon(t)) - V^\lambda(X^\varepsilon(t))|^2. \tag{5.42}$$

And then, in terms of (5.42), it follows from (5.41) that

$$\begin{aligned}
&\mathbb{E}|Y^\varepsilon(t) - X^\varepsilon(t)|^2 \\
&\quad \leq C_T\left(\delta^\alpha(\delta^\alpha + \varepsilon^{2\alpha})^\alpha + \delta(\varepsilon^2+\varepsilon^4)(\delta+\varepsilon^2)\right) + C_T(1+\varepsilon^2)\int_0^t \mathbb{E}|Y^\varepsilon(s) - X^\varepsilon(s)|^2\mathrm{d}s.
\end{aligned}$$

Next, using the Gronwall inequality, we have

$$\mathbb{E}|Y^\varepsilon(t) - X^\varepsilon(t)|^2 \leq C_T\left(\delta^\alpha(\delta^\alpha + \varepsilon^{2\alpha})^\alpha + \delta(\varepsilon^2+\varepsilon^4)(\delta+\varepsilon^2)\right)\mathrm{e}^{1+\varepsilon^2}. \tag{5.43}$$

Now we are left to estimate $\mathbb{E}|X^\varepsilon(t)-X^0(t)|^2$. Applying the same technique as above to deal with $V^\lambda(X^\varepsilon(t))$ in (5.34), one has

$$\mathrm{d}V^\lambda(X^0(t)) = \lambda u^\lambda(X^0(t))\mathrm{d}t - \frac{1}{2}\varepsilon^2\mathrm{Tr}\Big(\sigma(\sigma)^*\nabla^2 u^\lambda\Big)(X^0(t))\mathrm{d}t.$$

This, together with (5.34), yields that from the Hölder inequality and the Itô isometry

$$\begin{aligned}
&\mathbb{E}|V^\lambda(X^\varepsilon(t)) - V^\lambda(X^0(t))|^2 \\
&\leq CT\int_0^t \mathbb{E}|u^\lambda(X^\varepsilon(s)) - u^\lambda(X^0(s))|^2\mathrm{d}s \\
&\quad + C\varepsilon^2\int_0^t \mathbb{E}\|\nabla V^\lambda(X^\varepsilon(s))\sigma(X^\varepsilon(s))\|_{\mathrm{HS}}\mathrm{d}s \\
&\quad + CT\varepsilon^4\int_0^t \mathbb{E}\left|\mathrm{Tr}\Big(\sigma(\sigma)^*\nabla^2 u^\lambda\Big)(X^0(s))\right|^2\mathrm{d}s \\
&\leq CT\int_0^t \mathbb{E}|X^\varepsilon(s) - X^0(s)|^2\mathrm{d}s + CT\varepsilon^2 + CT^2\varepsilon^4 \\
&\leq C_T\left(\int_0^t \mathbb{E}|X^\varepsilon(s) - X^0(s)|^2\mathrm{d}s + \varepsilon^2 + \varepsilon^4\right).
\end{aligned} \tag{5.44}$$

Moreover, using the same technology as obtaining (5.42), we have

$$|V^\lambda(X^\varepsilon(t)) - V^\lambda(X^0(t))| \leq 4|X^\varepsilon(t) - V^0(t)|. \tag{5.45}$$

Then, we deal with the left side of (5.44) by (5.45), and we have

$$\mathbb{E}|X^\varepsilon(t) - X^0(t)|^2 \leq C_T\left(\int_0^t \mathbb{E}|X^\varepsilon(s) - X^0(s)|^2\mathrm{d}s + \varepsilon^2 + \varepsilon^4\right).$$

Applying the Gronwall inequality, one has

$$\mathbb{E}[|X^\varepsilon(t) - X^0(t)|^2] \leq C_T(\varepsilon^2 + \varepsilon^4). \tag{5.46}$$

Plugin (5.26), (5.43) and (5.46) into (5.31), we get

$$\begin{aligned}
&\mathbb{E}|Y^\varepsilon(t_\delta) - X^0(t)|^2 \\
&\leq C_T\left[\Big(\delta^\alpha(\delta^\alpha + \varepsilon^{2\alpha})^\alpha + \delta(\varepsilon^2+\varepsilon^4)(\delta+\varepsilon^2)\Big)\,\mathrm{e}^{1+\varepsilon^2} + (\varepsilon^2+\varepsilon^4) + \delta(\delta+\varepsilon^2)\right].
\end{aligned}$$

We have derived the desired assertions. □

Lemma 5.17 below plays a crucial role in revealing the asymptotic behavior of the LSE of the unknown parameter.

Lemma 5.17. *Let Assumptions 5.1 and 5.2 hold. Then we have*

$$\begin{aligned}\delta\sum_{k=1}^{n}&G^*(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\hat\sigma(Y^\varepsilon((k-1)\delta))G(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\\&\to H(\theta):=\int_0^T G^*(X^0(t),\theta,\theta_0)\hat\sigma(X^0(t))G(X^0(t),\theta,\theta_0)\mathrm{d}t,\end{aligned}\tag{5.47}$$

in L^1 *uniformly with respect to* θ *as* $\varepsilon\to0$ *and* $\delta\to0$, *in which* $(X^0(t))_{t\in[0,T]}$ *is the solution to ordinary differential equation* (5.3). *Moreover,*

$$\sum_{k=1}^{n}G^*(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\hat\sigma(Y^\varepsilon((k-1)\delta))F_k(\theta_0)\to0\tag{5.48}$$

in L^1 *uniformly with respect to* θ *as* $\varepsilon\to0$ *and* $\delta\to0$.

Proof. We begin with the proof of (5.47). It is straightforward to see that

$$\begin{aligned}\delta\sum_{k=1}^{n}&G^*(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\hat\sigma(Y^\varepsilon((k-1)\delta))G(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\\&-\int_0^T G^*(X^0(t),\theta,\theta_0)\hat\sigma(X^0(t))G(X^0(t),\theta,\theta_0)\mathrm{d}s\\=&\int_0^T\Big\{G^*(Y^\varepsilon(t_\delta),\theta,\theta_\theta)\hat\sigma(Y^\varepsilon(t_\delta))G(Y^\varepsilon(t_\delta),\theta,\theta_0)\\&-G^*(X^0(t),\theta,\theta_0)\hat\sigma(X^0(t))G(X^0(t),\theta,\theta_0)\Big\}\mathrm{d}s\\=&\int_0^T\Big(G(Y^\varepsilon(t_\delta),\theta,\theta_0)-G(X^0(t),\theta,\theta_0)\Big)^*\hat\sigma(Y^\varepsilon(t_\delta))G(Y^\varepsilon(t_\delta),\theta,\theta_0)\mathrm{d}t\\&+\int_0^T G^*(X^0(t),\theta,\theta_0)\Big(\hat\sigma(Y^\varepsilon(t_\delta))-\hat\sigma(X^0(t))\Big)G(Y^\varepsilon(t_\delta),\theta,\theta_0)\mathrm{d}t\\&+\int_0^T G^*(X^0(t),\theta,\theta_0)\hat\sigma(X^0(t))\Big(G(Y^\varepsilon(t_\delta),\theta,\theta_0)-G(X^0(t),\theta,\theta_0)\Big)\mathrm{d}t\\=:&\,J_1(T,\varepsilon,\delta)+J_2(T,\varepsilon,\delta)+J_3(T,\varepsilon,\delta).\end{aligned}$$

Before estimating

$$\mathbb{E}|J_i(T,\varepsilon,\delta)|\to0,\quad i=1,2,3,$$

as $\varepsilon\to0$, $\delta\to0$, we prepare the estimations for G and $\hat\sigma$. For any $x,y\in\mathbb{R}^d$, $\alpha\in(0,1]$, it follows from (5.13) and Assumption 5.1 that

$$\begin{aligned}|G(x,\theta,\theta_0)-G(y,\theta,\theta_0)|&\le|b(x,\theta_0)-b(y,\theta_0)|+|b(x,\theta)-b(y,\theta)|\\&\le C|x-y|^\alpha.\end{aligned}\tag{5.49}$$

For any $x, y \in \mathbb{R}^d$, we get from (5.8) and Assumption 5.2 that

$$\begin{aligned}
\|\hat{\sigma}(x) - \hat{\sigma}(y)\|_{\mathrm{HS}} &= \|(\sigma\sigma^*)^{-1}(x) - (\sigma\sigma^*)^{-1}(y)\|_{\mathrm{HS}} \\
&= \|(\sigma^*)^{-1}\sigma^{-1}(x) - (\sigma^*)^{-1}\sigma^{-1}(y)\|_{\mathrm{HS}} \\
&\le \|\nabla\sigma^{-1}\|_{\mathrm{HS}}\|\sigma^{-1}(x) - \sigma^{-1}(y)\|_{\mathrm{HS}} \\
&\le \|\nabla\sigma^{-1}\|_{\mathrm{HS}}^2|x-y| \\
&\le C|x-y|.
\end{aligned} \tag{5.50}$$

Hence, by (5.50) we have

$$\|\hat{\sigma}(x)\|_{\mathrm{HS}} \le \|\hat{\sigma}(x) - \hat{\sigma}(0)\|_{\mathrm{HS}} + \|\hat{\sigma}(0)\|_{\mathrm{HS}} \le C(1+|x|). \tag{5.51}$$

Consequently, combining with (5.49), (5.51), and Assumption 5.1, we deduce that when $\varepsilon \to 0$ and $\delta \to 0$

$$\begin{aligned}
&\mathbb{E}|J_1(T,\varepsilon,\delta)| \\
&\le C\mathbb{E}\int_0^T |(G(Y^\varepsilon(t_\delta),\theta,\theta_0) - G(X^0(t),\theta,\theta_0))^*| \cdot \|\hat{\sigma}(Y^\varepsilon(t_\delta))\|_{\mathrm{HS}} \cdot |G(Y^\varepsilon(t_\delta),\theta,\theta_0)|\mathrm{d}t \\
&\le C\mathbb{E}\int_0^T |Y^\varepsilon(t_\delta) - X^0(t)|^\alpha(1+|Y^\varepsilon(t_\delta)|)\mathrm{d}t \\
&\le C\int_0^T (\mathbb{E}[|Y^\varepsilon(t_\delta) - X^0(t)|^{2\alpha}])^{1/2}(1+\mathbb{E}|Y^\varepsilon(t_\delta)|^2)^{1/2}\mathrm{d}t \\
&\le C\int_0^T (\mathbb{E}[|Y^\varepsilon(t_\delta) - X^0(t)|^{2}])^{\alpha/2}(1+\mathbb{E}|Y^\varepsilon(t_\delta)|^2)^{1/2}\mathrm{d}t \to 0,
\end{aligned}$$

where the last three inequalities use the Hölder inequality, the Jensen inequality, (5.30) and (5.25) in order. Similarly, we have

$$\begin{aligned}
&\mathbb{E}|J_2(T,\varepsilon,\delta)| \\
&\quad\le C\mathbb{E}\int_0^T |G^*(t, X^0(t),\theta,\theta_0)| \cdot \|\hat{\sigma}(Y^\varepsilon(t_\delta)) - \hat{\sigma}(X^0(t))\|_{\mathrm{HS}} \cdot |G(Y^\varepsilon(t_\delta),\theta,\theta_0)|\mathrm{d}t \\
&\quad\le C\mathbb{E}\int_0^T |Y^\varepsilon(t_\delta) - X^0(t)|\mathrm{d}t \le C\int_0^T (\mathbb{E}[|Y^\varepsilon(t_\delta) - X^0(t)|^2])^{1/2}\mathrm{d}t \to 0
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}|J_3(T,\varepsilon,\delta)| \\
&\quad\le C\mathbb{E}\int_0^T |G^*(X^0(t),\theta,\theta_0)| \cdot \|\hat{\sigma}(X^0(t))\|_{\mathrm{HS}} \cdot |G(Y^\varepsilon(t_\delta),\theta,\theta_0) - G(X^0(t),\theta,\theta_0)|\mathrm{d}t \\
&\quad\le C\mathbb{E}\int_0^T |Y^\varepsilon(t_\delta) - X^0(t)|^\alpha\mathrm{d}t \le \int_0^T (\mathbb{E}|Y^\varepsilon(t_\delta) - X^0(t)|^2)^{\alpha/2}\mathrm{d}t \to 0.
\end{aligned}$$

Then, we derive that

$$\mathbb{E}|J_i(T,\varepsilon,\delta)| \to 0, \qquad i = 1,\ 2,\ 3, \tag{5.52}$$

whenever $\varepsilon \to 0$ and $\delta \to 0$. Hence, (5.47) follows immediately from (5.52).

In the sequel, we are going to show that (5.48) holds. Note that

$$\begin{aligned}
&\sum_{k=1}^{n} G^*(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\hat{\sigma}(Y^\varepsilon((k-1)\delta))F_k(\theta_0)\\
&\quad = \varepsilon\sum_{k=1}^{n} G^*(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\hat{\sigma}(Y^\varepsilon((k-1)\delta))\sigma(Y^\varepsilon((k-1)\delta))(W(k\delta)-W((k-1)\delta))\\
&\quad = \varepsilon\int_0^T G^*(Y^\varepsilon(t_\delta),\theta,\theta_0)\hat{\sigma}(Y^\varepsilon(t_\delta))\sigma(Y^\varepsilon(t_\delta))\mathrm{d}W(t),
\end{aligned}$$

where in the first step, we use the definition of $F_k(\theta_0)$. Next, via the Hölder inequality, Assumption 5.1 and Assumption 5.2, as well as the result in Lemma 5.13, one achieves that

$$\begin{aligned}
&\mathbb{E}\left|\sum_{k=1}^{n} G^*(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\hat{\sigma}(Y^\varepsilon((k-1)\delta))F_k(\theta_0)\right|\\
&\quad \le \varepsilon\Big(\int_0^T \mathbb{E}|G^*(Y^\varepsilon(t_\delta),\theta,\theta_0)\hat{\sigma}(Y^\varepsilon(t_\delta))\sigma(Y^\varepsilon(t_\delta))|^2\mathrm{d}t\Big)^{1/2}\\
&\quad \le C\varepsilon\Big(\int_0^T (1+\mathbb{E}|Y^\varepsilon(t_\delta)|^4)\mathrm{d}t\Big)^{1/2} \le C\varepsilon,
\end{aligned} \tag{5.53}$$

where we have applied (5.25) in the last procedure. Therefore, (5.48) is now available from (5.53). □

With Lemmas 5.13 - 5.17 in hand, we are in the position to complete the proof of Theorem 5.4.

Proof of Theorem 5.4. From (5.10), a straightforward calculation gives that

$$\begin{aligned}
&\Phi_{n,\varepsilon}(\theta)\\
&\quad = \delta^{-1}\sum_{k=1}^{n}\Big(F_k^*(\theta)\hat{\sigma}(Y^\varepsilon((k-1)\delta))F_k(\theta) - F_k^*(\theta_0)\hat{\sigma}(Y^\varepsilon((k-1)\delta))F_k(\theta_0)\Big)\\
&\quad = \delta^{-1}\sum_{k=1}^{n}\Big\{\Big(F_k(\theta_0)+G(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\delta\Big)^*\hat{\sigma}(Y^\varepsilon((k-1)\delta))\\
&\qquad \times\Big(F_k(\theta_0)+G(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\delta\Big) - F_k^*(\theta_0)\hat{\sigma}(Y^\varepsilon((k-1)\delta))F_k(\theta_0)\Big\} \qquad (5.54)\\
&\quad = \delta\sum_{k=1}^{n} G^*(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\hat{\sigma}(Y^\varepsilon((k-1)\delta))G(Y^\varepsilon((k-1)\delta),\theta,\theta_0)
\end{aligned}$$

$$+2\sum_{k=1}^{n} G^*(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\hat{\sigma}(Y^\varepsilon((k-1)\delta))F_k(\theta_0).$$

In terms of Lemmas 5.17, we deduce from the Chebyshev inequality that

$$\sup_{\theta\in\Theta}|-\Phi_{n,\varepsilon}(\theta)-(-H(\theta))|\to 0 \qquad \text{in probability.}$$

On the other hand, for any $\kappa>0$, notice that

$$\sup_{|\theta-\theta_0|\geq\kappa}(-H(\theta))<-H(\theta_0)=0$$

due to $H(\cdot)>0$. Moreover, according to the notion of $\hat{\theta}_{n,\varepsilon}$, one has $-\Phi_{n,\varepsilon}(\hat{\theta}_{n,\varepsilon})\geq -\Phi_{n,\varepsilon}(\theta_0)=0$. As far as our present model is concerned, all of the assumptions in Theorem 1.5 with $M_n(\cdot)=-\Phi_{n,\varepsilon}(\cdot)$ and $M(\cdot)=-H(\cdot)$ are fulfilled. As a consequence, we conclude that $\hat{\theta}_{n,\varepsilon}\to\theta_0$ in probability as $\varepsilon\to 0$ and $n\to\infty$, as required. □

We first prove the following lemmas, which will be used in the proof of Theorem 5.5.

Lemma 5.18. *Under Assumptions 5.1, 5.2 and 5.3,* Υ *is defined in* (5.15)*. Then we have*

$$\varepsilon^{-1}(\nabla_\theta\Phi_{n,\varepsilon})(\theta_0)\to -2\int_0^T \Upsilon(X^0(t),\theta_0)\mathrm{d}W(t) \quad \textit{in probability} \tag{5.55}$$

whenever $\varepsilon\to 0$ *and* $\delta\to 0$.

Proof. By the chain rule, one infers from (5.54), (5.7), (5.13) and (5.4) that

$$\begin{aligned}
&\varepsilon^{-1}(\nabla_\theta\Phi_{n,\varepsilon})(\theta_0)\\
&\quad=2\varepsilon^{-1}\sum_{k=1}^{n}(\nabla_\theta G)^*(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\hat{\sigma}(Y^\varepsilon((k-1)\delta))F_k(\theta_0)\\
&\quad=-2\sum_{k=1}^{n}(\nabla_\theta b)^*(Y^\varepsilon((k-1)\delta),\theta_0)\hat{\sigma}(Y^\varepsilon((k-1)\delta))\sigma(Y^\varepsilon((k-1)\delta))\Delta W_k\\
&\quad=-2\int_0^T \Upsilon(Y^\varepsilon((k-1)\delta),\theta_0)\mathrm{d}W(t)\\
&\quad=-2\int_0^T \Upsilon(Y^\varepsilon(t_\delta),\theta_0)\mathrm{d}W(t),
\end{aligned}$$

where in the second line we used (5.4), (5.7) and the fact that

$$(\nabla_\theta G)(Y^\varepsilon((k-1)\delta),\theta,\theta_0)=-(\nabla_\theta b)(Y^\varepsilon((k-1)\delta),\theta_0). \tag{5.56}$$

To achieve (5.55), in terms of Theorem 2.6 of [Fri75] for any $\rho>0$ and $\kappa>0$

$$\mathbb{P}\left(\left|\int_0^T(\Upsilon(Y^\varepsilon(t_\delta),\theta_0)-\Upsilon(X^0(t),\theta_0)\right|\geq\rho\right)$$

$$\leq \mathbb{P}\left(\int_0^T \|\Upsilon(Y^\varepsilon(t_\delta),\theta_0)-\Upsilon(X^0(t),\theta_0)\|^2\mathrm{d}t \geq \rho^2\kappa\right)+\kappa,$$

it is sufficient to claim that

$$\int_0^T \|\Upsilon(Y^\varepsilon(t_\delta),\theta_0)-\Upsilon(X^0(t),\theta_0)\|^2\mathrm{d}t \to 0 \quad \text{in probability} \tag{5.57}$$

as $\varepsilon \to 0$, $\delta \to 0$ and the arbitrariness of κ. Observe that

$$\begin{aligned}
&\|\Upsilon(Y^\varepsilon(t_\delta),\theta_0)-\Upsilon(X^0(t),\theta_0)\|_{\mathrm{HS}}^2\\
&\quad\leq \|[(\nabla_\theta b)^*(Y^\varepsilon(t_\delta),\theta_0)-(\nabla_\theta b)^*(X^0(t),\theta_0)]\hat{\sigma}(Y^\varepsilon(t_\delta)\sigma(Y^\varepsilon(t_\delta)\|_{\mathrm{HS}}^2\\
&\qquad+\|(\nabla_\theta b)^*(X^0(t),\theta_0)[\hat{\sigma}(Y^\varepsilon(t_\delta)-\hat{\sigma}(X^0(t))]\sigma(Y^\varepsilon(t_\delta))\|_{\mathrm{HS}}^2\\
&\qquad+\|(\nabla_\theta b)^*(X^0(t),\theta_0)\hat{\sigma}(X^0(t))[\sigma(Y^\varepsilon(t_\delta))-\sigma(X^0(t))]\|_{\mathrm{HS}}^2\\
&\quad=: \varsigma_1(t,\varepsilon,\delta)+\varsigma_2(t,\varepsilon,\delta)+\varsigma_3(t,\varepsilon,\delta).
\end{aligned}$$

By Assumption 5.3 when $j=1$, (5.51) and Assumption 5.2, we have

$$\varsigma_1(t,\varepsilon,\delta) \leq C|Y^\varepsilon(t_\delta)-X^0(t)|^{2\alpha}(1+|Y^\varepsilon(t_\delta)|^2). \tag{5.58}$$

By (5.50) and Assumption 5.2, one gets

$$\varsigma_2(t,\varepsilon,\delta) \leq C\int_0^T |Y^\varepsilon(t_\delta)-X^0(t)|^2(1+|Y^\varepsilon(t_\delta)|^2), \tag{5.59}$$

where we use the fact from Assumption 5.3

$$\|(\nabla_\theta b)^*(X^0(t),\theta)\| \leq C(1+|X^0(t)|). \tag{5.60}$$

By (5.60), (5.51) and $\|\nabla\sigma\|_{\mathrm{HS}}<\infty$ in Assumption 5.2, it follows that

$$\varsigma_3(t,\varepsilon,\delta) \leq C\int_0^T |Y^\varepsilon(t_\delta)-X^0(t)|^2(1+X^0(t)|^4).$$

So, in what follows, it remains to show

$$\begin{aligned}
&\mathbb{P}\Big(\int_0^T \|\Upsilon(Y^\varepsilon(t_\delta),\theta_0)-\Upsilon(X^0(t),\theta_0)\|^2\mathrm{d}t \geq \rho\Big)\\
&\quad\leq \mathbb{P}\left(\int_0^T \varsigma_1(t,\varepsilon,\delta)\mathrm{d}t \geq \rho/3\right)+\mathbb{P}\left(\int_0^T \varsigma_2(t,\varepsilon,\delta)\mathrm{d}t \geq \rho/3\right)\\
&\qquad+\mathbb{P}\left(\int_0^T \varsigma_3(t,\varepsilon,\delta)\mathrm{d}t \geq \rho/3\right).
\end{aligned} \tag{5.61}$$

First, let us show by the Chebyshev inequality, the properties of expectation and $t_\delta \leq t \leq T$, we imply that

$$\begin{aligned}
&\mathbb{P}\left(\int_0^T \varsigma_1(t,\varepsilon,\delta)\mathrm{d}t \geq \rho/3\right) \\
&\leq \frac{C}{\rho}\int_0^T \mathbb{E}\left[|Y^\varepsilon(t_\delta) - X^0(t)|^{2\alpha}(1+|Y^\varepsilon(t_\delta)|^2)\right]\mathrm{d}t \\
&= \frac{C}{\rho}\int_0^T \mathbb{E}\left[\mathbb{E}\left[|Y^\varepsilon(t_\delta) - X^0(t)|^{2\alpha}(1+|Y^\varepsilon(t_\delta)|^2)\right]\big|\mathcal{F}_{t_\delta}\right]\mathrm{d}t \qquad (5.62)\\
&= \frac{C}{\rho}\int_0^T \mathbb{E}\left[(1+|Y^\varepsilon(t_\delta)|^2)\right]\mathbb{E}\left[|Y^\varepsilon(t_\delta) - X^0(t)|^{2\alpha}\right]\mathrm{d}t \\
&\leq \frac{C}{\rho}\int_0^T (1+\mathbb{E}|Y^\varepsilon(t_\delta)|^2)\left(\mathbb{E}\left[|Y^\varepsilon(t_\delta) - X^0(t)|^2\right]\right)^\alpha \mathrm{d}t \to 0
\end{aligned}$$

as $\varepsilon \to 0$ and $\delta \to 0$. Note that in the last inequality, we use Lemmas 5.13 and 5.16.

Next, for ς_2, we use the same method as (5.62), and from (5.58), Lemmas 5.13 and 5.16 we have

$$\begin{aligned}
&\mathbb{P}\left(\int_0^T \varsigma_2(t,\varepsilon,\delta)\mathrm{d}t \geq \rho/3\right) \\
&\leq \frac{C}{\rho}\int_0^T \mathbb{E}\left[|Y^\varepsilon(t_\delta) - X^0(t)|^2(1+|Y^\varepsilon(t_\delta)|^2)\right]\mathrm{d}t \qquad (5.63)\\
&\leq \frac{C}{\rho}\int_0^T (1+\mathbb{E}|Y^\varepsilon(t_\delta)|^2)\mathbb{E}\left[|Y^\varepsilon(t_\delta) - X^0(t)|^2\right]\mathrm{d}t \to 0
\end{aligned}$$

as $\varepsilon \to 0$, $\delta \to 0$.

We continue to deal with ς_3,

$$\begin{aligned}
&\mathbb{P}\left(\int_0^T \varsigma_3(t,\varepsilon,\delta)\mathrm{d}t \geq \rho/3\right) \\
&\leq \frac{C}{\rho}\int_0^T \mathbb{E}\left[|Y^\varepsilon(t_\delta) - X^0(t)|^2(1+|X^0(t_\delta)|^4)\right]\mathrm{d}t \qquad (5.64)\\
&\leq \frac{C}{\rho}\int_0^T \mathbb{E}[|Y^\varepsilon(t_\delta) - X^0(t)|^2]\mathrm{d}t \to 0
\end{aligned}$$

as $\varepsilon \to 0$, $\delta \to 0$.

Substituting (5.62), (5.63) and (5.64) into the right of (5.61) respectively, the claim (5.57) holds and we have the desired result (5.55). $\square$

Lemma 5.19. *Let Assumptions 5.1, 5.2 and 5.3 hold. Then*

$$(\nabla_\theta^{(2)}\Phi_{n,\varepsilon})(\theta) \to \overline{K}(\theta) := K(\theta) + 2I(\theta) \quad \textit{in probability} \tag{5.65}$$

as $n \to \infty$ and $\varepsilon \to 0$, where $I(\cdot)$ and $K(\cdot)$ are introduced in (5.14) *and* (5.16)*, respectively.*

Proof. By the chain rule, we deduce from (5.56) that

$$\begin{aligned}
&(\nabla_\theta^{(2)}\Phi_{n,\varepsilon})(\theta)\\
&= 2\sum_{k=1}^{n}(\nabla_\theta^{(2)}G)^*(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\circ(\hat{\sigma}(Y^\varepsilon((k-1)\delta))F_k(\theta))\\
&\quad + 2\sum_{k=1}^{n}(\nabla_\theta G)^*(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\hat{\sigma}(Y^\varepsilon((k-1)\delta))(\nabla_\theta F_k)(\theta)\\
&= -2\sum_{k=1}^{n}(\nabla_\theta^{(2)}b)^*(Y^\varepsilon((k-1)\delta),\theta)\circ(\hat{\sigma}(Y^\varepsilon((k-1)\delta))F_k(\theta_0))\\
&\quad - 2\delta\sum_{k=1}^{n}\Big\{(\nabla_\theta^{(2)}b)^*(Y^\varepsilon((k-1)\delta),\theta)\circ(\hat{\sigma}(Y^\varepsilon((k-1)\delta))G(Y^\varepsilon((k-1)\delta),\theta,\theta_0)\\
&\quad - \sum_{k=1}^{n}(\nabla_\theta b)^*(Y^\varepsilon((k-1)\delta),\theta)\hat{\sigma}(Y^\varepsilon((k-1)\delta))(\nabla_\theta b)(Y^\varepsilon((k-1)\delta),\theta)\Big\}\\
&=: \Theta_1(\varepsilon,\delta) + \Theta_2(\varepsilon,\delta).
\end{aligned}$$

Taking Assumption 5.3 and the Chebyshev inequality into consideration, mimicking the argument of (5.48), we obtain that

$$\Theta_1(\varepsilon,\delta) \to 0 \quad \text{in probability} \;\; \text{as } \varepsilon \to 0, \delta \to 0.$$

Observe that

$$\begin{aligned}
\Theta_2(\varepsilon,\delta) &= -2\int_0^T (\nabla_\theta^{(2)}b)^*(Y^\varepsilon(t_\delta),\theta)\circ(\hat{\sigma}(Y^\varepsilon(t_\delta))G(Y^\varepsilon(t_\delta),\theta,\theta_0)\,\mathrm{d}t\\
&\quad + 2\int_0^T (\nabla_\theta b)^*(Y^\varepsilon(t_\delta),\theta)\hat{\sigma}(Y^\varepsilon(t_\delta)(\nabla_\theta b))(Y^\varepsilon(t_\delta),\theta)\mathrm{d}t\\
&= \vartheta_1(\varepsilon,\delta) + \vartheta_2(\varepsilon,\delta).
\end{aligned}$$

Carrying out an analogous argument to derive (5.47), together with Assumption 5.3 and the Chebyshev inequality, we infer that

$$\vartheta_1(\varepsilon,\delta) \to K(\theta) \quad \text{in probability} \;\; \text{as } \varepsilon \to 0, \delta \to 0, \tag{5.66}$$

and that

$$\vartheta_2(\varepsilon,\delta) \to 2I(\theta) \quad \text{in probability} \;\; \text{as } \varepsilon \to 0, \delta \to 0. \tag{5.67}$$

Thus, the desired assertion follows from (5.66) and (5.67) immediately. □

Now we start to finish the argument of Theorem 5.5 on the basis of the previous lemmas.

Proof of Theorem 5.5. The proof can be carried out as in the proof of Theorem 2.2 of [LSS13]. We prefer to clarify this in the thesis. According to the result of Theorem 5.4, there exists a sequence $\eta_{n,\varepsilon} \to 0$ as $n \to \infty$ and $\varepsilon \to 0$ such that $\hat{\theta}_{n,\varepsilon} \in B_{\eta_{n,\varepsilon}}(\theta_0) \subset \Theta$, $\mathbb{P}$-a.s., that is to say,

$$\mathbb{P}\Big(\hat{\theta}_{n,\varepsilon} \in B_{\eta_{n,\varepsilon}}(\theta_0)\Big) \to 1, \quad \text{as} \ \ n \to \infty, \ \ \varepsilon \to 0. \tag{5.68}$$

Then, it is easy to see that

$$(\nabla_\theta \Phi_{n,\varepsilon})(\hat{\theta}_{n,\varepsilon}) = (\nabla_\theta \Phi_{n,\varepsilon})(\theta_0) + F_{n,\varepsilon}(\hat{\theta}_{n,\varepsilon} - \theta_0), \quad \hat{\theta}_{n,\varepsilon} \in B_{\eta_{n,\varepsilon}}(\theta_0) \tag{5.69}$$

with

$$F_{n,\varepsilon} := \int_0^1 (\nabla_\theta^{(2)} \Phi_{n,\varepsilon})(\theta_0 + v(\hat{\theta}_{n,\varepsilon} - \theta_0))\mathrm{d}v, \qquad \hat{\theta}_{n,\varepsilon} \in B_{\eta_{n,\varepsilon}}(\theta_0),$$

owing to the Taylor expansion. In what follows we intend to deduce that

$$F_{n,\varepsilon} \to \overline{K}(\theta_0) \quad \ \mathbb{P} - \text{a.s.} \tag{5.70}$$

as $n \to \infty$ and $\varepsilon \to 0$. Indeed, for $\hat{\theta}_{n,\varepsilon} \in B_{\eta_{n,\varepsilon}}(\theta_0)$,

$$\begin{aligned}
&\|F_{n,\varepsilon} - \overline{K}(\theta_0)\| \\
&\quad \le \|F_{n,\varepsilon} - (\nabla_\theta^{(2)} \Phi_{n,\varepsilon})(\theta_0)\| + \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon})(\theta_0) - \overline{K}(\theta_0)\| \\
&\quad \le \int_0^1 \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon})(\theta_0 + v(\hat{\theta}_{n,\varepsilon} - \theta_0)) - (\nabla_\theta^{(2)} \Phi_{n,\varepsilon})(\theta_0)\|\mathrm{d}v \\
&\qquad + \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon})(\theta_0) - \overline{K}(\theta_0)\| \\
&\quad \le \sup_{\theta \in B_{\eta_{n,\varepsilon}}(\theta_0)} \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon})(\theta) - \overline{K}(\theta)\| + \sup_{\theta \in B_{\eta_{n,\varepsilon}}(\theta_0)} \|\overline{K}(\theta) - \overline{K}(\theta_0)\| \\
&\qquad + 2\|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon})(\theta_0) - \overline{K}(\theta_0)\|,
\end{aligned}$$

where $\overline{K}(\cdot)$ is shown in (5.65). This, together with Lemma 5.19 and the continuity of $\overline{K}(\cdot)$, yields that (5.70) holds. Next, we show the asymptotic distribution of $\hat{\theta}_{n,\varepsilon}$. Let

$$\mathcal{F}_{n,\varepsilon} = \{F_{n,\varepsilon} \text{ is invertible }, \hat{\theta}_{n,\varepsilon} \in B_{\eta_{n,\varepsilon}}(\theta_0)\}.$$

By Lemma 5.19, one gets, for some positive constant c,

$$\mathbb{P}\Big(\sup_{\theta \in B_{\eta_{n,\varepsilon}}(\theta_0)} \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon})(\theta) - \overline{K}(\theta_0)\| \le \frac{c}{2}\Big) \to 1 \tag{5.71}$$

as $n \to \infty$ and $\varepsilon \to 0$. What's more, by following the line of [LSS13, Theorem 2.2], we can deduce that $F_{n,\varepsilon}$ is invertible on the set

$$\Gamma_{n,\varepsilon} := \Big\{ \sup_{\theta \in B_{\eta_{n,\varepsilon}}(\theta_0)} \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon})(\theta) - \overline{K}(\theta_0)\| \le \frac{c}{2}, \quad \hat{\theta}_{n,\varepsilon} \in B_{\eta_{n,\varepsilon}}(\theta_0) \Big\}.$$

Clearly,

$$\begin{aligned} 1 \ge \mathbb{P}(\Gamma_{n,\varepsilon}) \ge & \mathbb{P}\Big(\sup_{\theta \in B_{\eta_{n,\varepsilon}}(\theta_0)} \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon})(\theta) - K_0(\theta_0)\| \le \frac{c}{2} \Big) \\ & + \mathbb{P}\Big(\hat{\theta}_{n,\varepsilon} \in B_{\eta_{n,\varepsilon}}(\theta_0) \Big) - 1. \end{aligned} \tag{5.72}$$

Thus, taking advantage of (5.71), (5.68) as well as (5.72), we deduce

$$\mathbb{P}(\mathcal{F}_{n,\varepsilon}) \ge \mathbb{P}(\Gamma_{n,\varepsilon}) \to 1 \quad \text{as} \quad n \to \infty, \quad \varepsilon \to 0. \tag{5.73}$$

Let

$$U_{n,\varepsilon} = F_{n,\varepsilon} \mathbf{1}_{\mathcal{F}_{n,\varepsilon}} + I_p \mathbf{1}_{(\mathcal{F}_{n,\varepsilon})^c} \ ,$$

where I_p is a $p \times p$ identity matrix. It follows from (5.69) that

$$\begin{aligned} & \varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0) \\ & =(\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0))\mathbf{1}_{\mathcal{F}_{n,\varepsilon}} + (\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0))\mathbf{1}_{(\mathcal{F}_{n,\varepsilon})^c} \\ & =(U_{n,\varepsilon})^{-1} F_{n,\varepsilon} (\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0))\mathbf{1}_{\mathcal{F}_{n,\varepsilon}} + (\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0))\mathbf{1}_{(\mathcal{F}_{n,\varepsilon})^c} \\ & =\varepsilon^{-1}(U_{n,\varepsilon})^{-1}\{(\nabla_\theta \Phi_{n,\varepsilon})(\hat{\theta}_{n,\varepsilon}) - (\nabla_\theta \Phi_{n,\varepsilon})(\theta_0)\}\mathbf{1}_{\mathcal{F}_{n,\varepsilon}} \\ & \quad + (\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0))\mathbf{1}_{(\mathcal{F}_{n,\varepsilon})^c} \\ & = - \varepsilon^{-1}(U_{n,\varepsilon})^{-1}(\nabla_\theta \Phi_{n,\varepsilon})(\theta_0)\mathbf{1}_{\mathcal{F}_{n,\varepsilon}} + (\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0))\mathbf{1}_{(\mathcal{F}_{n,\varepsilon})^c} \\ & \to I^{-1}(\theta_0) \int_0^T \Upsilon(X^0(t), \theta_0) \mathrm{dW}(t) \qquad \text{as} \ \ n \to \infty, \quad \varepsilon \to 0, \end{aligned}$$

where in the fourth step we have used Fermat's lemma and dropped the term $(\nabla_\theta \Phi_{n,\varepsilon})(\hat{\theta}_{n,\varepsilon})$, and in the last step we have utilized Lemma 5.18, (5.65), (5.70), and (5.73). The desired conclusion is obtained. □

5.3.2 Proof of Theorem 5.11

Theorem 5.11 extends the LSE from SDEs to SFDEs. Based on the proof of consistency and asymptotic normal properties in Subsection 5.3.1, we only need to prove the following necessary lemmas for establishing Theorem 5.11.

Lemma 5.20. *Let Assumptions 5.1, 5.2 and 5.4 be satisfied, and let $p \ge 2$, $\xi \in L^p([-\tau, 0]; \mathbb{R}^d)$. Then there exists a $C_T > 0$ such that*

$$\mathbb{E}||Y_t^\varepsilon||_{\mathscr{C}}^p \le C_T(1 + \mathbb{E}\|\xi\|_{\mathscr{C}}^p)e^{(1+\varepsilon^2)} < \infty, \quad t \in [0, T]. \tag{5.74}$$

Proof. By the Itô formula, we have

$$\begin{aligned}
&(1+|Y^{\varepsilon}(t)|^2)^{\frac{p}{2}}\\
&= (1+|\xi(0)|^2)^{\frac{p}{2}} + p\int_0^t (1+|Y^{\varepsilon}(s)|^2)^{\frac{p-2}{2}}(Y^{\varepsilon}(s))^*\{b(Y^{\varepsilon}(s_\delta),\theta)+Z(\hat{Y}^{\varepsilon}_s)\}\mathrm{d}s\\
&\quad + \frac{p\varepsilon^2}{2}\int_0^t (1+|Y^{\varepsilon}(s)|^2)^{\frac{p-2}{2}}|\sigma(Y^{\varepsilon}(s_\delta))|^2\mathrm{d}s\\
&\quad + \frac{p(p-2)\varepsilon^2}{2}\int_0^t (1+|Y^{\varepsilon}(s)|^2)^{\frac{p-4}{2}}|Y^{\varepsilon}(s)\sigma(Y^{\varepsilon}(s_\delta))|^2\mathrm{d}s \qquad (5.75)\\
&\quad + \varepsilon p\int_0^t (1+|Y^{\varepsilon}(s)|^2)^{\frac{p-2}{2}}(Y^{\varepsilon}(s))^*\sigma(Y^{\varepsilon}(s_\delta))\mathrm{d}W(s)\\
&\le 2^{\frac{p-2}{2}}(1+|\xi(0)|^p) + p\int_0^t (1+|Y^{\varepsilon}(s)|^2)^{\frac{p-2}{2}}\times\Big(\frac{\sqrt{M}}{2}|Y^{\varepsilon}(s)|^2\\
&\quad + \frac{1}{2\sqrt{M}}|b(Y^{\varepsilon}(s_\delta),\theta)+Z(\hat{Y}^{\varepsilon}_s)|^2 + \frac{(p-1)\varepsilon^2}{2}|\sigma(Y^{\varepsilon}(s_\delta))|^2\Big)\mathrm{d}s\\
&\quad + \varepsilon p\int_0^t (1+|Y^{\varepsilon}(s)|^2)^{\frac{p-2}{2}}(Y^{\varepsilon}(s))^*\sigma(Y^{\varepsilon}(s_\delta))\mathrm{d}W(s),
\end{aligned}$$

where in the last inequality we use the elementary inequality and the Young inequality.

By using the elementary inequality, Assumption 5.1, (5.22) and the Young inequality, we have

$$\begin{aligned}
|b(Y^{\varepsilon}(s_\delta),\theta)|^2 &\le |b(Y^{\varepsilon}(s_\delta),\theta)-b(0,\theta)|^2+|b(0,\theta)|^2\\
&\le \|Y^{\varepsilon}_s\|^{2\alpha}_{\mathscr{C}} + C \le \alpha\|Y^{\varepsilon}_s\|^2_{\mathscr{C}} + (1-\alpha) + C \le C(1+\|Y^{\varepsilon}_s\|^2_{\mathscr{C}}).
\end{aligned} \tag{5.76}$$

Similarly, by using the elementary inequality, Assumption 5.2, (5.22), we get

$$\|\sigma(Y^{\varepsilon}(s_\delta))\|^2_{\mathrm{HS}} \le C(1+\|Y^{\varepsilon}_s\|^2_{\mathscr{C}}). \tag{5.77}$$

Carrying out an analogous argument to derive (5.77), together with Assumption 5.4 and (5.23), we derive

$$|Z(\hat{Y}^{\varepsilon}_s)|^2 \le C(1+||\hat{Y}^{\varepsilon}_s||^2_{\mathscr{C}}) \le C(1+||Y^{\varepsilon}_s||^2_{\mathscr{C}}). \tag{5.78}$$

By preparing (5.76), (5.77) and (5.78), there exists a constant $M>0$ such that

$$\begin{aligned}
&p\int_0^t (1+|Y^{\varepsilon}(s)|^2)^{\frac{p-2}{2}}\times\Big(\frac{\sqrt{M}}{2}|Y^{\varepsilon}(s)|^2 + \frac{1}{2\sqrt{M}}|b(Y^{\varepsilon}(s_\delta),\theta)+Z(\hat{Y}^{\varepsilon}_s)|^2\\
&\quad + \frac{(p-1)\varepsilon^2}{2}|\sigma(Y^{\varepsilon}(s_\delta))|^2\Big)\mathrm{d}s
\end{aligned}$$

$$\leq p\int_0^t \frac{\sqrt{M}}{2}(1+||Y_s^\varepsilon||_{\mathscr{C}}^2)^{\frac{p}{2}}\mathrm{d}s + Cp\int_0^t \frac{1}{\sqrt{M}}(1+||Y_s^\varepsilon||_{\mathscr{C}}^2)^{\frac{p}{2}}\mathrm{d}s + Cp\int_0^t \frac{(p-1)\varepsilon^2}{2}(1+||Y_s^\varepsilon||_{\mathscr{C}}^2)^{\frac{p}{2}}\mathrm{d}s \leq C(1+\varepsilon^2)\int_0^t(1+||Y_s^\varepsilon||_{\mathscr{C}}^2)^{\frac{p}{2}}\mathrm{d}s. \tag{5.79}$$

Thus, substituting (5.79) into (5.75), it holds that

$$\begin{aligned}\mathbb{E}\Big(\sup_{0\leq s\leq t}(1+|Y^\varepsilon(s)|^2)^{\frac{p}{2}}\Big) &\leq 2^{\frac{p-2}{2}}(1+\mathbb{E}||\xi||_{\mathscr{C}}^p) + C(1+\varepsilon^2)\mathbb{E}\int_0^t(1+||Y_s^\varepsilon||_{\mathscr{C}}^2)^{\frac{p}{2}}\mathrm{d}s \\ &\quad + \varepsilon p\mathbb{E}\Big(\sup_{0\leq s\leq t}\int_0^s(1+|Y^\varepsilon(s)|^2)^{\frac{p-2}{2}}(Y^\varepsilon(s))^*\sigma(Y^\varepsilon(s_\delta))\mathrm{d}W(s)\Big).\end{aligned} \tag{5.80}$$

For the last term in (5.80) by using the BDG inequality and the Young inequality, we derive that

$$\begin{aligned}&\varepsilon p\mathbb{E}\Big(\sup_{0\leq s\leq t}\int_0^s(1+|Y^\varepsilon(s)|^2)^{\frac{p-2}{2}}(Y^\varepsilon(s))^*\sigma(Y^\varepsilon(s_\delta))\mathrm{d}W(s))\Big) \\ &\leq 4\sqrt{2}\varepsilon p\mathbb{E}\Big(\int_0^t(1+|Y^\varepsilon(s)|^2)^{p-2}|Y^\varepsilon(s)\sigma(Y^\varepsilon(s_\delta))|^2\mathrm{d}s\Big)^{\frac{1}{2}} \\ &\leq 4\sqrt{2}\varepsilon p\mathbb{E}\Big\{(\sup_{0\leq s\leq t}(1+|Y^\varepsilon(s)|^2))^{\frac{p}{2}}\int_0^t(1+|Y^\varepsilon(s)|^2)^{\frac{p-4}{2}}|Y^\varepsilon(s)|^2|\sigma(Y^\varepsilon(s_\delta))|^2\mathrm{d}s\Big\}^{\frac{1}{2}} \\ &\leq 4\sqrt{2}\varepsilon p\mathbb{E}\Big\{(\sup_{0\leq s\leq t}(1+|Y^\varepsilon(s)|^2))^{\frac{p}{2}}\int_0^t C(1+||Y_s^\varepsilon||_{\mathscr{C}}^2)^{\frac{p}{2}}\mathrm{d}s\Big\}^{\frac{1}{2}} \\ &\leq \frac{1}{2}\mathbb{E}\Big(\sup_{0\leq s\leq t}(1+|Y^\varepsilon(s)|^2)^{\frac{p}{2}}\Big) + 16\varepsilon^2p^2C\mathbb{E}\int_0^t(1+||Y_s^\varepsilon||_{\mathscr{C}}^2)^{\frac{p}{2}}\mathrm{d}s.\end{aligned} \tag{5.81}$$

Substituting (5.81) into (5.80), one has

$$\begin{aligned}&\mathbb{E}\Big(\sup_{0\leq s\leq t}(1+|Y^\varepsilon(s)|^2)^{\frac{p}{2}}\Big) \\ &\quad\leq 2^{\frac{p}{2}}(1+\mathbb{E}||\xi||_{\mathscr{C}}^p) + C(1+\varepsilon^2)\mathbb{E}\int_0^t(1+||Y_s^\varepsilon||_{\mathscr{C}}^2)^{\frac{p}{2}}\mathrm{d}s.\end{aligned} \tag{5.82}$$

Next, in order to obtain (5.74), we need to deal with the left side of (5.82). It holds that

$$\mathbb{E}(\sup_{-\tau\leq s\leq t}(1+|Y^\varepsilon(s)|^2)^{\frac{p}{2}})$$

$$
\begin{aligned}
&= \mathbb{E}(\sup_{-\tau\le s\le 0}(1+|Y^\varepsilon(s)|^2)^{\frac{p}{2}}) + \mathbb{E}(\sup_{0\le s\le t}(1+|Y^\varepsilon(s)|^2)^{\frac{p}{2}}) \\
&\le \mathbb{E}(1+||\xi||_{\mathscr{C}}^2)^{\frac{p}{2}} + \mathbb{E}(\sup_{0\le s\le t}(1+|Y^\varepsilon(s)|^2)^{\frac{p}{2}}) \qquad (5.83)\\
&\le 2^{\frac{p-2}{2}}(1+\mathbb{E}||\xi||_{\mathscr{C}}^p) + \mathbb{E}(\sup_{0\le s\le t}(1+|Y^\varepsilon(s)|^2)^{\frac{p}{2}}).
\end{aligned}
$$

Substituting (5.82) into the right side of (5.83), we derive

$$
\begin{aligned}
&\mathbb{E}(\sup_{-\tau\le s\le t}(1+|Y^\varepsilon(s)|^2)^{\frac{p}{2}}) \\
&\le 2^{\frac{p-2}{2}}(1+\mathbb{E}||\xi||_{\mathscr{C}}^p) + 2^{\frac{p}{2}}(1+\mathbb{E}||\xi||_{\mathscr{C}}^p) + C(1+\varepsilon^2)\mathbb{E}\int_0^t (1+||Y_s^\varepsilon||_{\mathscr{C}}^2)^{\frac{p}{2}}\mathrm{d}s \\
&\le C(1+\mathbb{E}||\xi||_{\mathscr{C}}^p) + C(1+\varepsilon^2)\int_0^t \mathbb{E}(\sup_{-\tau\le r\le s}(1+|Y^\varepsilon(r)|^2)^{\frac{p}{2}})\mathrm{d}s.
\end{aligned}
$$

An application of the Gronwall inequality implies that

$$
\mathbb{E}(\sup_{-\tau\le s\le t}(1+|Y^\varepsilon(s)|^2)^{\frac{p}{2}}) \le C_T(1+\mathbb{E}\|\xi\|_{\mathscr{C}}^p)e^{(1+\varepsilon^2)},
$$

and the desired assertion (5.74) follows. The proof is complete. □

Lemma 5.21. *Assume that Assumptions 5.1, 5.2 and 5.4 hold. Then, there exists a constant $C_T > 0$ such that*

$$
\sup_{0\le t\le T}\mathbb{E}\|\hat{Y}_t^\varepsilon - X_t^0\|_{\mathscr{C}}^2 \le C_T\left(\delta + (\delta^\alpha + \delta(1+\varepsilon^2+\varepsilon^4))\mathrm{e}^{(1+\varepsilon^2)} + (\varepsilon^2+\varepsilon^4)\right).
$$

Proof. The proof uses a similar idea of Lemma 5.16 with a little modification. By using the elementary inequality, we have

$$
\mathbb{E}\|\hat{Y}_t^\varepsilon - X_t^0\|_{\mathscr{C}}^2 \le 3\mathbb{E}\|\hat{Y}_t^\varepsilon - Y_t^\varepsilon\|_{\mathscr{C}}^2 + 3\mathbb{E}\|Y_t^\varepsilon - X_t^\varepsilon\|_{\mathscr{C}}^2 + 3\mathbb{E}\|X_t^\varepsilon - X_t^0\|_{\mathscr{C}}^2. \qquad (5.84)
$$

To proceed, we begin to estimate the first one, $\mathbb{E}\|\hat{Y}_t^\varepsilon - Y_t^\varepsilon\|_{\mathscr{C}}^2$.

The first estimation: by Lemma 5.20 with $p=2$, there exists a constant $C>0$ such that

$$
\mathbb{E}\|Y_t^\varepsilon\|_{\mathscr{C}}^2 \le C < \infty. \qquad (5.85)
$$

This, together with using the elementary inequality, the Hölder inequality and the BDG inequality, yields that

$$
\begin{aligned}
&\mathbb{E}\|Y_t^\varepsilon - \hat{Y}_t^\varepsilon\|_{\mathscr{C}}^2 \\
&= \mathbb{E}\left(\sup_{t-\tau\le s\le t}|Y^\varepsilon(s) - Y^\varepsilon(s\wedge t_\delta)|^2\right) \\
&= \mathbb{E}\left(\sup_{t-\tau\le s\le t}|Y^\varepsilon(s) - Y^\varepsilon(t_\delta)|^2\mathbf{1}_{\{s\ge t_\delta\}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left(\sup_{t-\tau\le s\le t}\left|\int_{t_\delta}^s b(Y^\varepsilon(u),\theta)\mathrm{d}u + \int_{t_\delta}^s Z(\hat{Y}_u^\varepsilon)\mathrm{d}u + \int_{t_\delta}^s \sigma(Y^\varepsilon(u))\mathrm{d}W(u)\right|^2\right) \quad (5.86)\\
&\le C\left\{\delta\left(\int_{t_\delta}^t \mathbb{E}|b(Y^\varepsilon(u),\theta)|^2\mathrm{d}u + \int_{t_\delta}^t \mathbb{E}|Z(\hat{Y}_u^\varepsilon)|^2\mathrm{d}u\right) + \int_{t_\delta}^t \mathbb{E}\|\sigma(Y^\varepsilon(u))\|_{\mathrm{HS}}^2\mathrm{d}u\right\}\\
&\le C\left\{\delta\int_{t_\delta}^t (1+\mathbb{E}\|Y_u^\varepsilon\|_{\mathscr{C}}^2)\mathrm{d}u + \int_{t_\delta}^t (1+\mathbb{E}\|Y_u^\varepsilon\|_{\mathscr{C}}^2)\mathrm{d}u\right\}\\
&\le C(\delta^2+\delta) \le C\delta,
\end{aligned}
$$

where in the second inequality we use (5.76), (5.77) and (5.78).

The second estimation: from (5.19) and (5.21), by the Itô formula and (5.27), we have

$$
\begin{aligned}
&\mathrm{d}V^\lambda(X^\varepsilon(t))\\
&\quad = \left\{\lambda u^\lambda(X^\varepsilon(t)) + \nabla V^\lambda(X^\varepsilon(t))Z(X_t^\varepsilon)\right\}\mathrm{d}t + \varepsilon\nabla V^\lambda(X^\varepsilon(t))\sigma(X^\varepsilon(t))\mathrm{d}W(t)
\end{aligned}
\tag{5.87}
$$

and

$$
\begin{aligned}
&\mathrm{d}V^\lambda(Y^\varepsilon(t))\\
&\quad = \lambda u^\lambda(Y^\varepsilon(t))\mathrm{d}t + \nabla u^\lambda(Y^\varepsilon(t))[b(Y^\varepsilon(t_\delta),\theta) - b(Y^\varepsilon(t),\theta)]\mathrm{d}t\\
&\qquad + \nabla V^\lambda(Y^\varepsilon(t))Z(\hat{Y}_t^\varepsilon)\mathrm{d}t + \varepsilon\nabla V^\lambda(Y^\varepsilon(t))\sigma(Y^\varepsilon(t_\delta))\mathrm{d}W(t)\\
&\qquad + \frac{1}{2}\varepsilon^2\mathrm{Tr}\Big([\sigma(Y^\varepsilon(t_\delta))\sigma^*(Y^\varepsilon(t_\delta)) - \sigma(Y^\varepsilon(t))\sigma^*(Y^\varepsilon(t))]\nabla^2u^\lambda(Y^\varepsilon(t)\Big)\mathrm{d}t,
\end{aligned}
\tag{5.88}
$$

where $V^\lambda(\cdot)$ is defined in (5.29). Taking the integral and expectation on both sides of (5.87) and (5.88), using the Hölder inequality and the BDG inequality, we have

$$
\begin{aligned}
&\mathbb{E}|V^\lambda(Y^\varepsilon(t)) - V^\lambda(X^\varepsilon(t))|^2\\
&\le CT\int_0^t \mathbb{E}|u^\lambda(Y^\varepsilon(s)) - u^\lambda(X^\varepsilon(s))|^2\mathrm{d}s\\
&\quad + CT\int_0^t \mathbb{E}|\nabla V^\lambda(Y^\varepsilon(s))\left(b(Y^\varepsilon(s_\delta),\theta) - b(Y^\varepsilon(s),\theta)\right)|^2\mathrm{d}s\\
&\quad + C\varepsilon^2\int_0^t \mathbb{E}\|\nabla V^\lambda(Y^\varepsilon(s))\sigma(Y^\varepsilon(s_\delta)) - \nabla V^\lambda(X^\varepsilon(s))\sigma(X^\varepsilon(s))\|_{\mathrm{HS}}^2\mathrm{d}s \quad (5.89)\\
&\quad + CT\varepsilon^4\int_0^t \mathbb{E}\Big|\mathrm{Tr}\Big([\sigma(Y^\varepsilon(s_\delta)\sigma^*(Y^\varepsilon(s_\delta)) - \sigma(Y^\varepsilon(s))\sigma^*(Y^\varepsilon(s))]\nabla^2u^\lambda(Y^\varepsilon(s)\Big)\Big|^2\mathrm{d}s\\
&\quad + CT\int_0^t \mathbb{E}|\nabla V^\lambda(Y^\varepsilon(s))Z(\hat{Y}_s^\varepsilon) - \nabla V^\lambda(X^\varepsilon(s))Z(X_s^\varepsilon)|^2\mathrm{d}s\\
&=: \sum_{i=1}^5 \Lambda_i(t).
\end{aligned}
$$

Next, we shall estimate $\Lambda_i(t)$, $i = 1, 2, 3, 4, 5$. First, for $\Lambda_i(t)$, $i = 1, 2, 3, 4$, using the same technology as the argument of $G_i(t)$, $i = 1, 2, 3, 4$ in Lemma 5.16, it holds that

$$\Lambda_1(t) \le CT \int_0^t \mathbb{E}|Y^\varepsilon(s) - X^\varepsilon(s)|^2 \mathrm{d}s \le CT \int_0^t \mathbb{E}\|Y_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s, \tag{5.90}$$

$$\Lambda_2(t) \le CT \int_0^t (\mathbb{E}|Y^\varepsilon(s_\delta) - Y^\varepsilon(s)|^2)^\alpha \mathrm{d}s \le CT \int_0^t (\mathbb{E}\|\hat{Y}_s^\varepsilon - Y_s^\varepsilon\|_{\mathscr{C}}^2)^\alpha \mathrm{d}s, \tag{5.91}$$

$$\begin{aligned}\Lambda_3(t) &\le C\varepsilon^2 \int_0^t (\mathbb{E}|Y^\varepsilon(s) - X^\varepsilon(s)|^2 + \mathbb{E}|Y^\varepsilon(s_\delta) - Y^\varepsilon(s)|^2) \mathrm{d}s \\ &\le C\varepsilon^2 \int_0^t \left(\mathbb{E}\|Y_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 + \mathbb{E}\|\hat{Y}_s^\varepsilon - Y_s^\varepsilon\|_{\mathscr{C}}^2\right) \mathrm{d}s\end{aligned} \tag{5.92}$$

and

$$\Lambda_4(t) \le CT\varepsilon^4 \int_0^t \mathbb{E}|Y^\varepsilon(s_\delta) - Y^\varepsilon(s)|^2 \mathrm{d}s \le CT\varepsilon^4 \int_0^t \mathbb{E}\|\hat{Y}_s^\varepsilon - Y_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s, \tag{5.93}$$

respectively. We are left to estimate $\Lambda_5(t)$. To do it, by using the elementary inequality, the Taylor expansion, (5.28) and Assumption 5.4, we have

$$\begin{aligned}\Lambda_5(t) &\le CT \int_0^t \mathrm{E}\left|\left(\nabla V^\lambda(Y^\varepsilon(s)) - \nabla V^\lambda(X^\varepsilon(s))\right) Z(\hat{Y}_s^\varepsilon)\right|^2 \mathrm{d}s \\ &\quad + CT \int_0^t \mathbb{E}\left|\nabla V^\lambda(X^\varepsilon(s)) \left(Z(\hat{Y}_s^\varepsilon) - Z(X_s^\varepsilon)\right)\right|^2 \mathrm{d}s \\ &\le CT \int_0^t \mathbb{E}|Y^\varepsilon(s) - X^\varepsilon(s)|^2 \mathrm{d}s + CT \int_0^T \mathrm{E}\|\hat{Y}_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s \\ &\le CT \int_0^t \mathbb{E}\|Y_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s + CT \int_0^T \mathrm{E}\|\hat{Y}_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s.\end{aligned} \tag{5.94}$$

In conclusion, combining (5.90)-(5.94) into (5.89), we have

$$\begin{aligned}&\mathbb{E}|V^\lambda(Y^\varepsilon(t)) - V^\lambda(X^\varepsilon(t))|^2 \\ &\quad \le C(T + \varepsilon^2) \int_0^t \mathbb{E}\|Y_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s + C \int_0^t (\mathbb{E}|\hat{Y}_s^\varepsilon - Y_s^\varepsilon\|_{\mathscr{C}}^2)^\alpha \mathrm{d}s \\ &\qquad + C(\varepsilon^2 + T\varepsilon^4) \int_0^t \mathbb{E}\|\hat{Y}_s^\varepsilon - Y_s^\varepsilon\|_{\mathscr{C}}^2 + CT \int_0^T \mathbb{E}\|\hat{Y}_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s \\ &\quad \le C(T + \varepsilon^2) \int_0^t \mathbb{E}\|Y_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s + C \int_0^t (\mathbb{E}|\hat{Y}_s^\varepsilon - Y_s^\varepsilon\|_{\mathscr{C}}^2)^\alpha \mathrm{d}s \\ &\qquad + C(T + \varepsilon^2 + T\varepsilon^4) \int_0^t \mathbb{E}\|\hat{Y}_s^\varepsilon - Y_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s,\end{aligned}$$

where in the last line we use the fact that

$$\mathbb{E}\|\hat{Y}_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 \le 2\mathbb{E}\|\hat{Y}_s^\varepsilon - Y_s^\varepsilon\|_{\mathscr{C}}^2 + 2\mathbb{E}\|Y_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2.$$

Moreover, it follows from (5.42) and (5.86) that

$$\begin{aligned}
&\mathbb{E}\|Y_t^\varepsilon - X_t^\varepsilon\|_{\mathscr{C}}^2 \\
&\quad \le C(T+\varepsilon^2)\int_0^t \mathbb{E}\|Y_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s + C\int_0^t (\mathbb{E}|\hat{Y}_s^\varepsilon - Y_s^\varepsilon\|_{\mathscr{C}}^2)^\alpha \mathrm{d}s \\
&\qquad + C(T+\varepsilon^2+T\varepsilon^4)\int_0^t \mathbb{E}\|\hat{Y}_s^\varepsilon - Y_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s \\
&\quad \le C(T+\varepsilon^2)\int_0^t \mathbb{E}\|Y_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s + CT\delta^\alpha + CT\delta^\alpha(T+\varepsilon^2+T\varepsilon^4) \\
&\quad \le C_T(1+\varepsilon^2)\int_0^t \mathbb{E}\|Y_s^\varepsilon - X_s^\varepsilon\|_{\mathscr{C}}^2 \mathrm{d}s + C_T(\delta^\alpha + \delta(1+\varepsilon^2+\varepsilon^4)).
\end{aligned}$$

By using the Gronwall inequality, one has

$$\mathbb{E}\|Y_t^\varepsilon - X_t^\varepsilon\|_{\mathscr{C}}^2 \le C_T(\delta^\alpha + \delta(1+\varepsilon^2+\varepsilon^4))\mathrm{e}^{(1+\varepsilon^2)}. \tag{5.95}$$

The third estimation: we claim that

$$\mathbb{E}\|X_t^\varepsilon - X_t^0\|_{\mathscr{C}}^2 \le C_T(\varepsilon^2+\varepsilon^4). \tag{5.96}$$

The proof of the third estimation is similar to the idea of the second estimation, so we omit it here.

Therefore, substituting (5.86), (5.95) and (5.96) into (5.84), we have

$$\mathbb{E}\|\hat{Y}_t^\varepsilon - X_t^0\|_{\mathscr{C}}^2 \le C_T\left(\delta + (\delta^\alpha + \delta(1+\varepsilon^2+\varepsilon^4))\mathrm{e}^{(1+\varepsilon^2)} + (\varepsilon^2+\varepsilon^4)\right).$$

The proof was completed. □

Proof of Theorem 5.11. Theorem 5.11 mainly states the results about the consistency and asymptotic normality of the least squares estimator $\hat{\theta}_{n,\varepsilon}^Z$ obtained from SFDEs when $n \to \infty$ and $\varepsilon \to 0$. In this proof, we need to use technical Lemmas 5.20 and 5.21. For the specific idea of proof, please refer to Theorems 5.4 and 5.5. □

5.4 Appendix

In this section, we will further explain how to understand the symbol "$\circ$" defined in (5.1) and appearing in the proof of Lemma 5.19; see, [RW21].

For a differentiable function $V(x) = (V_1(x), \ldots, V_d(x))^* : \mathbb{R}^m \to \mathbb{R}^d$, define its gradient operator $(\nabla_x V)(x) \in \mathbb{R}^d \otimes \mathbb{R}^m$ with respect to $x = (x_1, \ldots, x_m)^* \in \mathbb{R}^m$ by

$$(\nabla_x V)(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} V_1(x) & \frac{\partial}{\partial x_2} V_1(x) & \cdots & \frac{\partial}{\partial x_m} V_1(x) \\ \frac{\partial}{\partial x_1} V_2(x) & \frac{\partial}{\partial x_2} V_2(x) & \cdots & \frac{\partial}{\partial x_m} V_2(x) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial}{\partial x_1} V_d(x) & \frac{\partial}{\partial x_2} V_d(x) & \cdots & \frac{\partial}{\partial x_m} V_d(x) \end{pmatrix},$$

which enjoys the property $\nabla_x V^*(x) = (\nabla_x V)^*(x)$.

For a matrix-valued function $V(x) = (V_{ij}(x))_{m\times d} : \mathbb{R} \to \mathbb{R}^m \otimes \mathbb{R}^d$ be differentiable, its derivative $\frac{\partial}{\partial x} V(x) \in \mathbb{R}^m \otimes \mathbb{R}^d$ with respect to $x \in \mathbb{R}$ admits the form

$$\frac{\partial}{\partial x} V(x) = \begin{pmatrix} \frac{\partial}{\partial x} V_{11}(x) & \frac{\partial}{\partial x} V_{12}(x) & \cdots & \frac{\partial}{\partial x} V_{1d}(x) \\ \frac{\partial}{\partial x} V_{21}(x) & \frac{\partial}{\partial x} V_{22}(x) & \cdots & \frac{\partial}{\partial x} V_{2d}(x) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial}{\partial x} V_{m1}(x) & \frac{\partial}{\partial x} V_{m2}(x) & \cdots & \frac{\partial}{\partial x} V_{md}(x) \end{pmatrix}. \tag{5.97}$$

If $V(x) = (V_{ij}(x))_{m\times d} : \mathbb{R}^m \to \mathbb{R}^m \otimes \mathbb{R}^d$ is differentiable, its gradient operator $(\nabla_x V)(x) \in \mathbb{R}^m \otimes \mathbb{R}^{md}$ with respect to the variable $x = (x_1, \ldots, x_m)^* \in \mathbb{R}^m$ is written as

$$(\nabla_x V)(x) = \Big(\frac{\partial}{\partial x_1} V(x), \frac{\partial}{\partial x_2} V(x), \cdots, \frac{\partial}{\partial x_m} V(x)\Big),$$

where $\frac{\partial}{\partial x_i} V(x)$ is defined as in (5.97). Hence, there exists a matrix $A = (A_1, A_2, \ldots, A_p) \in \mathbb{R}^p \otimes \mathbb{R}^{pd}$ with $A_k \in \mathbb{R}^p \otimes \mathbb{R}^d$, $k = 1, 2, \ldots, p$, and $B \in \mathbb{R}^d$, we can define $A \circ B \in \mathbb{R}^p \otimes \mathbb{R}^p$ by

$$A \circ B = (A_1 B, A_2 B, \ldots, A_p B).$$

Chapter 6

Parameter estimations for McKean-Vlasov SDEs

This chapter aims to further solve the problem of parameter estimation for delay McKean-Vlasov stochastic differential equations (SDEs) with the coefficient exhibiting super-linear growth in the state component. Specifically, we propose a least squares estimator for an unknown parameter in the drift of a delay McKean-Vlasov SDEs with a small noise dispersion parameter by making use of time-discretized interacting particle systems and proving the weak convergence between the estimator and the true value, under suitable conditions. To achieve our main purposes on weak convergence, we give the approximation of the distribution of delay McKean-Vlasov SDEs at the discrete points and take advantage of calculating skills on the space of probability measures with finite order moments. Moreover, the asymptotic distribution of the least squares estimator is derived via the properties of solutions for the corresponding interacting particle systems.

This chapter is based on

[ZH22] Min Zhu and Yanyan Hu. Least squares estimation for delay McKean–Vlasov stochastic differential equations and interacting particle systems. *Commun. Math. Sci.*, 20(1):265–296, 2022

6.1 Introduction

The evolution of numerous stochastic systems depends not only on the microcosmic state of the particles, but also on the macrocosmic distribution of the particles. The McKean-Vlasov SDE is a kind of mathematical model, which can characterize the evolution of those stochastic systems. The pioneering work on McKean-Vlasov SDEs is initiated in [McK66], and McKean studied the propagation of chaos in physical systems of N-interacting particles related to Boltzmann's model for the statistical mechanics of rarefied gases in [McK67]. More concretely, McKean-Vlasov SDEs are a special class of SDEs, where the coefficients involved depend not only on the state process but also on their distribution. In response to the great needs, as a hot but difficult research topic, they have important application value in the fields of stochastic control, insurance, and mathematical finance, to name a few; see, for instance, [BLM17, CD18]. McKean-Vlasov SDEs have been extensively investigated by many authors, and various results on well-posedness, Harnack inequalities, Bismut formula, ergodicity, and other quantitative and qualitative properties have been proposed (e.g.[RZ21, Wan18, RW19a, EGZ19]). In contrast to the general McKean-Vlasov SDEs, there has not been much research on path-dependent McKean-Vlasov SDEs, but these have begun to gain attention recently. For works on well-posedness and Harnack-type inequalities, we refer to [HRW19, Hua21]. Huang and Yuan [HY21b] showed the existence and uniqueness of strong solutions to distribution-dependent neutral SFDEs and gave the comparison theorem of these equations. Most of the previous works are concerned with path-dependent McKean-Vlasov SDEs which do not contain unknown parameters. However, in many practical applications, these models may contain unknown parameters. Hence, we want to estimate deterministic quantities of these unknown parameters for SDEs, especially, path-dependent McKean-Vlasov SDEs.

Based on discrete and continuous time observations, there have been a number of attempts in the literature to develop methods on the parametric estimation for SDEs; see, e.g., [BBAKP18, BP16, LMS17]. Beyond that, estimation for stochastic delay differential equations (SDDEs) has been studied from various points of view, we refer to Kuechler and Soerensen [KS10], who proposed an estimator of drift parameters for affine stochastic delay differential equations by discretization of the continuous-time likelihood function; Reiss [Rei05] studied the problem of nonparametric estimation for affine SDDEs by continuous observation. Above all, the small diffusion asymptotic of SDEs has been discussed systematically and applied successively to real world problems; see, for instance, the monograph [Kut04] for more details. In general, the parametric estimation relied on continuous-time observations, which is a mathematical idealization, and no measuring device can follow continuously the sample paths of the diffusion processes involved (cf.[RW21]). So, from a practical standpoint in parametric inference, it is more meaningful to explore asymptotic estimation for diffusion processes with small dispersions based on discrete observations. Whereas multiple methods have been proposed, the simplest and most natural solution seems to be the one based on the least squares estimation (LSE) in cases of large-scale scattered data, see, e.g., [Ma10, LSS13, Kas88]. Concerning the LSE under various settings, we refer to, e.g., [LSS13] for SDEs driven by small Lévy noises with Lipschitz condition for the drift term, [Lon10] for SDEs driven by small α-stable noises with Lipschitz coefficients, and [PY19] for the α-stable Ornstein-Uhlenbeck process with a constant drift.

Compared with the general SDEs, the corresponding issues for McKean-Vlasov SDEs are rare. Recent attempts towards parameter inference of McKean-Vlasov SDEs (cf. [RW21, RW19b, WWMX16]) have led to renewed interest in the asymptotic theory of stochastic models. Inspired by their studies, we make a new attempt to study the problem of parameter estimation for delayed McKean-Vlasov SDEs with a small dispersion. Moreover, there is no published LSE for delayed McKean-Vlasov SDEs, to the best of our knowledge. What's more, for the problem of parameter estimation of delay McKean-Vlasov SDEs, the technique used for the general SDEs cannot directly be applied to obtain an asymptotically consistent estimation. This is because the McKean-Vlasov SDEs cannot be solved explicitly. A significant consequence of this fact is that we cannot obtain observations of the distribution of the path at regular space time points directly in most of our arguments. So, whereas the mechanisms of LSE are often relatively simple, caution needs to be exercised when approximating the distribution at every step of the analysis. Indeed, in many situations, due to the complicated dependence structure among discrete points, results from the execution of LSEs may differ considerably from the standard SDEs, affecting both the accuracy and precision of the LSE-based predictions.

References [RW19b, RW21] though, have succeeded in investigating parameter estimation for path-dependent McKean-Vlasov SDEs by an Euler-Maruyama type scheme. In particular, under the monotone condition, [RW21] studied LSE for path-dependent McKean-Vlasov SDEs by using the continuous time tamed EM method. It is worth noting that they simulated the segment process by the linear interpolations between the points on the gridpoints and approximated its distribution directly using the law of the associated segment process. Even so, the distribution cannot be simulated by the computer. Based on the macrocosmic property of the distribution of stochastic systems, we shall investigate parameter estimation for McKean-Vlasov SDEs by using an empirical distribution corresponding to stochastic interacting particle systems to approximate the distributions at each step. This method, based on stochastic interacting particle systems, has been successfully applied to the approximation of McKean-Vlasov SDEs in [BH22]. In the current work, by constructing an appropriate contrast function based on the associated interacting particle systems, we shall provide a new idea to derive the LSE consistency and asymptotic distribution for a class of McKean-Vlasov SDEs. Compared with the existing results in the work, the innovations of the Chapter lie in two aspects:

(a) We introduce stochastic particle systems to simulate delay McKean-Vlasov SDEs, and establish the contrast function;

(b) Our model is more applicable and practical as we are dealing with delay SDEs with superlinear growth coefficients which are distribution dependent.

6.2 Preliminaries and interacting particle systems

Throughout this Chapter, the following notation and terminology will be used. For $m, d \in \mathbb{N}$, the set of all positive integers, let $(\mathbb{R}^d, \langle\cdot,\cdot\rangle, |\cdot|)$ be the d-dimensional Euclidean space with the inner product $\langle\cdot,\cdot\rangle$ inducing the norm $|\cdot|$ and $\mathbb{R}^d \times \mathbb{R}^m$ the collection of all $d \times m$ matrixes with real entries, which is endowed with the Hilbert-Schmidt norm $\|\cdot\|$. $\mathbf{0} \in \mathbb{R}^d$ denotes the zero vector. For a matrix A, A^* denotes the transpose of A. Concerning a square matrix A, A^{-1} means the inverse of A provided that $\det(A) \neq 0$. For $p \in \mathbb{N}$, let Θ be an open bounded convex subset of $\mathbb{R}^p$, and $\overline{\Theta}$ the closure of Θ. For $r > 0$ and $x \in \mathbb{R}^p$, $B_r(x)$ represents the closed ball centered at x with the radius r. $\lfloor a \rfloor$ stands for the integer part of the real number $a \geq 0$. For a random variable ξ, $\mathscr{L}_\xi$ denotes its law. For given $\tau > 0$, $\mathscr{C} := C([-\tau, 0]; \mathbb{R}^d)$ means the family of all continuous functions $\xi : [-\tau, 0] \to \mathbb{R}^d$ with the uniform norm $\|\xi\|_\infty := \sup_{-\tau \leq \theta \leq 0} |f(\theta)|$. For $p > 0$, $\mathscr{P}_p(\mathbb{R}^d)$ stands for the space of all probability measures on $\mathbb{R}^d$ with the finite p-th moment, i.e., $\mu(|\cdot|^p) := \int_{\mathbb{R}^d} |x|^p \mu(\mathrm{d}x) < \infty$ for $\mu \in \mathscr{P}_p(\mathbb{R}^d)$. Define the $\mathbb{W}_p$-Wasserstein distance on $\mathscr{P}_p(\mathbb{R}^d)$ by

$$\mathbb{W}_p(\mu, \nu) = \inf_{\pi \in \mathcal{C}(\mu,\nu)} \Big(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |x - y|^p \pi(\mathrm{d}x, \mathrm{d}y) \Big)^{\frac{1}{1 \vee p}}, \qquad \mu, \nu \in \mathscr{P}_p(\mathbb{R}^d),$$

where $\mathcal{C}(\mu, \nu)$ signifies the set of all couplings of μ and ν. $L_p^0(\mathbb{R}^d)$ denotes the space of $\mathbb{R}^d$-valued, $\mathscr{F}_0$-measurable random variables X with $\mathbb{E}|X|^p < \infty$. Let $(W_t)_{t \geq 0}$ be an m-dimensional Brownian motion defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with the filtration $(\mathscr{F}_t)_{t \geq 0}$ satisfying the usual condition (i.e., $\mathscr{F}_0$ contains all $\mathbb{P}$-null sets and $\mathscr{F}_t = \mathscr{F}_{t+} := \bigcap_{s>t} \mathscr{F}_s$).

For a fixed time horizon $T > 0$ and scale parameter $\varepsilon \in (0, 1)$, we consider a *delay McKean-Vlasov SDE* on $\mathbb{R}^d$

$$\begin{cases} \mathrm{d}X_t^\varepsilon = b(X_t^\varepsilon, X_{t-\tau}^\varepsilon, \mu_t^\varepsilon, \mu_{t-\tau}^\varepsilon, \theta)\mathrm{d}t + \varepsilon\, \sigma(X_t^\varepsilon, X_{t-\tau}^\varepsilon, \mu_t^\varepsilon, \mu_{t-\tau}^\varepsilon)\mathrm{d}W_t, \quad t \in (0, T] \\ X_s^\varepsilon = \xi(s), \;\; s \in [-\tau, 0] \end{cases} \tag{6.1}$$

where $\mu_\cdot^\varepsilon := \mathscr{L}_{X_\cdot^\varepsilon}$ denotes the law of $X_\cdot^\varepsilon$; $b : \mathbb{R}^d \times \mathbb{R}^d \times \mathscr{P}_2(\mathbb{R}^d) \times \mathscr{P}_2(\mathbb{R}^d) \times \Theta \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \times \mathbb{R}^d \times \mathscr{P}_2(\mathbb{R}^d) \times \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}^d \times \mathbb{R}^m$ are continuous. In (6.1), we assume that the drift term b and the diffusion term σ are known apart from the parameter $\theta \in \Theta$. We stipulate that $\theta_0 \in \Theta$ is the true value of $\theta \in \Theta$.

For $i \in S_N := \{1, \cdots, N\}$, $N \geq 1$, let (X_0^i, W_t^i) be i.i.d copies of (X_0, W_t). We introduce the stochastic interacting particle to approximate (6.1). First, for $t \in (0, T]$, consider the following *stochastic non-interacting particle systems* associated with (6.1)

$$\begin{cases} \mathrm{d}X_t^{\varepsilon,i} = b(X_t^{\varepsilon,i}, X_{t-\tau}^{\varepsilon,i}, \mu_t^{\varepsilon,i}, \mu_{t-\tau}^{\varepsilon,i}, \theta)\mathrm{d}t + \varepsilon\, \sigma(X_t^{\varepsilon,i}, X_{t-\tau}^{\varepsilon,i}, \mu_t^{\varepsilon,i}, \mu_{t-\tau}^{\varepsilon,i})\mathrm{d}W_t^i, \\ X_s^{\varepsilon,i} = \xi(s), \quad s \in [-\tau, 0], \quad i \in \mathbf{S}_N, \end{cases} \tag{6.2}$$

where $\mu_\cdot^{\varepsilon,i} := \mathscr{L}_{X_\cdot^{\varepsilon,i}}$ denotes the law of $X_\cdot^{\varepsilon,i}$, $i \in \mathbf{S}_N$. By virtue of the weak uniqueness due to Theorem 6.1, it is easy to see that $\mu_\cdot^\varepsilon = \mu_\cdot^{\varepsilon,i}$, $i \in \mathbf{S}_N$. Let $\tilde{\mu}_\cdot^{\varepsilon,N}$ be the empirical distribution corresponding to $X_\cdot^{\varepsilon,1}, X_\cdot^{\varepsilon,2}, \cdots, X_\cdot^{\varepsilon,N}$, namely,

$$\tilde{\mu}_t^{\varepsilon,N}(dx) = \frac{1}{N}\sum_{j=1}^{N}\delta_{X_t^{\varepsilon,j}}(dx), \qquad t \geq -\tau.$$

Consider the following *deterministic ordinary differential equation*

$$\begin{cases} \mathrm{d}X_t^{0,i} = b(X_t^{0,i}, X_{t-\tau}^{0,i}, \mu_t^0, \mu_{t-\tau}^0, \theta_0)\mathrm{d}t, \quad t > 0, \\ X_s^0 = \xi(s), \quad s \in [-\tau, 0], \quad i \in \mathbf{S}_N \end{cases} \tag{6.3}$$

where $\mu_\cdot^0 = \mu_\cdot^{0,i} := \mathscr{L}_{X_\cdot^{0,i}}$ denotes the law of $X_\cdot^{0,i}$.

Second, for $t \in (0, T]$, *stochastic interacting particle systems* can be described as

$$\begin{cases} \mathrm{d}X_t^{\varepsilon,i,N} = b(X_t^{\varepsilon,i,N}, X_{t-\tau}^{\varepsilon,i,N}, \mu_t^{\varepsilon,N}, \mu_{t-\tau}^{\varepsilon,N}, \theta)\mathrm{d}t + \varepsilon\,\sigma(X_t^{\varepsilon,i,N}, X_{t-\tau}^{\varepsilon,i,N}, \mu_t^{\varepsilon,N}, \mu_{t-\tau}^{\varepsilon,N})\mathrm{d}W_t^i, \\ X_s^{\varepsilon,i,N} = \xi(s), \ s \in [-\tau, 0], \quad i \in \mathbf{S}_N, \end{cases} \tag{6.4}$$

where $\mu_\cdot^{\varepsilon,N}$ stands for the empirical distribution corresponding to $X_\cdot^{\varepsilon,1,N}, \cdots, X_\cdot^{\varepsilon,N,N}$, namely,

$$\mu_t^{\varepsilon,N}(dx) = \frac{1}{N}\sum_{j=1}^{N}\delta_{X_t^{\varepsilon,j,N}}(dx), \qquad t \geq -\tau. \tag{6.5}$$

It is worth pointing out that (6.1), (6.2), (6.3) and (6.4) share the same initial data. Set

$$B(x, y, \theta_0, \theta) := b(x, y, \mu, \nu, \theta_0) - b(x, y, \mu, \nu, \theta)$$

and

$$\Lambda(x, y) := (\sigma\sigma^*)(x, y, \mu, \nu),$$

for any $x,\ y \in \mathbb{R}^d$ and $\mu,\ \nu \in \mathscr{P}(\mathbb{R}^d)$.

For a fixed time horizon $T > 0$, we give a uniform time discretization of $[-\tau, T]$ with mesh-size $\delta = \frac{T}{n} = \frac{\tau}{M} \in (0, 1)$, where n, $M > 1$. In order to approximate the measure $\mu_\cdot^\varepsilon$ and improve the simulation precision of (6.1), by virtue of the interacting particle system (6.4) we construct the following *contrast function*

$$\Psi_{n,\varepsilon}^{i,N}(\theta) = \varepsilon^{-2}\delta^{-1}\sum_{n=1}^{n}(P_k^{\varepsilon,i,N}(\theta))^*\Lambda^{-1}(X_{(k-1)\delta}^{\varepsilon,i,N}, X_{(k-1)\delta-\tau}^{\varepsilon,i,N})P_k^{\varepsilon,i,N}(\theta), \tag{6.6}$$

where

$$P_k^{\varepsilon,i,N}(\theta) = X_{k\delta}^{\varepsilon,i,N} - X_{(k-1)\delta}^{\varepsilon,i,N} - b(X_{(k-1)\delta}^{\varepsilon,i,N}, X_{(k-1)\delta-\tau}^{\varepsilon,i,N}, \mu_{(k-1)\delta}^{\varepsilon,N}, \mu_{(k-1)\delta-\tau}^{\varepsilon,N}, \theta)\delta, \tag{6.7}$$

for $k = 1, 2, \cdots, n$.

According to the principle of the least squares method, to achieve the least squares estimation of $\theta \in \Theta$, we need to seek an argument $\hat{\theta}^{i,N}_{n,\varepsilon} \in \Theta$ such that

$$\Psi^{i,N}_{n,\varepsilon}(\hat{\theta}^{i,N}_{n,\varepsilon}) = \min_{\theta \in \Theta} \Psi^{i,N}_{n,\varepsilon}(\theta), \tag{6.8}$$

namely,

$$\hat{\theta}^{i,N}_{n,\varepsilon} = \arg\min_{\theta \in \Theta} \Psi^{i,N}_{n,\varepsilon}(\theta).$$

Let $\theta_0 \in \Theta$ be the true value of θ and

$$\Phi^{i,N}_{n,\varepsilon}(\theta) = \varepsilon^2(\Psi^{i,N}_{n,\varepsilon}(\theta) - \Psi^{i,N}_{n,\varepsilon}(\theta_0)).$$

Then, from (6.8), one has

$$\hat{\theta}^{i,N}_{n,\varepsilon} = \arg\min_{\theta \in \Theta} \Phi^{i,N}_{n,\varepsilon}(\theta). \tag{6.9}$$

That is to say, $\hat{\theta}^{i,N}_{n,\varepsilon}$ satisfying (6.9) is called LSE of $\theta \in \Theta$.

To obtain the main results, we give the following assumptions. Let $K_i : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ such that

$$K_i(x, y) \le L_i(1 + |x|^{r_i} + |y|^{r_i}), \qquad i = 1, 2, 3, 4, 5, \tag{6.10}$$

for some constants $L_i > 0$, $r_i \ge 1$ and any $x, y \in \mathbb{R}^d$. Furthermore, for any $x_i, y_i \in \mathbb{R}^d$ and $\mu_i, \nu_i \in \mathscr{P}_2(\mathbb{R}^d)$, $i = 1, 2$, we assume that:

Assumption 6.1. *For any $\theta \in \overline{\Theta}$, there exists a $C_1 > 0$ such that*

$$\begin{aligned}\langle x_1 - x_2, b(x_1, y_1, \mu, \nu, \theta) - b(x_2, y_2, \mu, \nu, \theta)\rangle \le& C_1\Big(|x_1 - x_2|^2 + |y_1 - y_2|^2\Big);\\ |b(x, y, \mu_1, \nu_1, \theta) - b(x, y, \mu_2, \nu_2, \theta)| \le& C_1\Big(\mathbb{W}_2(\mu_1, \mu_2) + \mathbb{W}_2(\nu_1, \nu_2)\Big);\\ |b(x_1, y_1, \mu, \nu, \theta) - b(x_2, y_2, \mu, \nu, \theta)| \le& C_1|x_1 - x_2| + K_1(y_1, y_2)|y_1 - y_2|;\end{aligned}$$

Assumption 6.2. *There exists a $C_2 > 0$ such that*

$$\|\sigma(x_1, y_1, \mu, \nu) - \sigma(x_2, y_2, \mu, \nu)\| \le C_2|x_1 - x_2| + K_2(y_1, y_2)|y_1 - y_2|$$

and

$$\|\sigma(x, y, \mu_1, \nu_1) - \sigma(x, y, \mu_2, \nu_2)\| \le C_2\Big(\mathbb{W}_2(\mu_1, \mu_2) + \mathbb{W}_2(\nu_1, \nu_2)\Big);$$

Assumption 6.3. *$\sigma\sigma^*$ is invertible, and there exists a $C_3 > 0$ such that*

$$\begin{aligned}&\|(\sigma\sigma^*)^{-1}(x_1, y_1, \mu_1, \nu_1) - (\sigma\sigma^*)^{-1}(x_2, y_2, \mu_2, \nu_2)\|\\ &\quad\le C_3\Big(|x_1 - x_2| + \mathbb{W}_2(\mu_1, \mu_2) + \mathbb{W}_2(\nu_1, \nu_2)\Big) + K_3(y_1, y_2)|y_1 - y_2|;\end{aligned}$$

Assumption 6.4. *There exists a $C_4 > 0$ such that*

$$\begin{aligned}&\sup_{\theta \in \overline{\Theta}} \|(\nabla_\theta b)(x_1, y_1, \mu_1, \nu_1, \theta) - (\nabla_\theta b)(x_2, y_2, \mu_2, \nu_2, \theta)\|\\ &\quad\le C_4\Big(|x_1 - x_2| + \mathbb{W}_2(\mu_1, \mu_2) + \mathbb{W}_2(\nu_1, \nu_2)\Big) + K_4(y_1, y_2)|y_1 - y_2|,\end{aligned}$$

where $(\nabla_\theta b)$ means the gradient operator w.r.t. the fifth spatial variable;

Assumption 6.5. *There exists a* $C_5 > 0$ *such that*

$$\begin{aligned}
&\sup_{\theta\in\overline{\Theta}} \|(\nabla_\theta^{(2)} b^*)(x_1, y_1, \mu_1, \nu_1, \theta) - (\nabla_\theta^{(2)} b^*)(x_2, y_2, \mu_2, \nu_2, \theta)\| \\
&\quad \le C_5\Big(|x_1 - x_2| + \mathbb{W}_2(\mu_1, \mu_2) + \mathbb{W}_2(\nu_1, \nu_2)\Big) + K_5(y_1, y_2)|y_1 - y_2|,
\end{aligned}$$

where $\nabla_\theta^{(2)} b =: \nabla_\theta(\nabla_\theta b)$*;*

Assumption 6.6. *There exists a constant* $q > p$ *such that*

$$\sup_{0\le t\le T} \mathbb{E}|X_t|^q < \infty.$$

Assumptions 6.1 and 6.2 are used to guarantee the well-posedness of (6.1) and the corresponding stochastic interacting particle system (6.4) (see Theorem 6.1 and Lemma 6.6 below). Besides, Assumptions 6.3 and 6.6 also play an important role in the analysis of the consistency of the LSE for the unknown parameter θ. Assumption 6.6 is set to ensure that strong convergence between stochastic interacting particle systems and non-interacting particle systems in the p-th moment holds, which improves the result in [BH22] on the convergence, in two aspects: First, the Assumption 6.6 is more applicable than the conditions of [BH22, Theorem 1.4]; Second, in the current work we only need Assumption 6.6 to hold for some $q > p$, wherein it is easier to seek a constant q under Assumptions 6.1 and 6.2, and it is not confined to $p > 4$ as [BH22]. Assumptions 6.4 and 6.5 are used to establish the asymptotic distribution of the LSE.

6.3 Main results

Under the framework of non-Lipschitz condition, the tamed Euler scheme is adopted to establish contrast function for LSE in [RW21]. Here we approximate (6.1) by a particle system, and investigate the consistency and asymptotic distribution of the LSE under a super-linear condition. First, we show the following result on the strong well-posedness of (6.1), where the drift and diffusion terms have polynomial growth for the delay variables.

Theorem 6.1. *Assume that Assumptions 6.1 and 6.2 hold, for any initial value* $X_0^\varepsilon = \xi \in L^0_{p_1 q_1}(\mathscr{C})$*, where* p_1 *and* q_1 *will be shown in the proof, then* (6.1) *possesses a unique global strong solution* $(X_t^\varepsilon)_{t\ge-\tau}$ *with*

$$\mathbb{E}\Big(\sup_{-\tau\le t\le T} |X_t^\varepsilon|^p\Big) \le C < \infty, \qquad p \ge 2. \tag{6.11}$$

The strong well-posedness of McKean-Vlasov SDEs under various conditions has been studied largely, e.g., [dRST19] with a drift of polynomial growth, [Wan18] for continuity, monotonicity, and growth of coefficients, [BH22] with Hölder continuous coefficients. In the meantime, strong well-posedness of path-dependent McKean-Vlasov

SDEs has got more and more attention, e.g., [HRW19] under the condition of continuity, monotonicity, and growth of coefficients, [RW19b] for one-side Lipschitz drifts and Lipschitz diffusions. Theorem 6.1 above shows that the delay McKean-Vlasov SDE is strong well-posedness when both the coefficients have super-linear growth.

The second result in the current work shows the consistency of the LSE with high frequency and small dispersion. To display this result, we analyze strong convergence between stochastic interacting particle systems and non-interacting particle systems corresponding to delay McKean-Vlasov SDEs (6.1) whenever the particle number goes to infinity and the stepsize closes to zero. For the sake of simplicity, we set

$$\Pi(\theta) := \int_0^T B^*(X_t^{0,i}, X_{t-\tau}^{0,i}, \theta_0, \theta)\Lambda^{-1}(X_t^{0,i}, X_{t-\tau}^{0,i})B(X_t^{0,i}, X_{t-\tau}^{0,i}, \theta_0, \theta)\mathrm{d}t, \tag{6.12}$$

$$\Upsilon(x, y, \theta_0) := (\nabla_\theta b)^*(x, y, \mu, \nu, \theta_0)\Lambda^{-1}(x, y)\sigma(x, y, \mu, \nu), x, y \in \mathbb{R}^d, \mu, \nu \in \mathscr{P}_p(\mathbb{R}^d) \tag{6.13}$$

and

$$\begin{aligned} I(\theta) := \int_0^T & (\nabla_\theta b)^*(X_t^{0,i}, X_{t-\tau}^{0,i}, \mu_t^{0,i}, \mu_{t-\tau}^{0,i}, \theta)\Lambda^{-1}(X_t^{0,i}, X_{t-\tau}^{0,i}) \\ & \times (\nabla_\theta b)(X_t^{0,i}, X_{t-\tau}^{0,i}, \mu_t^{0,i}, \mu_{t-\tau}^{0,i}, \theta)\mathrm{d}t. \end{aligned} \tag{6.14}$$

Theorem 6.2. *Under Assumptions 6.1, 6.2, 6.3 and 6.6. If, for any $\theta \in \Theta$, $\Pi(\theta) \geq 0$, then*

$$\hat{\theta}_{n,\varepsilon}^{i,N} \to \theta_0 \quad \textit{in probability as} \quad N, n \to \infty \quad \textit{and} \quad \varepsilon \to 0.$$

The last result focuses on the asymptotic distribution of the LSE $\hat{\theta}_{n,\varepsilon}^{i,N}$.

Theorem 6.3. *Under the assumptions of Theorem 6.2, suppose that Assumptions 6.4 and 6.5 hold. Then,*

$$\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon}^{i,N} - \theta_0) \to I^{-1}(\theta_0)\int_0^T \Upsilon(X_t^{0,i}, X_{t-\tau}^{0,i}, \theta_0)\mathrm{d}W_t^i \qquad \mathbb{P} - a.s.$$

as $n, N \to \infty$ and $\varepsilon \to 0$, where $I(\cdot)$ and $\Upsilon(\cdot)$ are continuous.

6.4 An illustrative example

In this section, we intend to provide an example to demonstrate our results. We first give the setup of numerical examples as follows.

Example 6.4. Let $\theta = (\theta^{(1)}, \theta^{(2)})^* \in \Theta_0 := (c_1, c_2) \times (c_3, c_4) \subset \mathbb{R}^2$ for some $c_1 < c_2$ and $c_3 < c_4$. For any $\varepsilon \in (0, 1)$, consider the following delay McKean-Vlasov SDE

$$\begin{aligned} \mathrm{d}X^\varepsilon(t) = & \Big\{\theta^{(1)} + \theta^{(2)}\Big(X_t^\varepsilon - (X_{t-\tau}^\varepsilon)^3 + X_{t-\tau}^\varepsilon + \mathbb{E}X_{t-\tau}^\varepsilon\Big)\Big\}\mathrm{d}t \\ & + \varepsilon\Big\{1 + |X_{t-\tau}^\varepsilon|^3 + |X_{t-\tau}^\varepsilon| + \mathbb{E}|X_{t-\tau}^\varepsilon|\Big\}\mathrm{d}W(t), \quad t \in (0, T] \end{aligned} \tag{6.15}$$

with the initial value $X_0^\varepsilon = \xi$, where $\theta \in \Theta_0$ is an unknown parameter with the true value $\theta_0^* = (\theta_0^{(1)}, \theta_0^{(2)}) \in \Theta_0$. Let $\hat{\theta}_{n,\varepsilon}^{i,N} \in \Theta$ be the least squares estimation for the unknown parameter θ. For any $x, y \in \mathbb{R}$ and $\mu, \nu \in \mathscr{P}_2(\mathbb{R})$, set

$$\tilde{b}(x,y) := x - y^3 + y + \mathbb{E}y, \tag{6.16}$$

$$b(x,y,\mu,\nu,\theta) := \theta^{(1)} + \theta^{(2)}\tilde{b}(x,y) \tag{6.17}$$

and

$$\sigma(x,y,\mu,\nu,\theta) := 1 + |y|^3 + |y| + \mathbb{E}|y|. \tag{6.18}$$

Then, (6.15) can be reformulated as (6.1). Furthermore, according to Theorem 6.2 and Theorem 6.3, we get

$$\hat{\theta}_{n,\varepsilon}^{i,N} \to \theta_0$$

and

$$\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon}^{i,N} - \theta_0) \to I^{-1}(\theta_0) \int_0^T \Upsilon(X_s^{0,i}, X_{s-\tau}^{0,i}, \theta_0) \mathrm{d}W_s^i$$

in probability as $N, n \to \infty$ and $\varepsilon \to 0$. Here

$$I(\theta_0) = \int_0^T \frac{1}{(1 + 2|X_{s-\tau}^{0,i}| + |X_{s-\tau}^{0,i}|^3)^2} \begin{pmatrix} 1 & \tilde{b}(X_s^{0,i}, X_{s-\tau}^{0,i}) \\ \tilde{b}(X_s^{0,i}, X_{s-\tau}^{0,i}) & \tilde{b}(X_s^{0,i}, X_{s-\tau}^{0,i})^2 \end{pmatrix} \mathrm{d}s$$

and

$$\Upsilon(X_s^{0,i}, X_{s-\tau}^{0,i}, \theta_0) = \frac{1}{1 + 2|X_{s-\tau}^{0,i}| + |X_{s-\tau}^{0,i}|^3} \Big(1, \tilde{b}(X_s^{0,i}, X_{s-\tau}^{0,i})\Big)^*.$$

Next, concerning (6.15) we aim to examine that all the assumptions imposed on Theorem 6.2 and Theorem 6.3 apply very well. Indeed, by a direct calculation, for any $\mu, \nu, \mu_i, \nu_i \in \mathscr{P}_2(\mathbb{R})$ and $x, y, x_i, y_i \in \mathbb{R}$, $i = 1, 2$, it follows from (6.16), (6.17) and the Hölder inequality that there exists a constant $c > 0$

$$\begin{aligned}
&|b(x,y,\mu_1,\nu_1,\theta) - b(x,y,\mu_2,\nu_2,\theta)| \\
&\quad = |\theta^{(2)}| \cdot |\mathbb{E}(y_1 - y_2)| \le |\theta^{(2)}| \Big(\mathbb{E}|y_1 - y_2|^2\Big)^{\frac{1}{2}} \\
&\quad \le c\mathbb{W}_2(\nu_1,\nu_2), \\
&\langle x_1 - x_2, b(x_1,y_1,\mu,\nu,\theta) - b(x_2,y_2,\mu,\nu,\theta)\rangle \\
&\qquad \le |\theta^{(2)}| \Big(|x_1 - x_2|^2 + |y_1 - y_2|^2\Big) \\
&\qquad \le c\Big(|x_1 - x_2|^2 + |y_1 - y_2|^2\Big)
\end{aligned}$$

and

$$\begin{aligned}
&|b(x_1,y_1,\mu,\nu,\theta) - b(x_2,y_2,\mu,\nu,\theta)| \\
&\quad \le |\theta^{(2)}| \cdot \Big(|x_1 - x_2| + |y_1^3 - y_2^3| + |y_1 - y_2|\Big) \\
&\quad \le c\Big(|x_1 - x_2| + |y_1 - y_2|(1 + |y_1|^2 + |y_2|^2)\Big).
\end{aligned}$$

On the other hand, it holds that, by (6.18),

$$|\sigma(x_1, y_1, \mu, \nu) - \sigma(x_2, y_2, \mu, \nu)| \le |y_1 - y_2|(1 + |y_1|^2 + |y_2|^2)$$

and

$$|\sigma(x, y, \mu_1, \nu_1) - \sigma(x, y, \mu_2, \nu_2)| \le \mathbb{E}|y_1 - y_2| \le c\mathbb{W}_2(\nu_1, \nu_2).$$

Hence, Assumptions 6.1 and 6.2 hold for (6.15). Next, note that

$$\begin{aligned}
&|\sigma^{-2}(x_1, y_1, \mu_1, \nu_1) - \sigma^{-2}(x_2, y_2, \mu_2, \nu_2)| \\
&\quad = \Big|\frac{1}{(1 + |y_1|^3 + |y_1| + \mathbb{E}|y_1|)^2} - \frac{1}{(1 + |y_2|^3 + |y_2| + \mathbb{E}|y_2|)^2}\Big| \\
&\quad \le 4\Big||y_1|^3 + |y_1| + \mathbb{E}|y_1| - |y_2|^3 - |y_2| - \mathbb{E}|y_2|\Big| \\
&\quad \le c\Big(|y_1 - y_2|(1 + |y_1|^2 + |y_2|^2) + \mathbb{W}_2(\nu_1, \nu_2)\Big).
\end{aligned}$$

So, Assumption 6.3 is fulfilled. Furthermore, it follows from (6.17) that

$$(\nabla_\theta b)(\zeta, \mu, \theta) = \Big(1, \tilde{b}(x, y)\Big)^* \quad \text{and} \quad (\nabla_\theta(\nabla_\theta b))(\zeta, \mu, \theta) = \mathbf{0}_{2\times 2}, \tag{6.19}$$

where $\mathbf{0}_{2\times 2}$ stands for the 2×2-zero matrix. Thus, (6.16) yields that both Assumptions 6.4 and 6.5 hold. As a consequence, concerning (6.15), Assumptions 6.1 - 6.5 hold, respectively. In terms of (6.6), the contrast function enjoys the form

$$\begin{aligned}
\Psi_{n,\varepsilon}^{i,N}(\theta) =& \varepsilon^{-2}\delta^{-1}\sum_{k=1}^{n}\frac{1}{(1 + |X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|^3 + |X_{(k-1)\delta-\tau}^{\varepsilon,i,N}| + \mathbb{E}|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|)^2} \\
&\times \Big|X_{k\delta}^{\varepsilon,i,N} - X_{(k-1)\delta}^{\varepsilon,i,N} - \Big(\theta^{(1)} + \theta^{(2)}\tilde{b}(X_{(k-1)\delta}^{\varepsilon,i,N}, X_{(k-1)\delta-\tau}^{\varepsilon,i,N})\Big)\delta\Big|^2.
\end{aligned}$$

Note that

$$\begin{aligned}
\frac{\partial}{\partial\theta^{(1)}}\Psi_{n,\varepsilon}^{i,N}(\theta) =& -2\,\varepsilon^{-2}\sum_{k=1}^{n}\frac{1}{(1 + |X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|^3 + |X_{(k-1)\delta-\tau}^{\varepsilon,i,N}| + \mathbb{E}|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|)^2} \\
&\times \Big\{X_{k\delta}^{\varepsilon,i,N} - X_{(k-1)\delta}^{\varepsilon,i,N} - \Big(\theta^{(1)} + \theta^{(2)}\tilde{b}(X_{(k-1)\delta}^{\varepsilon,i,N}, X_{(k-1)\delta-\tau}^{\varepsilon,i,N})\Big)\delta\Big\}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial}{\partial\theta^{(2)}}\Psi_{n,\varepsilon}^{i,N}(\theta) =& -2\,\varepsilon^{-2}\sum_{k=1}^{n}\frac{1}{(1 + |X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|^3 + |X_{(k-1)\delta-\tau}^{\varepsilon,i,N}| + \mathbb{E}|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|)^2} \\
&\times \Big\{X_{k\delta}^{\varepsilon,i,N} - X_{(k-1)\delta}^{\varepsilon,i,N} - \Big(\theta^{(1)} + \theta^{(2)}\tilde{b}(X_{(k-1)\delta}^{\varepsilon,i,N}, X_{(k-1)\delta-\tau}^{\varepsilon,i,N})\Big)\delta\Big\} \\
&\times \tilde{b}(X_{(k-1)\delta}^{\varepsilon,i,N}, X_{(k-1)\delta-\tau}^{\varepsilon,i,N}).
\end{aligned}$$

Setting

$$\frac{\partial}{\partial\theta^{(1)}}\Psi_{n,\varepsilon}^{i,N}(\theta)=\frac{\partial}{\partial\theta^{(2)}}\Psi_{n,\varepsilon}^{i,N}(\theta)=0,$$

we obtain that the LSE $\hat{\theta}_{n,\varepsilon}^{i,N}=(\hat{\theta}_{n,\varepsilon}^{i,N,(1)},\hat{\theta}_{n,\varepsilon}^{i,N,(2)})^*$ of the unknown parameter $\theta=(\theta^{(1)},\theta^{(2)})^*\in\Theta_0$ possesses the formula

$$\hat{\theta}_{n,\varepsilon}^{i,N,(1)}=\frac{A_2A_5-A_3A_4}{\delta(A_1A_5-A_4^2)}\qquad\text{and}\qquad\hat{\theta}_{n,\varepsilon}^{i,N,(2)}=\frac{A_1A_3-A_2A_4}{\delta(A_1A_5-A_4^2)},$$

where

$$A_1:=\sum_{k=1}^{n}\frac{1}{(1+|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|^3+|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|+\mathbb{E}|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|)^2},$$

$$A_2:=\sum_{k=1}^{n}\frac{X_{k\delta}^{\varepsilon,i,N}-X_{(k-1)\delta}^{\varepsilon,i,N}}{(1+|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|^3+|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|+\mathbb{E}|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|)^2},$$

$$A_3:=\sum_{k=1}^{n}\frac{(X_{k\delta}^{\varepsilon,i,N}-X_{(k-1)\delta}^{\varepsilon,i,N})\tilde{b}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})}{(1+|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|^3+|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|+\mathbb{E}|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|)^2},$$

$$A_4:=\sum_{k=1}^{n}\frac{\tilde{b}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})}{(1+|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|^3+|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|+\mathbb{E}|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|)^2}$$

and

$$A_5:=\sum_{k=1}^{n}\frac{\Big(\tilde{b}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})\Big)^2}{(1+|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|^3+|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|+\mathbb{E}|X_{(k-1)\delta-\tau}^{\varepsilon,i,N}|)^2}.$$

In terms of Theorem 6.2, $\hat{\theta}_{n,\varepsilon}^{i,N}\to\theta$ in probability as $\varepsilon\to0$ and $n\to\infty$. Next, from (6.19), it follows that

$$I(\theta_0)=\int_0^T\frac{1}{(1+2|X_{s-\tau}^{0,i}|+|X_{s-\tau}^{0,i}|^3)^2}\begin{pmatrix}1 & \tilde{b}(X_s^{0,i},X_{s-\tau}^{0,i})\\ \tilde{b}(X_s^{0,i},X_{s-\tau}^{0,i}) & \tilde{b}(X_s^{0,i},X_{s-\tau}^{0,i})^2\end{pmatrix}\mathrm{d}s,$$

and, for $\zeta\in\mathscr{C}$,

$$\int_0^T\Upsilon(X_s^{0,i},X_{s-\tau}^{0,i},\theta_0)\mathrm{d}W(s)=\int_0^T\frac{1}{1+2|X_{s-\tau}^{0,i}|+|X_{s-\tau}^{0,i}|^3|}\begin{pmatrix}1\\ \tilde{b}(X_s^{0,i},X_{s-\tau}^{0,i})\end{pmatrix}\mathrm{d}W_s.$$

At last, according to Theorem 6.2, we conclude that

$$\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon}^{i,N}-\theta_0)\to I^{-1}(\theta_0)\int_0^T\Upsilon(X_t^{0,i},X_{t-\tau}^{0,i},\theta_0)\mathrm{d}W_t^i\qquad\mathbb{P}-\text{a.s.}$$

as $\varepsilon\to0$ and $n\to\infty$ provided that $I(\cdot)$ is positive definite.

6.5 Proof of main results

6.5.1 Proof of Theorem 6.1.

The more popular methods to argue regarding the existence of solutions of McKean-Vlasov SDEs need to seek the convergence of the corresponding distribution-iterated SDEs; see, for instance [HRW19, Wan18]. However, it is hard to verify the convergence of the distribution-iterated SDEs for (6.1), due to the coefficients satisfying polynomial growth with respect to the delay variables. So, we will adopt an interval iteration method to overcome this difficulty in terms of the structure of (6.1).

Proof of Theorem 6.1. Under Assumptions 6.1 and 6.2, firstly, we shall show the well-posedness of the delay McKean-Vlasov SDE (6.1). For each $t \in [0, \tau]$, (6.1) can be reformulated as

$$\begin{cases} \mathrm{d}X_t^\varepsilon = b(X_t^\varepsilon, \xi_{t-\tau}, \mu_t^\varepsilon, \mu_{t-\tau}^0, \theta)\mathrm{d}t + \varepsilon\sigma(X_t^\varepsilon, \xi_{t-\tau}, \mu_t^\varepsilon, \mu_{t-\tau}^0)\mathrm{d}W_t, \\ X_s^\varepsilon = \xi_s, \quad s \in [-\tau, 0], \end{cases} \tag{6.20}$$

where $\mu_\cdot^0 := \mathscr{L}_{\xi_\cdot}$. Then (6.20) is a non-delay SDE. In terms of Assumptions 6.1 and 6.2, it holds that the coefficients of (6.20) are Lipschitz continuous, then this SDE has a unique strong solution on the interval $[0, \tau]$.
On the interval $t \in [\tau, 2\tau]$, SDE (6.1) can be written as

$$\mathrm{d}X_t^\varepsilon = b(X_t^\varepsilon, X_t^{\varepsilon,(1)}, \mu_t^\varepsilon, \mu_t^1, \theta)\mathrm{d}t + \varepsilon\sigma(X_t^\varepsilon, X_t^{\varepsilon,(1)}, \mu_t^\varepsilon, \mu_t^1)\mathrm{d}W_t \tag{6.21}$$

with the initial value X_τ^ε, where $X_t^{\varepsilon,(1)} = X_{t-\tau}^\varepsilon$ and $\mu_\cdot^1 := \mathscr{L}_{X_\cdot^{\varepsilon,(1)}}$. Obviously, the delay McKean-Vlasov SDE (6.1) becomes a general McKean-Vlasov SDE with Lipschitz condition. Then this equation (6.21) has a unique strong solution on the interval $[\tau, 2\tau]$. Duplicating this procedure over the intervals $[n\tau, (n+1)\tau]$, where $2 < n \leq \lfloor T/\tau \rfloor$.

In addition, for any $x, y \in \mathbb{R}^d$ and $\mu, \nu \in \mathscr{P}_p(\mathbb{R}^d)$, by Assumptions 6.1 and 6.2, it is easy to see that there is a constant $C > 0$ such that

$$|b(x, y, \mu, \nu, \theta)| \leq C(1 + |x| + |y| + |y|^{r_1+1} + \mathbb{W}_2(\mu, \delta_0) + \mathbb{W}_2(\nu, \delta_0)) \tag{6.22}$$

and

$$\|\sigma(x, y, \mu, \nu)\| \leq C(1 + |x| + |y| + |y|^{r_2+1} + \mathbb{W}_2(\mu, \delta_0) + \mathbb{W}_2(\nu, \delta_0)), \tag{6.23}$$

where δ_0 is the Dirac measure at point $\mathbf{0} \in \mathbb{R}^n$.

Secondly, we shall show that the p-th moment of the solution is uniformly bounded in a finite time interval. In fact, set $(X_t^\varepsilon)_{t\geq-\tau}$ to be a solution to (6.1) with initial data $X_0^\varepsilon = \xi \in L^0_{p_1 q_1}(\mathscr{C})$. Let $\tau_m = \inf\{t > 0 : |X_t^\varepsilon| \geq m\}$, for $m \geq 1$. Then, by the Burkhold-Davis-Gundy inequality and Hölder inequality, together with (6.22) and (6.23), one gets

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le t\wedge\tau_m}|X_s^\varepsilon|^p\Big)\\
&\le C\mathbb{E}\|\xi\|_\infty^p + C\mathbb{E}\Big(\int_0^{t\wedge\tau_m}|b(X_s^\varepsilon, X_{s-\tau}^\varepsilon, \mu_s^\varepsilon, \mu_{s-\tau}^\varepsilon, \theta)|^p\mathrm{d}s\Big)\\
&\quad + C\mathbb{E}\Big(\int_0^{t\wedge\tau_m}\|\sigma(X_s^\varepsilon, X_{s-\tau}^\varepsilon, \mu_s^\varepsilon, \mu_{s-\tau}^\varepsilon)\|^p\mathrm{d}s\Big)\\
&\le C\Big\{1+\mathbb{E}\int_0^{t\wedge\tau_m}(1+|X_s^\varepsilon|^p)\mathrm{d}s + \mathbb{E}\int_0^{t\wedge\tau_m}\Big(|X_{s-\tau}^\varepsilon|^{(r_1+1)p}\\
&\quad + |X_{s-\tau}^\varepsilon|^{(r_2+1)p}\Big)\mathrm{d}s + \mathbb{E}\int_0^{t\wedge\tau_m}(\mathbb{W}_2(\mu_s^\varepsilon,\delta_0)^p + \mathbb{W}_2(\mu_{s-\tau}^\varepsilon,\delta_0)^p)\mathrm{d}s\Big\}.
\end{aligned}$$

Let $m\to\infty$, then we get

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le t}|X_s^\varepsilon|^p\Big)\\
&\quad\le C\Big\{1+\int_0^t\mathbb{E}|X_s^\varepsilon|^p\mathrm{d}s + \int_0^t(\mathbb{E}|X_{s-\tau}^\varepsilon|^{\gamma_1 p} + \mathbb{E}|X_{s-\tau}^\varepsilon|^{\gamma_2 p})\mathrm{d}s\\
&\qquad + \int_0^t(\mathbb{E}\mathbb{W}_2(\mu_s^\varepsilon,\delta_0)^p + \mathbb{E}\mathbb{W}_2(\mu_{s-\tau}^\varepsilon,\delta_0)^p)\mathrm{d}s\Big\}\\
&\quad\le C\Big\{1+\int_0^t\mathbb{E}|X_s^\varepsilon|^p\mathrm{d}s + \int_0^t(\mathbb{E}|X_{s-\tau}^\varepsilon|^{\gamma_1 p} + \mathbb{E}|X_{s-\tau}^\varepsilon|^{\gamma_2 p})\mathrm{d}s\Big\},
\end{aligned}$$

where $\gamma_1 := r_1+1$ and $\gamma_2 := r_2+1$. Then the Gronwall inequality yields

$$\mathbb{E}\Big(\sup_{0\le s\le t}|X_s^\varepsilon|^p\Big) \le C\Big\{1+\int_0^t(\mathbb{E}|X_{s-\tau}^\varepsilon|^{\gamma_1 p} + \mathbb{E}|X_{s-\tau}^\varepsilon|^{\gamma_2 p})\mathrm{d}s\Big\}. \tag{6.24}$$

Set $q_1 := \gamma_1\vee\gamma_2$ and

$$p_i = (\lfloor T/\tau\rfloor + 2 - i)pq_1^{\lfloor T/\tau\rfloor+1-i},\quad i=1,2,\cdots,\lfloor T/\tau\rfloor+1.$$

Then there exists a finite sequence $\{p_1,p_2,\cdots,p_{\lfloor T/\tau\rfloor+1}\}$ such that

$$p_i\ge 2,\quad p_{i+1}q_1<p_i\quad\text{and}\quad p_{\lfloor T/\tau\rfloor+1}=p,\quad i=1,2,\cdots,\lfloor T/\tau\rfloor.$$

Further, for $X_0^\varepsilon=\xi\in L^0_{p_1q_1}(\mathscr{C})$, one has

$$\mathbb{E}\Big(\sup_{0\le s\le\tau}|X_s^\varepsilon|^{p_1}\Big)\le C(1+E\|\xi\|_\infty^{p_1\gamma_1}+E\|\xi\|_\infty^{p_1\gamma_2})\le C, \tag{6.25}$$

which leads to

$$\mathbb{E}\Big(\sup_{-\tau\le s\le\tau}|X_s^\varepsilon|^{p_1}\Big)\le\mathbb{E}\Big(\sup_{-\tau\le s\le 0}|X_s^\varepsilon|^{p_1}\Big)+\mathbb{E}\Big(\sup_{0\le s\le\tau}|X_s^\varepsilon|^{p_1}\Big)\le C.$$

It follows from (6.24), (6.25) and the Hölder inequality that

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{-\tau\le s\le 2\tau}|X_s^\varepsilon|^{p_2}\Big)\\
&\le C\Big\{1+\int_0^{2\tau}(\mathbb{E}|X_{s-\tau}^\varepsilon|^{p_2\gamma_1}+\mathbb{E}|X_{s-\tau}^\varepsilon|^{p_2\gamma_2})\mathrm{d}s\Big\}\\
&\le C\Big\{1+\int_0^{\tau}\Big((\mathbb{E}|X_s^\varepsilon|^{p_1})^{p_2\gamma_1/p_1}+(\mathbb{E}|X_s^\varepsilon|^{p_1})^{p_2\gamma_2/p_1}\Big)\mathrm{d}s\Big\}\\
&\le C<\infty.
\end{aligned}$$

Carrying out the previous procedures gives (6.11). □

Remark 6.5. Obviously, $\mathbb{E}\Big(\sup_{-\tau\le t\le T}|X_t^0|^p\Big)\le C<\infty (p\ge 2)$ if the coefficients may be polynomial of any degree $r\ge 1$ with respect to the delay variables.

6.5.2 Proof of Theorem 6.2.

Next, to derive the consistency of LSE, we display some auxiliary results in the form of lemmas.

Lemma 6.6. *Assump Assumptions 6.1 and 6.2 hold. Then stochastic interacting particle system* (6.4) *has a strong solution with*

$$\sup_{i\in\mathbf{S}_N}\mathbb{E}\Big(\sup_{-\tau\le t\le T}|X_t^{\varepsilon,i,N}|^p\Big)\le C<\infty,\qquad\qquad p\ge 2.$$

Proof. For $x:=(x_1,x_2,\cdots,x_N)\in\mathbb{R}^d\otimes\mathbb{R}^N$, $y:=(y_1,y_2,\cdots,y_N)\in\mathbb{R}^d\otimes\mathbb{R}^N$, $x_i,y_i\in\mathbb{R}^d$, $i=1,2,\cdots,N$, define

$$\mu_x^N=\frac{1}{N}\sum_{i=1}^N\delta_{x_i},\qquad \mu_y^N=\frac{1}{N}\sum_{i=1}^N\delta_{y_i},$$

$$\begin{aligned}
\tilde{b}(x,y)&=(b(x_1,y_1,\mu_x^N,\mu_y^N,\theta),\cdots,b(x_N,y_N,\mu_x^N,\mu_y^N,\theta))^*,\\
\tilde{\sigma}(x,y)&=\mathrm{diag}\Big(\sigma(x_1,y_1,\mu_x^N,\mu_y^N),\cdots,\sigma(x_N,y_N,\mu_x^N,\mu_y^N)\Big)
\end{aligned}$$

and

$$\tilde{W}_t=\Big(W_t^1,\cdots,W_t^N\Big)^*.$$

Then, (6.4) can be redescribed as

$$\mathrm{d}X_t=\tilde{b}(X_t,X_{t-\tau})\mathrm{d}t+\varepsilon\tilde{\sigma}(X_t,X_{t-\tau})\mathrm{d}\tilde{W}_t,\quad t\ge 0. \tag{6.26}$$

Note that for any $x,y\in\mathbb{R}^d\otimes\mathbb{R}^N$

$$\mathbb{W}_2(\mu_x^N,\mu_y^N)\le\Big(\frac{1}{N}\sum_{j=1}^N|x_j-y_j|^2\Big)^{\frac{1}{2}}.$$

This, together with Assumptions 6.1 and 6.2, for any $x, x', y, y' \in \mathbb{R}^d \otimes \mathbb{R}^N$, leads to

$$|\tilde{b}(x,y) - \tilde{b}(x',y')| \le C|x-x'| + K_1(y_1,y_2)|y-y'|$$

and

$$|\tilde{\sigma}(x,y) - \tilde{\sigma}(x',y')| \le C|x-x'| + K_2(y_1,y_2)|y-y'|.$$

Then by [BY13, Lemma 2.2], it can be readily seen that (6.26) admits a unique global strong solution with

$$\mathbb{E}\Big(\sup_{-\tau\le t\le T}|X_t|^p\Big) \le C, \qquad p \ge 2.$$

Consequently, we conclude that (6.4) has a unique strong solution with

$$\sup_{i\in\mathbf{S}^N}\mathbb{E}\Big(\sup_{-\tau\le t\le T}|X_t^{\varepsilon,i,N}|^p\Big) \le C < \infty, \qquad p \ge 2.$$

The proof is completed. □

Remark 6.7. In [BH22, Lemma 3.1], Bao and Huang have investigated the question of the well-posedness of the stochastic N-interacting particle systems associated with McKean-Vlasov SDEs. We extend the idea used in [BH22] to the case of delay McKean-Vlasov SDEs.

Lemma 6.8. *Let Assumptions 6.1, 6.2 and 6.6 hold. Then, for initial value $X_0^\varepsilon = \xi \in L^0_{p_1q_1}(\mathscr{C})$, $p \ge 2$,*

$$\sup_{i\in\mathbf{S}_N}\mathbb{E}(\sup_{0\le t\le T}|X_t^{\varepsilon,i,N} - X_t^{\varepsilon,i}|^p) \le C(C_N + C_N^{\frac{1}{2}}), \tag{6.27}$$

where C_N is a decreasing function with respect to N and is defined as (6.35)*.*

Proof. For fixed $\lambda > 1$ and arbitrary $\epsilon \in (0,1)$, there exists a continuous nonnegative function $\psi_{\lambda\epsilon}$, $x \ge 0$, with support $[\epsilon/\lambda, \epsilon]$ such that

$$\int_{\epsilon/\lambda}^{\epsilon}\psi_{\lambda\epsilon}(x)\mathrm{d}x = 1 \qquad \text{and} \qquad 0 \le \psi_{\lambda\epsilon}(x) \le \frac{2}{x\ln\lambda}, \quad x > 0.$$

Let

$$\phi_{\lambda\epsilon}(x) = \int_0^x\int_0^y\psi_{\lambda\epsilon}(z)\mathrm{d}z\mathrm{d}y, \quad x > 0.$$

Then $\phi_{\lambda\epsilon}$ is C^2 and satisfies

$$x - \epsilon \le \phi_{\lambda\epsilon}(x) \le x, \quad x > 0 \tag{6.28}$$

and

$$0 \le \phi'_{\lambda\epsilon}(x) \le 1, \quad \phi''_{\lambda\epsilon}(x) \le \frac{2}{x\ln\lambda}\mathbf{1}_{[\epsilon/\lambda,\epsilon]}(x), \quad x > 0. \tag{6.29}$$

Define

$$V_{\lambda\epsilon}(x) = \phi_{\lambda\epsilon}(|x|), \qquad x \in \mathbb{R}^d. \tag{6.30}$$

Then, by the definition of $\phi_{\lambda\epsilon}$, it holds that $V_{\lambda\epsilon} \in C^2(\mathbb{R}^d;\mathbb{R}^+)$. For $x \in \mathbb{R}^d$, a direct calculation leads to

$$\frac{\partial V_{\lambda\epsilon}(x)}{\partial x_i} = \phi'_{\lambda\epsilon}(|x|)\frac{x_i}{|x|}$$

and

$$\frac{\partial^2 V_{\lambda\epsilon}(x)}{\partial x_i \partial x_j} = \phi'_{\lambda\epsilon}(|x|)(\delta_{ij}|x|^2 - x_i x_j)|x|^{-3} + \phi''_{\lambda\epsilon}(|x|)x_i x_j|x|^{-2}, \quad i,j = 1,2,\cdots,d,$$

where $\delta_{ij} = 1$ if $i = j$ or otherwise 0. Set

$$(V_{\lambda\epsilon})_x(x) := \Big(\frac{\partial V_{\lambda\epsilon}(x)}{\partial x_1},\cdots,\frac{\partial V_{\lambda\epsilon}(x)}{\partial x_d}\Big) \ \text{ and } \ (V_{\lambda\epsilon})_{xx}(x) := \Big(\frac{\partial^2 V_{\lambda\epsilon}(x)}{\partial x_i \partial x_j}\Big)_{d\times d}, \quad x \in \mathbb{R}^d.$$

According to (6.29) and (6.30), it holds

$$0 \le |(V_{\lambda\epsilon})_x(x)| \le 1 \text{ and } \|(V_{\lambda\epsilon})_{xx}(x)\| \le 2d\Big(1+\frac{1}{\ln\lambda}\Big)\frac{1}{|x|}\mathbf{1}_{[\epsilon/\lambda,\epsilon]}(|x|),\ x \in \mathbb{R}^d. \quad (6.31)$$

Set $Z^{i,N}_\cdot := X^{\varepsilon,i,N}_\cdot - X^{\varepsilon,i}_\cdot$ and $\overline{Z}^{i,N}_\cdot := (X^{\varepsilon,i,N}_\cdot, X^{\varepsilon,i}_\cdot) \in \mathbb{R}^{2d}$. For any $t \in [0,T]$, by the Itô formula, one gets

$$\begin{aligned}
&V_{\lambda\epsilon}(Z^{i,N}_t)\\
&\quad= \int_0^t \langle (V_{\lambda\epsilon})_x(Z^{i,N}_s), b(X^{\varepsilon,i,N}_s, X^{\varepsilon,i,N}_{s-\tau}, \mu^{\varepsilon,N}_s, \mu^{\varepsilon,N}_{s-\tau},\theta) - b(X^{\varepsilon,i}_s, X^{\varepsilon,i}_{s-\tau}, \mu^{\varepsilon}_s, \mu^{\varepsilon}_{s-\tau},\theta)\rangle \mathrm{d}s\\
&\qquad + \frac{\varepsilon^2}{2}\int_0^t \mathrm{trace}\{(\sigma(X^{\varepsilon,i,N}_s, X^{\varepsilon,i,N}_{s-\tau}, \mu^{\varepsilon,N}_s, \mu^{\varepsilon,N}_{s-\tau}) - \sigma(X^{\varepsilon,i}_s, X^{\varepsilon,i}_{s-\tau}, \mu^{\varepsilon}_s, \mu^{\varepsilon}_{s-\tau}))^*\\
&\qquad \times (V_{\lambda\epsilon})_{xx}(Z^{i,N}_s)(\sigma(X^{\varepsilon,i,N}_s, X^{\varepsilon,i,N}_{s-\tau}, \mu^{\varepsilon,N}_s, \mu^{\varepsilon,N}_{s-\tau}) - \sigma(X^{\varepsilon,i}_s, X^{\varepsilon,i}_{s-\tau}, \mu^{\varepsilon}_s, \mu^{\varepsilon}_{s-\tau}))\}\mathrm{d}s\\
&\qquad + \varepsilon\int_0^t \langle (V_{\lambda\epsilon})_x(Z^{i,N}_s), \sigma(X^{\varepsilon,i,N}_s, X^{\varepsilon,i,N}_{s-\tau}, \mu^{\varepsilon,N}_s, \mu^{\varepsilon,N}_{s-\tau}) - \sigma(X^{\varepsilon,i}_s, X^{\varepsilon,i}_{s-\tau}, \mu^{\varepsilon}_s, \mu^{\varepsilon}_{s-\tau})\mathrm{d}W^i_s\rangle\\
&\quad =: \sum_{i=1}^3 Q_i(t).
\end{aligned}$$

By means of Assumption 6.1, (6.31) and the Hölder inequality, we derive that, for any $t \in [0,T]$

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le t}|Q_1(s)|^p\Big)\\
&\quad \le \int_0^t \mathbb{E}|b(X^{\varepsilon,i,N}_s, X^{\varepsilon,i,N}_{s-\tau}, \mu^{\varepsilon,N}_s, \mu^{\varepsilon,N}_{s-\tau},\theta) - b(X^{\varepsilon,i}_s, X^{\varepsilon,i}_{s-\tau}, \mu^{\varepsilon}_s, \mu^{\varepsilon}_{s-\tau},\theta)|^p \mathrm{d}s\\
&\quad \le C\int_0^t \Big\{\mathbb{E}|Z^{i,N}_s|^p + \Big(\mathbb{E}K_1^{2p}(\overline{Z}^{i,N}_{s-\tau})\Big)^{\frac{1}{2}}\Big(\mathbb{E}|Z^{i,N}_{s-\tau}|^{2p}\Big)^{\frac{1}{2}}
\end{aligned} \quad (6.32)$$

$$+\,\mathbb{E}\mathbb{W}_2(\mu_s^{\varepsilon,N},\mu_s^\varepsilon)^p+\mathbb{E}\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,N},\mu_{s-\tau}^\varepsilon)^p\Big\}\mathrm{d}s.$$

By means of Assumption 6.2, (6.31) and the Hölder inequality, it holds that

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le t}|Q_2(s)|^p\Big)\\
&\le C\varepsilon^{2p}\mathbb{E}\int_0^t\frac{1}{|Z_s^{i,N}|^p}\Big\{|Z_s^{i,N}|^{2p}+K_2^{2p}(\overline{Z}_{s-\tau}^{i,N})|Z_{s-\tau}^{i,N}|^{2p}+\mathbb{W}_2(\mu_s^{\varepsilon,N},\mu_s^\varepsilon)^{2p}\\
&\quad+\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,N},\mu_{s-\tau}^\varepsilon)^{2p}\Big\}\mathbf{1}_{[\epsilon/\lambda,\epsilon]}(|Z_s^{i,N}|)\mathrm{d}s\\
&\le C\varepsilon^{2p}\int_0^t\Big\{\mathbb{E}|Z_s^{i,N}|^p+\frac{1}{\epsilon^p}\Big(\mathbb{E}K_2^{4p}(\overline{Z}_{s-\tau}^{i,N})\Big)^{\frac12}\Big(\mathbb{E}|Z_{s-\tau}^{i,N}|^{4p}\Big)^{\frac12} \qquad (6.33)\\
&\quad+\frac{1}{\epsilon^p}\mathbb{E}\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,N},\mu_{s-\tau}^\varepsilon)^{2p}+\mathbb{E}|Z_s^{i,N}|^{-p}\mathbb{W}_2(\mu_s^{\varepsilon,N},\mu_s^\varepsilon)^{2p}\mathbf{1}_{[\epsilon/\lambda,\epsilon]}(|Z_s^{i,N}|)\Big\}\mathrm{d}s\\
&\le C\varepsilon^{2p}\int_0^T\Big\{\mathbb{E}|Z_s^{i,N}|^p+\frac{1}{\epsilon^p}\Big(\mathbb{E}K_2^{4p}(\overline{Z}_{s-\tau}^{i,N})\Big)^{\frac12}\Big(\mathbb{E}|Z_{s-\tau}^{i,N}|^{4p}\Big)^{\frac12}\\
&\quad+\frac{1}{\epsilon^p}\Big(\mathbb{E}\mathbb{W}_2(\mu_s^{\varepsilon,N},\mu_s^\varepsilon)^{2p}+\mathbb{E}\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,N},\mu_{s-\tau}^\varepsilon)^{2p}\Big)\Big\}\mathrm{d}s.
\end{aligned}$$

By virtue of Assumption 6.2, the Burkhold-Davis-Gundy inequality, the Hölder inequality and the Young inequality, we derive that

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le t}|Q_3(s)|^p\Big)\\
&\le C\varepsilon^p\int_0^t\Big\{\mathbb{E}|Z_s^{i,N}|^p+\Big(\mathbb{E}K_2^{2p}(\overline{Z}_{s-\tau}^{i,N})\Big)^{\frac12}\Big(\mathbb{E}|Z_{s-\tau}^{i,N}|^{2p}\Big)^{\frac12} \qquad (6.34)\\
&\quad+\mathbb{E}\mathbb{W}_2(\mu_s^{\varepsilon,N},\mu_s^\varepsilon)^p+\mathbb{E}\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,N},\mu_{s-\tau}^\varepsilon)^p\Big\}\mathrm{d}s.
\end{aligned}$$

In addition, it follows from (6.10), (6.11) and Lemma 6.6 that

$$\mathbb{E}K_1^{2p}(\overline{Z}_{s-\tau}^{i,N})+\mathbb{E}K_2^{4p}(\overline{Z}_{s-\tau}^{i,N})\le C.$$

This, together with (6.32), (6.33) and (6.34), it holds from (6.28) that, for any $t\in[0,T]$ and $p\ge 2$

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le t}|Z_s^{i,N}|^p\Big)\\
&\le 2^{p-1}\Big\{\epsilon^p+\mathbb{E}\Big(\sup_{0\le s\le t}V_{\lambda\epsilon}^p(Z_s^{i,N})\Big)\Big\}\\
&\le C\Big\{\epsilon^p+\int_0^t\Big\{\mathbb{E}|Z_s^{i,N}|^p+\Big(\mathbb{E}|Z_{s-\tau}^{i,N}|^{2p}\Big)^{\frac12}+\frac{1}{\epsilon^p}\Big(\mathbb{E}|Z_{s-\tau}^{i,N}|^{4p}\Big)^{\frac12}\\
&\quad+\mathbb{E}\mathbb{W}_2(\mu_s^{\varepsilon,N},\mu_s^\varepsilon)^p+\mathbb{E}\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,N},\mu_{s-\tau}^\varepsilon)^p
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{\epsilon^p}\Big(\mathbb{E}\mathbb{W}_2(\mu_s^{\varepsilon,N},\mu_s^\varepsilon)^{2p}+\mathbb{E}\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,N},\mu_{s-\tau}^\varepsilon)^{2p}\Big)\Big\}\mathrm{d}s\\
\le{}& C\Big\{\epsilon^p+\int_0^t\Big\{\mathbb{E}|Z_s^{i,N}|^p+\Big(\mathbb{E}|Z_{s-\tau}^{i,N}|^{2p}\Big)^{\frac12}+\frac{1}{\epsilon^p}\Big(\mathbb{E}|Z_{s-\tau}^{i,N}|^{4p}\Big)^{\frac12}\\
&+\mathbb{E}\mathbb{W}_p(\tilde\mu_s^{\varepsilon,N},\mu_s^\varepsilon)^p+\mathbb{E}\mathbb{W}_p(\tilde\mu_{s-\tau}^{\varepsilon,N},\mu_{s-\tau}^\varepsilon)^p\\
&+\frac{1}{\epsilon^p}\Big(\mathbb{E}\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,N},\tilde\mu_{s-\tau}^{\varepsilon,N})^{2p}+\mathbb{E}\mathbb{W}_{2p}(\tilde\mu_{s-\tau}^{\varepsilon,N},\mu_{s-\tau}^\varepsilon)^{2p}\\
&+\mathbb{E}\mathbb{W}_{2p}(\tilde\mu_s^{\varepsilon,N},\mu_s^\varepsilon)^{2p}\Big)\Big\}\mathrm{d}s,
\end{aligned}
$$

we used the Hölder inequality and the fact that

$$
\mathbb{E}\mathbb{W}_2(\mu_\cdot^{\varepsilon,N},\mu_\cdot^\varepsilon)^p\le\mathbb{E}\mathbb{W}_2(\mu_\cdot^{\varepsilon,N},\tilde\mu_\cdot^{\varepsilon,N})^p+\mathbb{E}\mathbb{W}_2(\tilde\mu_\cdot^{\varepsilon,N},\mu_\cdot^\varepsilon)^p\le\mathbb{E}|Z_\cdot^{i,N}|^p+\mathbb{E}\mathbb{W}_2(\tilde\mu_\cdot^{\varepsilon,N},\mu_\cdot^\varepsilon)^p
$$

since $(Z_\cdot^{i,N})_{1\le j\le N}$ are identically distributed. Moreover, according to [FG15, Theorem 1] and Assumption 6.6, it holds that

$$
\begin{aligned}
\mathbb{E}\mathbb{W}_p(\tilde\mu_\cdot^{\varepsilon,N},\mu_\cdot^\varepsilon)^p\le\tilde C&\begin{cases}N^{-1/2}+N^{\frac{p}{q}-1}, & \text{if } p>\frac d2,\ q\neq 2p,\\ N^{-1/2}\log(1+N)+N^{\frac{p}{q}-1}, & \text{if } \ p=\frac d2,\ \ q\neq 2p,\\ N^{-p/d}+N^{\frac{p}{q}-1}, & \text{if } p\in(0,\frac d2),\ q\neq\frac{d}{d-p}.\end{cases}\\
&=:C_N.
\end{aligned}\tag{6.35}
$$

Thus, it follows from the Gronwall inequality that

$$
\mathbb{E}\Big(\sup_{0\le s\le t}|Z_s^{i,N}|^p\Big)\le C\Big\{\epsilon^p+\int_0^t\Big\{\Big(\mathbb{E}|Z_{s-\tau}^{i,N}|^{2p}\Big)^{\frac12}+\frac{1}{\epsilon^p}\Big(\mathbb{E}|Z_{s-\tau}^{i,N}|^{4p}\Big)^{\frac12}+C_N+\frac{1}{\epsilon^p}C_N\Big\}\mathrm{d}s\Big\}.
$$

Set, for any $p\ge2$,

$$
p_i=(\lfloor T/\tau\rfloor+2-i)p4^{\lfloor T/\tau\rfloor+1-i},\quad i=1,2,\cdots,\lfloor T/\tau\rfloor+1.
$$

Then it is easy to see that

$$
p_i\ge2,\quad 4p_{i+1}<p_i\quad\text{and}\quad p_{\lfloor T/\tau\rfloor+1}=p,\quad i=1,2,\cdots,\lfloor T/\tau\rfloor. \tag{6.36}
$$

For $s\in[0,\tau]$, $Z_{s-\tau}^{i,N}=0$, which, and taking $\epsilon=C_N^{\frac{1}{2p_1}}$ implies that

$$
\mathbb{E}\Big(\sup_{0\le s\le\tau}|Z_s^{i,N}|^{p_1}\Big)\le C(C_N+C_N^{\frac12}).
$$

This fact, together with (6.36) and the Hölder inequality, implies

$$
\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le 2\tau}|Z_s^{i,N}|^{p_2}\Big)\\
&\le C\Big\{\epsilon^{p_2}+\int_0^{2\tau}\Big\{\Big(\mathbb{E}|Z_{s-\tau}^{i,N}|^{p_1}\Big)^{\frac{p_2}{p_1}}+\frac{1}{\epsilon^{p_2}}\Big(\mathbb{E}|Z_{s-\tau}^{i,N}|^{p_1}\Big)^{\frac{2p_2}{p_1}}+C_N+\frac{1}{\epsilon^{p_2}}C_N\Big\}\mathrm{d}s\Big\}
\end{aligned}
$$

$$\leq C(C_N + C_N^{\frac{1}{2}}),$$

by setting $\epsilon = C_N^{\frac{1}{2p_2}}$. Repeating the previous procedures gives the desired assertion (6.27). □

Remark 6.9. In terms of Lemma 6.8, it is desirable to measure the convergence between stochastic interacting particle systems and the corresponding non-interacting particle systems in the sense of the p-moment. This result plays an important role in the process of establishing the consistency of the LSE.

Lemma 6.10. *Let Assumptions 6.1, 6.2 and 6.6 hold. Then, for initial value $X_0^\varepsilon = \xi \in L^0_{p_1 q_1}(\mathscr{C})$, $p \geq 2$, there is $C > 0$ such that*

$$\mathbb{E}\Big(\sup_{0\leq t\leq T} |X_{t_\delta}^{\varepsilon,i} - X_t^{0,i}|^p\Big) \leq C\delta(\delta^{p-1} + \varepsilon^p) + C\varepsilon^p, \quad i \in \mathbf{S}_N, \tag{6.37}$$

where $t_\delta := \lfloor t/\delta \rfloor \delta$.

Proof. For any $t \in [0, T]$,

$$|X_{t_\delta}^{\varepsilon,i} - X_t^{0,i}|^p \leq 2^{p-1}|X_{t_\delta}^{\varepsilon,i} - X_t^{\varepsilon,i}|^p + 2^{p-1}|X_t^{\varepsilon,i} - X_t^{0,i}|^p. \tag{6.38}$$

Now, for any $t \in [0, T]$, there exists an integer $k_0 \in [0, n-1]$ such that $t \in [k_0\delta, (k_0 + 1)\delta]$. Obviously, $k_0 = [t/\delta]$. Next, the Hölder inequality and Burkhold-Davis-Gundy inequality, together with Assumptions 6.1 and 6.2, yield that

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\leq t\leq T} |X_{t_\delta}^{\varepsilon,i} - X_t^{\varepsilon,i}|^p\Big) \\
&= \mathbb{E}\Big(\sup_{0\leq t\leq T} |X_{k_0\delta}^{\varepsilon,i} - X_t^{\varepsilon,i}|^p\Big) \\
&\leq 2^{p-1}\delta^{p-1}\mathbb{E}\int_{k_0\delta}^T |b(X_s^{\varepsilon,i}, X_{s-\tau}^{\varepsilon,i}, \mu_s^{\varepsilon,i}, \mu_{s-\tau}^{\varepsilon,i}, \theta)|^p \mathrm{d}s + 2^{p-1}\varepsilon^p C_{p,T} \\
&\quad \times \mathbb{E}\int_{k_0\delta}^T \|\sigma(X_s^{\varepsilon,i}, X_{s-\tau}^{\varepsilon,i}, \mu_s^{\varepsilon,i}, \mu_{s-\tau}^{\varepsilon,i})\|^p \mathrm{d}s \qquad (6.39) \\
&\leq C(2\delta)^{p-1}\int_{k_0\delta}^T \Big\{1 + \mathbb{E}|X_s^{\varepsilon,i}|^p + \mathbb{E}|X_{s-\tau}^{\varepsilon,i}|^p + \mathbb{E}|X_{s-\tau}^{\varepsilon,i}|^{p(r_1+1)} + \mathbb{E}\mathbb{W}_2(\mu_s^{\varepsilon,i}, \delta_0)^p \\
&\quad + \mathbb{E}\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,i}, \delta_0)^p\Big\}\mathrm{d}s + C\varepsilon^p\int_{k_0\delta}^T \Big\{1 + \mathbb{E}|X_s^{\varepsilon,i}|^p + \mathbb{E}|X_{s-\tau}^{\varepsilon,i}|^p \\
&\quad + \mathbb{E}|X_{s-\tau}^{\varepsilon,i}|^{p(r_2+1)} + \mathbb{E}\mathbb{W}_2(\mu_s^{\varepsilon,i}, \delta_0)^p + \mathbb{E}\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,i}, \delta_0)^p\Big\}\mathrm{d}s \\
&\leq C(\delta^{p-1} + \varepsilon^p)\int_{k_0\delta}^T \Big\{1 + \mathbb{E}|X_s^{\varepsilon,i}|^p + \mathbb{E}|X_{s-\tau}^{\varepsilon,i}|^{pq_1}\Big\}\mathrm{d}s \\
&\leq C\delta(\delta^{p-1} + \varepsilon^p).
\end{aligned}$$

Moreover, set $Z_t^i := X_t^{\varepsilon,i} - X_t^{0,i}$ and $\overline{Z}_\cdot^i := (X_\cdot^{\varepsilon,i}, X_\cdot^{0,i}) \in \mathbb{R}^{2d}$. Applying the Itô formula implies that

$$
\begin{aligned}
&V_{\lambda\epsilon}(Z_t^i)\\
&= \int_0^t \langle (V_{\lambda\epsilon})_x(Z_s^i), b(X_s^{\varepsilon,i}, X_{s-\tau}^{\varepsilon,i}, \mu_s^{\varepsilon,i}, \mu_{s-\tau}^{\varepsilon,i}, \theta) - b(X_s^{0,i}, X_{s-\tau}^{0,i}, \mu_s^0, \mu_{s-\tau}^0, \theta)\rangle \mathrm{d}s\\
&\quad + \frac{\varepsilon^2}{2}\int_0^t \mathrm{trace}\{(\sigma(X_s^{\varepsilon,i}, X_{s-\tau}^{\varepsilon,i}, \mu_s^{\varepsilon,i}, \mu_{s-\tau}^{\varepsilon,i}))^*(V_{\lambda\epsilon})_{xx}(Z_s^i)\sigma(X_s^{\varepsilon,i}, X_{s-\tau}^{\varepsilon,i}, \mu_s^{\varepsilon,i}, \mu_{s-\tau}^{\varepsilon,i})\}\mathrm{d}s\\
&\quad + \varepsilon\int_0^t \langle (V_{\lambda\epsilon})_x(Z_s^i), \sigma(X_s^{\varepsilon,i}, X_{s-\tau}^{\varepsilon,i}, \mu_s^{\varepsilon}, \mu_{s-\tau}^{\varepsilon})\mathrm{d}W_s^i\rangle\\
&=: \sum_{i=1}^3 \overline{Q}_i(t).
\end{aligned}
$$

By means of Assumption 6.1, (6.31) and the Hölder inequality, we derive that, for any $t \in [0,T]$

$$
\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le t} |\overline{Q}_1(s)|^p\Big)\\
&\le \int_0^t \mathbb{E}|b(X_s^{\varepsilon,i}, X_{s-\tau}^{\varepsilon,i}, \mu_s^{\varepsilon,i}, \mu_{s-\tau}^{\varepsilon,i}, \theta) - b(X_s^{0,i}, X_{s-\tau}^{0,i}, \mu_s^0, \mu_{s-\tau}^0, \theta)|^p \mathrm{d}s\\
&\le C\int_0^t \Big\{\mathbb{E}|Z_s^i|^p + \Big(\mathbb{E}K_1^{2p}(\overline{Z}_{s-\tau}^i)\Big)^{\frac12}\Big(\mathbb{E}|Z_{s-\tau}^i|^{2p}\Big)^{\frac12} + \mathbb{E}\mathbb{W}_2(\mu_s^{\varepsilon,i}, \mu_s^0)^p \qquad (6.40)\\
&\quad + \mathbb{E}\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,i}, \mu_{s-\tau}^0)^p\Big\}\mathrm{d}s\\
&\le C\int_0^t \Big\{\mathbb{E}|Z_s^i|^p + \mathbb{E}|Z_{s-\tau}^i|^p + \Big(\mathbb{E}|Z_{s-\tau}^i|^{2p}\Big)^{\frac12}\Big\}\mathrm{d}s.
\end{aligned}
$$

By means of Assumption 6.2, the elementary inequality, the Hölder inequality and (6.31), it holds that

$$
\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le t} |\overline{Q}_2(s)|^p\Big)\\
&\le \frac{\varepsilon^{2p}}{2}\mathbb{E}\int_0^t \{\|(V_{\lambda\epsilon})_{xx}(Z_s^i)\|\|\sigma(X_s^{\varepsilon,i}, X_{s-\tau}^{\varepsilon,i}, \mu_s^{\varepsilon,i}, \mu_{s-\tau}^{\varepsilon,i})\|^2\}^p \mathrm{d}s\\
&\le C\varepsilon^{2p}\mathbb{E}\int_0^t \frac{1}{|Z_s^i|^p}\Big\{1 + |Z_s^i|^{2p} + |Z_{s-\tau}^i|^{2p} + |Z_{s-\tau}^i|^{2p(r_2+1)} + \mathbb{W}_2(\mu_s^{\varepsilon,i}, \mu_s^0)^{2p}\\
&\quad + \mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,i}, \mu_{s-\tau}^0)^{2p}\Big\}\mathbf{1}_{[\epsilon/\lambda,\epsilon]}(|Z_s^i|)\mathrm{d}s \qquad (6.41)\\
&\le C\varepsilon^{2p}\int_0^t \Big\{\epsilon^p + \mathbb{E}|Z_s^i|^p + \frac{1}{\epsilon^p}\Big(1 + \mathbb{E}|Z_{s-\tau}^i|^{2p} + \mathbb{E}|Z_{s-\tau}^i|^{2p(r_2+1)}\Big)\Big\}\mathrm{d}s.
\end{aligned}
$$

By virtue of Assumption 6.2, the Burkhold-Davis-Gundy inequality, the Hölder inequality and the elementary inequality, we derive that

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le t}|\overline{Q}_3(s)|^p\Big)\\
&\le C\varepsilon^p\mathbb{E}\int_0^t\|\sigma(X_s^{\varepsilon,i},X_{s-\tau}^{\varepsilon,i},\mu_s^{\varepsilon,i},\mu_{s-\tau}^{\varepsilon,i})\|^p\mathrm{d}s\\
&\le C\varepsilon^p\mathbb{E}\int_0^t\Big\{1+|X_s^{\varepsilon,i}|^p+|X_{s-\tau}^{\varepsilon,i}|^p+|X_{s-\tau}^{\varepsilon,i}|^{p(r_2+1)}\\
&\quad+\mathbb{W}_2(\mu_s^{\varepsilon,i},\mu_s^0)^p+\mathbb{W}_2(\mu_{s-\tau}^{\varepsilon,i},\mu_{s-\tau}^0)^p\Big\}\mathrm{d}s\\
&\le C\varepsilon^p\int_0^t\Big\{1+\mathbb{E}|Z_s^i|^p+\mathbb{E}|Z_{s-\tau}^i|^p+\mathbb{E}|Z_{s-\tau}^i|^{p(r_2+1)}\Big\}\mathrm{d}s.
\end{aligned}\tag{6.42}$$

Furthermore, in view of (6.40), (6.41) and (6.42), we derive that, for any $t\in[0,T]$ and $p\ge 2$,

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le t}|Z_s^i|^p\Big)\\
&\le 2^{p-1}\Big\{\epsilon^p+\mathbb{E}\Big(\sup_{0\le s\le t}V_{\lambda\epsilon}^p(Z_s^i)\Big)\Big\}\\
&\le C\Big\{\epsilon^p+\int_0^t\Big\{\mathbb{E}|Z_s^i|^p+\mathbb{E}|Z_{s-\tau}^i|^p+\Big(\mathbb{E}|Z_{s-\tau}^i|^{2p}\Big)^{\frac12}+\epsilon^p\varepsilon^{2p}\\
&\quad+\frac{\varepsilon^{2p}}{\epsilon^p}\Big(1+\mathbb{E}|Z_{s-\tau}^i|^{2p}+\mathbb{E}|Z_{s-\tau}^i|^{2p(r_2+1)}\Big)+\varepsilon^p+\varepsilon^p\mathbb{E}|Z_{s-\tau}^i|^{p(r_2+1)}\Big\}\mathrm{d}s.
\end{aligned}$$

Then, the Gronwall inequality implies that

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le t}|Z_s^i|^p\Big)\\
&\le C\Big\{\epsilon^p+\int_0^t\Big\{\epsilon^p\varepsilon^{2p}+\varepsilon^p+\mathbb{E}|Z_{s-\tau}^i|^p+\Big(\mathbb{E}|Z_{s-\tau}^i|^{2p}\Big)^{\frac12}\\
&\quad+\frac{\varepsilon^{2p}}{\epsilon^p}\Big(1+\mathbb{E}|Z_{s-\tau}^i|^{2p}+\mathbb{E}|Z_{s-\tau}^i|^{2p(r_2+1)}\Big)+\varepsilon^p\mathbb{E}|Z_{s-\tau}^i|^{p(r_2+1)}\Big\}\mathrm{d}s.
\end{aligned}\tag{6.43}$$

Set, for any $p\ge 2$,

$$p_i=(\lfloor T/\tau\rfloor+2-i)p(2r_2+2)^{\lfloor T/\tau\rfloor+1-i},\quad i=1,2,\cdots,\lfloor T/\tau\rfloor+1.$$

Then it is easy to see that

$$p_i\ge 2,\quad 2(r_2+1)p_{i+1}<p_i\ \text{ and }\ p_{\lfloor T/\tau\rfloor+1}=p,\quad i=1,2,\cdots,\lfloor T/\tau\rfloor.\tag{6.44}$$

For $s\in[0,\tau]$, $Z_{s-\tau}^i=0$, which, and taking $\epsilon=\varepsilon$ implies that

$$\mathbb{E}\Big(\sup_{0\le s\le\tau}|Z_s^i|^{p_1}\Big)\le C\varepsilon^{p_1}.$$

This fact together with (6.43), (6.44) and the Hölder inequality implies, by setting $\epsilon = \varepsilon$,

$$
\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le 2\tau}|Z_s^i|^{p_2}\Big)\\
&\le C\Big\{\epsilon^{p_2}+\int_0^{2\tau}\Big\{\epsilon^{p_2}\varepsilon^{2p_2}+\varepsilon^{p_2}+\mathbb{E}|Z_{s-\tau}^i|^{p_2}+\Big(\mathbb{E}|Z_{s-\tau}^i|^{2p_2}\Big)^{\frac{1}{2}}\\
&\quad+\frac{\varepsilon^{2p_2}}{\epsilon^{p_2}}\Big(1+\mathbb{E}|Z_{s-\tau}^i|^{2p_2}+\mathbb{E}|Z_{s-\tau}^i|^{2p_2(r_2+1)}\Big)+\varepsilon^{p_2}\mathbb{E}|Z_{s-\tau}^i|^{p_2(r_2+1)}\Big\}\mathrm{d}s\\
&\le C\Big\{\epsilon^{p_2}+\int_0^{2\tau}\Big\{\epsilon^{p_2}\varepsilon^{2p_2}+\varepsilon^{p_2}+\Big(\mathbb{E}|Z_{s-\tau}^i|^{p_1}\Big)^{\frac{p_2}{p_1}}+\Big(\mathbb{E}|Z_{s-\tau}^i|^{p_1}\Big)^{\frac{p_2}{p_1}}\\
&\quad+\frac{\varepsilon^{2p_2}}{\epsilon^{p_2}}\Big(1+\Big(\mathbb{E}|Z_{s-\tau}^i|^{p_1}\Big)^{\frac{2p_2}{p_1}}+\Big(\mathbb{E}|Z_{s-\tau}^i|^{p_1}\Big)^{\frac{2p_2(r_2+1)}{p_1}}\Big)+\varepsilon^{p_2}\Big(\mathbb{E}|Z_{s-\tau}^i|^{p_1}\Big)^{\frac{p_2(r_2+1)}{p_1}}\Big\}\mathrm{d}s\\
&\le C\varepsilon^{p_2}.
\end{aligned}
$$

Following the previous procedures implies that

$$
\mathbb{E}\Big(\sup_{0\le t\le T}|Z_t^i|^{p}\Big)\le C\varepsilon^{p}. \tag{6.45}
$$

Plugging (6.39) and (6.45) into (6.38) yields (6.37). □

Lemma 6.11. *Let Assumptions 6.1, 6.2 and 6.3 hold. Then, for any initial value* $X_0^\varepsilon=\xi\in L^0_{p_1q_1}(\mathscr{C})$,

$$
\Phi_{n,\varepsilon}^{i,N,(1)}:=\sum_{k=1}^n B^*(X^{\varepsilon,i,N}_{(k-1)\delta},X^{\varepsilon,i,N}_{(k-1)\delta-\tau},\theta_0,\theta)\Lambda^{-1}(X^{\varepsilon,i,N}_{(k-1)\delta},X^{\varepsilon,i,N}_{(k-1)\delta-\tau})P_k^{\varepsilon,i,N}(\theta_0)\to 0,
$$

in L^1 *as* $\varepsilon\to0$ *and* $n,N\to\infty$.

Proof. According to (6.5), we get

$$
\mathbb{W}_2(\mu_s^{\varepsilon,N},\delta_0)^2\le\frac{1}{N}\sum_{i=1}^N|X_s^{\varepsilon,i,N}|^2,\qquad s\ge-\tau. \tag{6.46}
$$

In view of (6.4) and (6.7), one has

$$
\begin{aligned}
\Phi_{n,\varepsilon}^{i,N,(1)}=&\varepsilon\sum_{k=1}^n B^*(X^{\varepsilon,i,N}_{(k-1)\delta},X^{\varepsilon,i,N}_{(k-1)\delta-\tau},\theta_0,\theta)\Lambda^{-1}(X^{\varepsilon,i,N}_{(k-1)\delta},X^{\varepsilon,i,N}_{(k-1)\delta-\tau})\\
&\times\sigma(X^{\varepsilon,i,N}_{(k-1)\delta},X^{\varepsilon,i,N}_{(k-1)\delta-\tau},\mu^{\varepsilon,N}_{(k-1)\delta},\mu^{\varepsilon,N}_{(k-1)\delta-\tau})(W^i_{k\delta}-W^i_{(k-1)\delta})\\
=&\varepsilon\int_0^T B^*(X^{\varepsilon,i,N}_{s_\delta},X^{\varepsilon,i,N}_{s_\delta-\tau},\theta_0,\theta)\Lambda^{-1}(X^{\varepsilon,i,N}_{s_\delta},X^{\varepsilon,i,N}_{s_\delta-\tau})
\end{aligned}
$$

$$\times\,\sigma(X_{s_\delta}^{\varepsilon,i,N},X_{s_\delta-\tau}^{\varepsilon,i,N},\mu_{s_\delta}^{\varepsilon,N},\mu_{s_\delta-\tau}^{\varepsilon,N})\mathrm{d}W_s^i,$$

where $s_\delta:=\lfloor s/\delta\rfloor\delta$. This, together with the Hölder inequality and [CD18, Theorem 7.1], further implies that

$$\begin{aligned}\mathbb{E}|\Phi_{n,\varepsilon}^{i,N,(1)}|=&\varepsilon\mathbb{E}\Big|\int_0^T B^*(X_{s_\delta}^{\varepsilon,i,N},X_{s_\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta)\Lambda^{-1}(X_{s_\delta}^{\varepsilon,i,N},X_{s_\delta-\tau}^{\varepsilon,i,N})\\&\times\sigma(X_{s_\delta}^{\varepsilon,i,N},X_{s_\delta-\tau}^{\varepsilon,i,N},\mu_{s_\delta}^{\varepsilon,N},\mu_{s_\delta-\tau}^{\varepsilon,N})\mathrm{d}W_s^i\Big|\\\le&C\varepsilon\Big(\mathbb{E}\int_0^T|B^*(X_{s_\delta}^{\varepsilon,i,N},X_{s_\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta)|^2\|\Lambda^{-1}(X_{s_\delta}^{\varepsilon,i,N},X_{s_\delta-\tau}^{\varepsilon,i,N})\|^2\\&\times\|\sigma(X_{s_\delta}^{\varepsilon,i,N},X_{s_\delta-\tau}^{\varepsilon,i,N},\mu_{s_\delta}^{\varepsilon,N},\mu_{s_\delta-\tau}^{\varepsilon,N})\|^2\mathrm{d}s\Big)^{\frac{1}{2}}.\end{aligned}\tag{6.47}$$

One the other hand, for any $x,y\in\mathbb{R}^d$ and $\mu,\nu\in\mathscr{P}(\mathbb{R}^d)$, $i=1,2$, by Assumption 6.3, it is easy to see that there is a constant $\overline{L}>0$ such that

$$\|(\sigma\sigma^*)^{-1}(x,y,\mu,\nu)\|\le\overline{L}\Big\{1+|x|+|y|+|y|^{r_3+1}+\mathbb{W}_2(\mu,\delta_0)+\mathbb{W}_2(\nu,\delta_0)\Big\}.\tag{6.48}$$

Now, it follows from (6.22) that

$$\begin{aligned}|B^*(X_{s_\delta}^{\varepsilon,i,N},X_{s_\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta)|^2\le&C\Big\{1+|X_{s_\delta}^{\varepsilon,i,N}|^2+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^2+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^{2(r_1+1)}\\&+\mathbb{W}_2(\mu_{s_\delta}^{\varepsilon,N},\delta_0)^2+\mathbb{W}_2(\mu_{s_\delta-\tau}^{\varepsilon,N},\delta_0)^2\Big\}.\end{aligned}\tag{6.49}$$

Due to (6.23), we obtain

$$\begin{aligned}\|\sigma(X_{s_\delta}^{\varepsilon,i,N},X_{s_\delta-\tau}^{\varepsilon,i,N},\mu_{s_\delta}^{\varepsilon,N},\mu_{s_\delta-\tau}^{\varepsilon,N})\|^2\le C\Big\{&1+|X_{s_\delta}^{\varepsilon,i,N}|^2+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^2+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^{2(r_2+1)}\\&+\mathbb{W}_2(\mu_{s_\delta}^{\varepsilon,N},\delta_0)^2+\mathbb{W}_2(\mu_{s_\delta-\tau}^{\varepsilon,N},\delta_0)^2\Big\}.\end{aligned}\tag{6.50}$$

From (6.48), one has

$$\begin{aligned}\|\Lambda^{-1}(X_{s_\delta}^{\varepsilon,i,N},X_{s_\delta-\tau}^{\varepsilon,i,N})\|^2\le&C\Big\{1+|X_{s_\delta}^{\varepsilon,i,N}|^2+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^2+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^{2(r_3+1)}\\&+\mathbb{W}_2(\mu_{s_\delta}^{\varepsilon,N},\delta_0)^2+\mathbb{W}_2(\mu_{s_\delta-\tau}^{\varepsilon,N},\delta_0)^2\Big\}.\end{aligned}\tag{6.51}$$

Substituting these inequalities into (6.47), by the Hölder inequality and using (6.46) lead to

$$\begin{aligned}&\mathbb{E}|\Phi_{n,\varepsilon}^{i,N,(1)}|\\&\le C\varepsilon\Big\{\mathbb{E}\int_0^T\Big\{1+|X_{s_\delta}^{\varepsilon,i,N}|^8+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^8+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^{8q_1}+\mathbb{W}_2(\mu_{s_\delta}^{\varepsilon,N},\delta_0)^8\\&\quad+\mathbb{W}_2(\mu_{s_\delta-\tau}^{\varepsilon,N},\delta_0)^8+|X_{s_\delta}^{\varepsilon,i,N}|^4+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^4+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^{4(r_3+1)}\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta}, \delta_0)^4 + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta-\tau}, \delta_0)^4\Big\}\mathrm{d}s\Big\}^{\frac{1}{2}} \\
\le\ & C\varepsilon\Big\{\mathbb{E}\int_0^T \Big\{1 + |X^{\varepsilon,i,N}_{s_\delta}|^8 + |X^{\varepsilon,i,N}_{s_\delta-\tau}|^8 + |X^{\varepsilon,i,N}_{s_\delta-\tau}|^{8q_1} + \Big(\frac{1}{N}\sum_{j=1}^N \mathbb{E}|X^{\varepsilon,j,N}_{s_\delta}|^8\Big) \\
&+ \Big(\frac{1}{N}\sum_{j=1}^N \mathbb{E}|X^{\varepsilon,j,N}_{s_\delta-\tau}|^8\Big) + |X^{\varepsilon,i,N}_{s_\delta}|^4 + |X^{\varepsilon,i,N}_{s_\delta-\tau}|^4 + |X^{\varepsilon,i,N}_{s_\delta-\tau}|^{4(r_3+1)} \\
&+ \Big(\frac{1}{N}\sum_{j=1}^N \mathbb{E}|X^{\varepsilon,j,N}_{s_\delta}|^4\Big) + \Big(\frac{1}{N}\sum_{j=1}^N \mathbb{E}|X^{\varepsilon,j,N}_{s_\delta-\tau}|^4\Big)\Big\}\mathrm{d}s\Big\}^{\frac{1}{2}} \\
\le\ & C\varepsilon,
\end{aligned}
$$

where the last step is due to Lemma 6.6. Hence, the desired result holds by taking ε sufficiently small and n, N sufficiently large. □

Lemma 6.12. *Let Assumptions 6.1, 6.2, 6.3 and 6.6 hold. Then, for any initial value $X^\varepsilon_0 = \xi \in L^0_{p_1q_1}(\mathscr{C})$,*

$$
\begin{aligned}
\delta\sum_{k=1}^n & B^*(X^{\varepsilon,i,N}_{(k-1)\delta}, X^{\varepsilon,i,N}_{(k-1)\delta-\tau}, \theta_0, \theta)\Lambda^{-1}(X^{\varepsilon,i,N}_{(k-1)\delta}, X^{\varepsilon,i,N}_{(k-1)\delta-\tau}) \\
&\times B(X^{\varepsilon,i,N}_{(k-1)\delta}, X^{\varepsilon,i,N}_{(k-1)\delta-\tau}, \theta_0, \theta) \to \Pi(\theta)
\end{aligned}
$$

in L^1 as $\varepsilon \to 0$, $N \to \infty$ and $\delta \to 0$, where $\Pi(\theta)$ is defined in (6.12).

Proof. Obviously,

$$
\begin{aligned}
&\Phi^{i,N,(2)}_{n,\varepsilon}(\theta) \\
&:=\delta\sum_{k=1}^n B^*(X^{\varepsilon,i,N}_{(k-1)\delta}, X^{\varepsilon,i,N}_{(k-1)\delta-\tau}, \theta_0, \theta)\Lambda^{-1}(X^{\varepsilon,i,N}_{(k-1)\delta}, X^{\varepsilon,i,N}_{(k-1)\delta-\tau}) \\
&\quad\times B(X^{\varepsilon,i,N}_{(k-1)\delta}, X^{\varepsilon,i,N}_{(k-1)\delta-\tau}, \theta_0, \theta) \\
&= \int_0^T B^*(X^{\varepsilon,i,N}_{s_\delta}, X^{\varepsilon,i,N}_{s_\delta-\tau}, \theta_0, \theta)\Lambda^{-1}(X^{\varepsilon,i,N}_{s_\delta}, X^{\varepsilon,i,N}_{s_\delta-\tau})B(X^{\varepsilon,i,N}_{s_\delta}, X^{\varepsilon,i,N}_{s_\delta-\tau}, \theta_0, \theta)\mathrm{d}s.
\end{aligned}
$$

Thus, by calculating directly, one has

$$
\begin{aligned}
&\Phi^{i,N,(2)}_{n,\varepsilon}(\theta) - \Pi(\theta) \\
&= \int_0^T \Big\{B(X^{\varepsilon,i,N}_{s_\delta}, X^{\varepsilon,i,N}_{s_\delta-\tau}, \theta_0, \theta) - B(X^{0,i}_s, X^{0,i}_{s-\tau}, \theta_0, \theta)\Big\}^*\Lambda^{-1}(X^{\varepsilon,i,N}_{s_\delta}, X^{\varepsilon,i,N}_{s_\delta-\tau}) \\
&\quad\times B(X^{\varepsilon,i,N}_{s_\delta}, X^{\varepsilon,i,N}_{s_\delta-\tau}, \theta_0, \theta)\mathrm{d}s + \int_0^T B^*(X^{0,i}_s, X^{0,i}_{s-\tau}, \theta_0, \theta)\Big\{\Lambda^{-1}(X^{\varepsilon,i,N}_{s_\delta}, X^{\varepsilon,i,N}_{s_\delta-\tau})
\end{aligned}
$$

$$
\begin{aligned}
&- \Lambda^{-1}(X_s^{0,i}, X_{s-\tau}^{0,i})\Big\}B(X_{s_\delta}^{\varepsilon,i,N}, X_{s_\delta-\tau}^{\varepsilon,i,N}, \theta_0, \theta)\mathrm{d}s + \int_0^T B^*(X_s^{0,i}, X_{s-\tau}^{0,i}, \theta_0, \theta) \\
&\times \Lambda^{-1}(X_s^{0,i}, X_{s-\tau}^{0,i})\Big\{B(X_{s_\delta}^{\varepsilon,i,N}, X_{s_\delta-\tau}^{\varepsilon,i,N}, \theta_0, \theta) - B(X_s^{0,i}, X_{s-\tau}^{0,i}, \theta_0, \theta)\Big\}\mathrm{d}s \\
=: &\sum_{i=1}^{3} J_i.
\end{aligned}
$$

In addition, for any $x_i, y_i \in \mathbb{R}^d$ and $\mu_{x_i}, \mu_{y_i} \in \mathscr{P}_2(\mathbb{R}^d)$, $i = 1, 2$, it follows from Assumption 6.1 that

$$
\begin{aligned}
&|B(x_1, y_1, \theta_0, \theta) - B(x_2, y_2, \theta_0, \theta)| \\
&\quad \le C\{|x_1 - x_2| + (1 + |y_1|^{r_1} + |y_2|^{r_1})|y_1 - y_2| \\
&\qquad + \mathbb{W}_2(\mu_{x_1}, \mu_{x_2}) + \mathbb{W}_2(\mu_{y_1}, \mu_{y_2})\}.
\end{aligned} \tag{6.52}
$$

This leads to

$$
\begin{aligned}
&|B(X_{s_\delta}^{\varepsilon,i,N}, X_{s_\delta-\tau}^{\varepsilon,i,N}, \theta_0, \theta) - B(X_s^{0,i}, X_{s-\tau}^{0,i}, \theta_0, \theta)| \\
&\quad \le C\{|X_{s_\delta}^{\varepsilon,i,N} - X_s^{0,i}| + (1 + |X_{s_\delta-\tau}^{\varepsilon,i,N}|^{r_1} + |X_{s-\tau}^{0,i}|^{r_1})|X_{s_\delta-\tau}^{\varepsilon,i,N} - X_{s-\tau}^{0,i}| \\
&\qquad + \mathbb{W}_2(\mu_{s_\delta}^{\varepsilon,N}, \mu_s^{0,i}) + \mathbb{W}_2(\mu_{s_\delta-\tau}^{\varepsilon,N}, \mu_{s-\tau}^{0,i})\}.
\end{aligned} \tag{6.53}
$$

Set $q_2 = (r_1 + 1) \vee (r_3 + 1)$. By (6.49) and (6.51), one has

$$
\begin{aligned}
&\|\Lambda^{-1}(X_{s_\delta}^{\varepsilon,i,N}, X_{s_\delta-\tau}^{\varepsilon,i,N})\| \cdot |B(X_{s_\delta}^{\varepsilon,i,N}, X_{s_\delta-\tau}^{\varepsilon,i,N}, \theta_0, \theta)| \\
&\quad \le C\Big\{1 + |X_{s_\delta}^{\varepsilon,i,N}|^2 + |X_{s_\delta-\tau}^{\varepsilon,i,N}|^2 + |X_{s_\delta-\tau}^{\varepsilon,i,N}|^{2q_2} \\
&\qquad + \frac{1}{N}\sum_{j=1}^{N} |X_{s_\delta}^{\varepsilon,j,N}|^2 + \frac{1}{N}\sum_{j=1}^{N} |X_{s_\delta-\tau}^{\varepsilon,j,N}|^2\Big\},
\end{aligned} \tag{6.54}
$$

where we have used (6.46). Then, the Hölder inequality implies that

$$
\begin{aligned}
&\mathbb{E}|J_1| \\
&\le C\mathbb{E}\Big|\int_0^T \Big\{|X_{s_\delta}^{\varepsilon,i,N} - X_s^{0,i}| + (1 + |X_{s_\delta-\tau}^{\varepsilon,i,N}|^{r_1} + |X_{s-\tau}^{0,i}|^{r_1})|X_{s_\delta-\tau}^{\varepsilon,i,N} - X_{s-\tau}^{0,i}| \\
&\quad + \mathbb{W}_2(\mu_{s_\delta}^{\varepsilon,N}, \mu_s^{0,i}) + \mathbb{W}_2(\mu_{s_\delta-\tau}^{\varepsilon,N}, \mu_{s-\tau}^{0,i})\Big\}\Big\{1 + |X_{s_\delta}^{\varepsilon,i,N}|^2 + |X_{s_\delta-\tau}^{\varepsilon,i,N}|^2 + |X_{s_\delta-\tau}^{\varepsilon,i,N}|^{2q_2} \\
&\quad + \frac{1}{N}\sum_{j=1}^{N} |X_{s_\delta}^{\varepsilon,j,N}|^2 + \frac{1}{N}\sum_{j=1}^{N} |X_{s_\delta-\tau}^{\varepsilon,j,N}|^2\Big\}\mathrm{d}s\Big| \\
&\le C\mathbb{E}\Big|\int_0^T \Big\{|X_{s_\delta}^{\varepsilon,i,N} - X_s^{0,i}|\Big(1 + |X_{s_\delta}^{\varepsilon,i,N}|^2 + |X_{s_\delta-\tau}^{\varepsilon,i,N}|^{2q_2} + \frac{1}{N}\sum_{j=1}^{N} |X_{s_\delta}^{\varepsilon,j,N}|^2 \\
&\quad + \frac{1}{N}\sum_{j=1}^{N} |X_{s_\delta-\tau}^{\varepsilon,j,N}|^2\Big) + |X_{s_\delta-\tau}^{\varepsilon,i,N} - X_{s-\tau}^{0,i}|\Big(1 + |X_{s_\delta}^{\varepsilon,i,N}|^2 + |X_{s_\delta-\tau}^{\varepsilon,i,N}|^{2q_2}
\end{aligned} \tag{6.55}
$$

$$+\frac{1}{N}\sum_{j=1}^{N}|X_{s_\delta}^{\varepsilon,j,N}|^2+\frac{1}{N}\sum_{j=1}^{N}|X_{s_\delta-\tau}^{\varepsilon,j,N}|^2\Big)^2+\Big(\mathbb{W}_2(\mu_{s_\delta}^{\varepsilon,N},\mu_s^{0,i})+\mathbb{W}_2(\mu_{s_\delta-\tau}^{\varepsilon,N},\mu_{s-\tau}^{0,i})\Big)$$

$$\times\Big(1+|X_{s_\delta}^{\varepsilon,i,N}|^2+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^{2q_2}+\frac{1}{N}\sum_{j=1}^{N}|X_{s_\delta}^{\varepsilon,j,N}|^2+\frac{1}{N}\sum_{j=1}^{N}|X_{s_\delta-\tau}^{\varepsilon,j,N}|^2\Big)\Big\}\mathrm{d}s\Big|$$

$$\leq C\int_0^T\Big\{\sqrt{\mathbb{E}|X_{s_\delta}^{\varepsilon,i,N}-X_s^{0,i}|^2}+\Big(\mathbb{E}\mathbb{W}_2(\mu_{s_\delta}^{\varepsilon,N},\mu_s^{0,i})^2+\mathbb{E}\mathbb{W}_2(\mu_{s_\delta-\tau}^{\varepsilon,N},\mu_{s-\tau}^{0,i})^2\Big)^{\frac{1}{2}}\Big\}\mathrm{d}s$$

$$\leq C\int_0^T\Big\{(C_N+C_N^{\frac{1}{2}})+C\delta(\delta+\varepsilon^2)\Big\}^{\frac{1}{2}}\mathrm{d}s,$$

where, in the first step we used (6.53) and (6.54), and in the last step we utilized lemmas 6.8 and 6.10. To obtain the estimate of J_2, we firstly seek some inequalities of the integrands. By (6.22), we find out

$$\begin{aligned}&|B^*(X_s^{0,i},X_{s-\tau}^{0,i},\theta_0,\theta)|\\&\quad\leq L\Big\{1+|X_s^{0,i}|+|X_{s-\tau}^{0,i}|+|X_{s-\tau}^{0,i}|^{(r_1+1)}+\mathbb{W}_2(\mu_s^{0,i},\delta_0)+\mathbb{W}_2(\mu_{s-\tau}^{0,i},\delta_0)\Big\}.\end{aligned}\tag{6.56}$$

By means of Assumption 6.3, one gets

$$\begin{aligned}&\|\Lambda^{-1}(X_{s_\delta}^{\varepsilon,i,N},X_{s_\delta-\tau}^{\varepsilon,i,N})-\Lambda^{-1}(X_s^{0,i},X_{s-\tau}^{0,i})\|\\&\quad\leq C_3\Big\{|X_{s_\delta}^{\varepsilon,i,N}-X_s^{0,i}|+(1+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^{r_3}+|X_{s-\tau}^{0,i}|^{r_3})|X_{s_\delta-\tau}^{\varepsilon,i,N}-X_{s_\delta-\tau}^{0,i}|\\&\qquad+\mathbb{W}_2(\mu_{s_\delta}^{\varepsilon,N},\mu_s^{0,i})+\mathbb{W}_2(\mu_{s_\delta-\tau}^{\varepsilon,N},\mu_{s-\tau}^{0,i})\Big\},\end{aligned}$$

$$\begin{aligned}&\mathbb{E}\mathbb{W}_2(\mu_{s_\delta}^{\varepsilon,N},\mu_s^{0,i})^2+\mathbb{E}\mathbb{W}_2(\mu_{s_\delta-\tau}^{\varepsilon,N},\mu_{s-\tau}^{0,i})^2\\&\quad\leq\mathbb{E}|X_{s_\delta}^{\varepsilon,i,N}-X_{s_\delta}^{\varepsilon,i}|^2+\mathbb{E}\mathbb{W}_2(\tilde{\mu}_{s_\delta}^{\varepsilon,N},\mu_{s_\delta}^{\varepsilon})^2+\mathbb{E}|X_{s_\delta-\tau}^{\varepsilon,i,N}-X_{s_\delta-\tau}^{\varepsilon,i}|^2\\&\qquad+\mathbb{E}\mathbb{W}_2(\tilde{\mu}_{s_\delta-\tau}^{\varepsilon,N},\mu_{s_\delta-\tau}^{\varepsilon})^2+\mathbb{E}|X_{s_\delta}^{\varepsilon,i}-X_s^{0,i}|^2+\mathbb{E}|X_{s_\delta-\tau}^{\varepsilon,i}-X_{s-\tau}^{0,i}|^2\end{aligned}$$

and

$$\mathbb{E}|X_{s_\delta}^{\varepsilon,i,N}-X_s^{0,i}|^2\leq 2\mathbb{E}|X_{s_\delta}^{\varepsilon,i,N}-X_s^{\varepsilon,i}|^2+2\mathbb{E}|X_{s_\delta}^{\varepsilon,i}-X_s^{0,i}|^2.\tag{6.57}$$

In view of the results obtained above, we find out

$$\begin{aligned}&\mathbb{E}|J_2|\\&\leq C\mathbb{E}\Big|\int_0^T\Big\{1+|X_s^{0,i}|+|X_{s-\tau}^{0,i}|+|X_{s-\tau}^{0,i}|^{(r_1+1)}+\mathbb{W}_2(\mu_s^{0,i},\delta_0)+\mathbb{W}_2(\mu_{s-\tau}^{0,i},\delta_0)\Big\}\\&\quad\times\Big\{|X_{s_\delta}^{\varepsilon,i,N}-X_s^{0,i}|+(1+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^{r_3}+|X_{s-\tau}^{0,i}|^{r_3})|X_{s_\delta-\tau}^{\varepsilon,i,N}-X_{s_\delta-\tau}^{0,i}|\\&\quad+\mathbb{W}_2(\mu_{s_\delta}^{\varepsilon,N},\mu_s^{0,i})+\mathbb{W}_2(\mu_{s_\delta-\tau}^{\varepsilon,N},\mu_{s-\tau}^{0,i})\Big\}\times\Big\{1+|X_{s_\delta}^{\varepsilon,i,N}|+|X_{s_\delta-\tau}^{\varepsilon,i,N}|+|X_{s_\delta-\tau}^{\varepsilon,i,N}|^{(r_1+1)}\end{aligned}$$

$$
\begin{aligned}
&\quad + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta},\delta_0) + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta-\tau},\delta_0)\Big\}\mathrm{d}s\Big| \\
&\le C\mathbb{E}\int_0^T \Big|\Big\{|X^{\varepsilon,i,N}_{s_\delta} - X^{0,i}_s|\Big(1+|X^{\varepsilon,i,N}_{s_\delta}| + |X^{\varepsilon,i,N}_{s_\delta-\tau}|^{q_2} + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta},\delta_0) \\
&\quad + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta-\tau},\delta_0)\Big) + \Big(1+|X^{\varepsilon,i,N}_{s_\delta}| + |X^{\varepsilon,i,N}_{s_\delta-\tau}|^{q_1} + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta},\delta_0) \\
&\quad + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta-\tau},\delta_0)\Big)^2 |X^{\varepsilon,i,N}_{s_\delta-\tau} - X^{0,i}_{s_\delta-\tau}| + \Big(\mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta},\mu^{0,i}_s) + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta-\tau},\mu^{0,i}_{s-\tau})\Big) \\
&\quad \times \Big(1+|X^{\varepsilon,i,N}_{s_\delta}| + |X^{\varepsilon,i,N}_{s_\delta-\tau}|^{q_2} + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta},\delta_0) + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta-\tau},\delta_0)\Big)\Big\}\Big|\mathrm{d}s \qquad (6.58) \\
&\le C\int_0^T \Big\{\sqrt{\mathbb{E}|X^{\varepsilon,i,N}_{s_\delta} - X^{0,i}_s|^2} + \sqrt{\mathbb{E}|X^{\varepsilon,i,N}_{s_\delta-\tau} - X^{0,i}_{s_\delta-\tau}|^2} + \Big(\mathbb{E}\mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta},\mu^{0,i}_s)^2 \\
&\quad + \mathbb{E}\mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta-\tau},\mu^{0,i}_{s-\tau})^2\Big)^{\frac{1}{2}}\Big\}\mathrm{d}s \\
&\le C\int_0^T \Big\{(C_N + C_N^{\frac{1}{2}}) + C\delta(\delta+\varepsilon^2)\Big\}^{\frac{1}{2}}\mathrm{d}s,
\end{aligned}
$$

where, in the last step we have used the inequalities (6.27) and (6.37). Moreover, the inequality (6.48) leads to

$$
\begin{aligned}
&\|\Lambda^{-1}(X^{0,i}_s, X^{0,i}_{s-\tau})\| \\
&\quad \le \overline{L}\Big\{1+|X^{0,i}_s| + |X^{0,i}_{s-\tau}| + |X^{0,i}_{s-\tau}|^{(r_3+1)} + \mathbb{W}_2(\mu^{0,i}_s,\delta_0) + \mathbb{W}_2(\mu^{0,i}_{s-\tau},\delta_0)\Big\},
\end{aligned} \tag{6.59}
$$

which, together with (6.53) and (6.56), further leads to

$$
\begin{aligned}
&\mathbb{E}|J_3| \\
&\quad \le C\mathbb{E}\Big|\int_0^T \Big\{1+|X^{0,i}_s| + |X^{0,i}_{s-\tau}| + |X^{0,i}_{s-\tau}|^{q_2} + \mathbb{W}_2(\mu^{0,i}_s,\delta_0) + \mathbb{W}_2(\mu^{0,i}_{s-\tau},\delta_0)\Big\}^2 \\
&\qquad \times \Big\{|X^{\varepsilon,i,N}_{s_\delta} - X^{0,i}_s| + (1+|X^{\varepsilon,i,N}_{s_\delta-\tau}|^{r_1} + |X^{0,i}_{s-\tau}|^{r_1})|X^{\varepsilon,i,N}_{s_\delta-\tau} - X^{0,i}_{s-\tau}| \\
&\qquad + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta},\mu^{0,i}_s) + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta-\tau},\mu^{0,i}_{s-\tau})\Big\}\mathrm{d}s\Big| \\
&\quad \le C\mathbb{E}\int_0^T \Big\{|X^{\varepsilon,i,N}_{s_\delta} - X^{0,i}_s| + (1+|X^{\varepsilon,i,N}_{s_\delta-\tau}|^{r_1} + |X^{0,i}_{s-\tau}|^{r_1})|X^{\varepsilon,i,N}_{s_\delta-\tau} - X^{0,i}_{s-\tau}| \\
&\qquad + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta},\mu^{0,i}_s) + \mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta-\tau},\mu^{0,i}_{s-\tau})\Big\}\mathrm{d}s \qquad (6.60) \\
&\quad \le C\int_0^T \Big\{\mathbb{E}|X^{\varepsilon,i,N}_{s_\delta} - X^{0,i}_s| + (1+\mathbb{E}|X^{\varepsilon,i,N}_{s_\delta-\tau}|^{2r_1})^{\frac{1}{2}}\sqrt{\mathbb{E}|X^{\varepsilon,i,N}_{s_\delta-\tau} - X^{0,i}_{s-\tau}|^2} \\
&\qquad + \mathbb{E}\mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta},\mu^{0,i}_s) + \mathbb{E}\mathbb{W}_2(\mu^{\varepsilon,N}_{s_\delta-\tau},\mu^{0,i}_{s-\tau})\Big\}\mathrm{d}s \\
&\quad \le C\int_0^T \Big\{(C_N + C_N^{\frac{1}{2}} + \delta(\delta+\varepsilon^2))^{\frac{1}{2}} + C_N + C_N^{\frac{1}{2}} + \delta(\delta+\varepsilon^2)\Big\}\mathrm{d}s.
\end{aligned}
$$

Therefore, from (6.55), (6.58) and (6.60), we conclude that the desired result holds. □

Proof of Theorem 6.2.

$$\begin{aligned}
\Phi_{n,\varepsilon}^{i,N}(\theta) =&\varepsilon^2(\Psi_{n,\varepsilon}^{i,N}(\theta)-\Psi_{n,\varepsilon}^{i,N}(\theta_0))\\
=&\delta^{-1}\sum_{k=1}^{n}\Big\{\Big(P_k^{\varepsilon,i,N}(\theta_0)+b(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\mu_{(k-1)\delta}^{\varepsilon,N},\mu_{(k-1)\delta-\tau}^{\varepsilon,N},\theta_0)\delta\\
&-b(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\mu_{(k-1)\delta}^{\varepsilon,N},\mu_{(k-1)\delta-\tau}^{\varepsilon,N},\theta)\delta\Big)^*\Lambda^{-1}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})\\
&\times\Big(P_k^{\varepsilon,i,N}(\theta_0)+\delta(b(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\mu_{(k-1)\delta}^{\varepsilon,N},\mu_{(k-1)\delta-\tau}^{\varepsilon,N},\theta_0)\\
&-b(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\mu_{(k-1)\delta}^{\varepsilon,N},\mu_{(k-1)\delta-\tau}^{\varepsilon,N},\theta)\Big)-(P_k^{\varepsilon,i,N}(\theta_0))^*\\
&\times\Lambda^{-1}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})P_k^{\varepsilon,i,N}(\theta_0)\Big\}\\
=&2\sum_{k=1}^{n}B^*(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta)\Lambda^{-1}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})P_k^{\varepsilon,i,N}(\theta_0)\\
&+\delta\sum_{k=1}^{n}B^*(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta)\Lambda^{-1}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})\\
&\times B^*(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta)\\
=:&2\Phi_{n,\varepsilon}^{i,N,(1)}(\theta)+\Phi_{n,\varepsilon}^{i,N,(2)}(\theta).
\end{aligned}$$

In view of Lemmas 6.11 and 6.12, together with the Chebyshev inequality, we deduce that

$$\sup_{\theta\in\Theta}|-\Phi_{n,\varepsilon}^{i,N}(\theta)-(-\Pi(\theta))|\to 0 \quad \text{in probability.} \tag{6.61}$$

According to (6.9), we find out $0=\Phi_{n,\varepsilon}^{i,N}(\theta_0)\geq\Phi_{n,\varepsilon}^{i,N}(\hat\theta_{n,\varepsilon}^{i,N})$, i.e., $0=-\Phi_{n,\varepsilon}^{i,N}(\theta_0)\leq -\Phi_{n,\varepsilon}^{i,N}(\hat\theta_{n,\varepsilon}^{i,N})$. In addition, due to $\Pi(\cdot)\geq 0$, we get

$$\sup_{|\theta-\theta_0|\geq\iota}(-\Pi(\theta))<-\Pi(\theta_0)=0, \qquad \text{for any } \ \iota>0. \tag{6.62}$$

In terms of Theorem 1.5, and combining with (6.61) and (6.62), we deduce that $\hat\theta_{n,\varepsilon}^{i,N}\to\theta_0$ in probability as $N,n\to\infty$ and $\varepsilon\to 0$. □

6.5.3 Proof of Theorem 6.3.

To make the deduction of the asymptotic distribution of the LSE $\hat\theta_{n,\varepsilon}^{i,N}$ clearer, we divide the proof of Theorem 6.3 into several auxiliary lemmas.

Lemma 6.13. *Assume that Assumptions 6.1– 6.6 hold. Then, for* $X_0^\varepsilon = \xi \in L^0_{p_1q_1}(\mathscr{C})$,

$$\int_0^T \Upsilon(X_{t_\delta}^{\varepsilon,i,N}, X_{t_\delta-\tau}^{\varepsilon,i,N}, \theta_0)\mathrm{d}W_t^i \to \int_0^T \Upsilon(X_t^{0,i}, X_{t-\tau}^{0,i}, \theta_0)\mathrm{d}W_t^i \quad \mathbb{P}-a.s. \tag{6.63}$$

as $\varepsilon \to 0$, $\delta \to 0$ *and* $N \to \infty$. *Moreover,*

$$\varepsilon^{-1}(\nabla_\theta \Phi^{i,N}_{n,\varepsilon})(\theta_0) \to -2\int_0^T \Upsilon(X_t^{0,i}, X_{t-\tau}^{0,i}, \theta_0)\, W_t^i \quad \mathbb{P}-a.s.$$

whenever $\varepsilon \to 0$ *and* $n, N \to \infty$.

Proof. In view of Assumption 6.4, we see that, for any $x, y \in \mathbb{R}^d$ and $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, there exists a constant $C > 0$ such that

$$\begin{aligned}&\sup_{\theta\in\overline{\Theta}} \|(\nabla_\theta b)(x, y, \mu, \nu, \theta)\| \\ &\quad\le C\{1 + |x| + |y| + |y|^{r_4+1} + \mathbb{W}_2(\mu, \delta_0) + \mathbb{W}_2(\nu, \delta_0)\}.\end{aligned} \tag{6.64}$$

We first claim that

$$\int_0^T \|\Upsilon(X_{t_\delta}^{\varepsilon,i,N}, X_{t_\delta-\tau}^{\varepsilon,i,N}, \theta_0) - \Upsilon(X_t^{0,i}, X_{t-\tau}^{0,i}, \theta_0)\|^2 \mathrm{d}t \to 0 \qquad \mathbb{P}-\text{a.s.} \tag{6.65}$$

as $\varepsilon \to 0$, $\delta \to 0$ and $N \to \infty$. According to (6.13), one gets

$$\begin{aligned}
&\|\Upsilon(X_{t_\delta}^{\varepsilon,i,N}, X_{t_\delta-\tau}^{\varepsilon,i,N}, \theta_0) - \Upsilon(X_t^{0,i}, X_{t-\tau}^{0,i}, \theta_0)\|^2 \\
&\quad\le 3\|\{(\nabla_\theta b)^*(X_{t_\delta}^{\varepsilon,i,N}, X_{t_\delta-\tau}^{\varepsilon,i,N}, \mu_{t_\delta}^{\varepsilon,N}, \mu_{t_\delta-\tau}^{\varepsilon,N}, \theta_0) - (\nabla_\theta b)^*(X_t^{0,i}, X_{t-\tau}^{0,i}, \mu_t^{0,i}, \mu_{t-\tau}^{0,i}, \theta_0)\} \\
&\qquad\times \Lambda^{-1}(X_{t_\delta}^{\varepsilon,i,N}, X_{t_\delta-\tau}^{\varepsilon,i,N})\sigma(X_{t_\delta}^{\varepsilon,i,N}, X_{t_\delta-\tau}^{\varepsilon,i,N}, \mu_{t_\delta}^{\varepsilon,N}, \mu_{t_\delta-\tau}^{\varepsilon,N})\|^2 \\
&\quad+ 3\|(\nabla_\theta b)^*(X_t^{0,i}, X_{t-\tau}^{0,i}, \mu_t^{0,i}, \mu_{t-\tau}^{0,i}, \theta_0) \times \{\Lambda^{-1}(X_{t_\delta}^{\varepsilon,i,N}, X_{t_\delta-\tau}^{\varepsilon,i,N}) \\
&\qquad- \Lambda^{-1}(X_t^{0,i}, X_{t-\tau}^{0,i})\}\sigma(X_{t_\delta}^{\varepsilon,i,N}, X_{t_\delta-\tau}^{\varepsilon,i,N}, \mu_{t_\delta}^{\varepsilon,N}, \mu_{t_\delta-\tau}^{\varepsilon,N})\|^2 \\
&\quad+ 3\|(\nabla_\theta b)^*(X_t^{0,i}, X_{t-\tau}^{0,i}, \mu_t^0, \mu_{t-\tau}^0, \theta_0)\Lambda^{-1}(X_t^{0,i}, X_{t-\tau}^{0,i}) \\
&\qquad\times \{\sigma(X_{t_\delta}^{\varepsilon,i,N}, X_{t_\delta-\tau}^{\varepsilon,i,N}, \mu_{t_\delta}^{\varepsilon,N}, \mu_{t_\delta-\tau}^{\varepsilon,N}) - \sigma(X_t^{0,i}, X_{t-\tau}^{0,i}, \mu_t^{0,i}, \mu_{t-\tau}^{0,i})\}\|^2 \\
&\quad=: \sum_{k=1}^3 G_k.
\end{aligned}$$

For the first term G_1, from Assumption 6.4 we first give the below result

$$\begin{aligned}
&\|(\nabla_\theta b)^*(X_{t_\delta}^{\varepsilon,i,N}, X_{t_\delta-\tau}^{\varepsilon,i,N}, \mu_{t_\delta}^{\varepsilon,N}, \mu_{t_\delta-\tau}^{\varepsilon,N}, \theta_0) - (\nabla_\theta b)^*(X_t^{0,i}, X_{t-\tau}^{0,i}, \mu_t^{0,i}, \mu_{t-\tau}^{0,i}, \theta_0)\|^2 \\
&\quad\le C\Big\{|X_{t_\delta}^{\varepsilon,i,N} - X_t^{0,i}|^2 + (1 + |X_{t_\delta-\tau}^{\varepsilon,i,N}|^{2r_4} + |X_{t-\tau}^{0,i}|^{2r_4})|X_{t_\delta-\tau}^{\varepsilon,i,N} - X_{t-\tau}^{0,i}|^2 \\
&\qquad+ \mathbb{W}_2(\mu_{t_\delta}^{\varepsilon,N}, \mu_t^{0,i})^2 + \mathbb{W}_2(\mu_{t_\delta-\tau}^{\varepsilon,N}, \mu_{t-\tau}^{0,i})^2\Big\}.
\end{aligned}$$

This, combining (6.50) with (6.51), leads to

$$
\begin{aligned}
&G_1\\
&\leq C\Big\{1+|X_{t_\delta}^{\varepsilon,i,N}|^2+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^2+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{2(r_2+1)}+\mathbb{W}_2(\mu_{t_\delta}^{\varepsilon,N},\delta_0)^2+\mathbb{W}_2(\mu_{t_\delta-\tau}^{\varepsilon,N},\delta_0)^2\Big\}\\
&\quad\times\Big\{1+|X_{t_\delta}^{\varepsilon,i,N}|^2+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^2+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{2(r_3+1)}+\mathbb{W}_2(\mu_{t_\delta}^{\varepsilon,N},\delta_0)^2+\mathbb{W}_2(\mu_{t_\delta-\tau}^{\varepsilon,N},\delta_0)^2\Big\}\\
&\quad\times\Big\{|X_{t_\delta}^{\varepsilon,i,N}-X_t^{0,i}|^2+|X_{t_\delta-\tau}^{\varepsilon,i,N}-X_{t-\tau}^{0,i}|^2(1+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{2r_4}+|X_{t-\tau}^{0,i}|^{2r_4})\\
&\quad+\mathbb{W}_2(\mu_{t_\delta}^{\varepsilon,N},\mu_t^{0,i})^2+\mathbb{W}_2(\mu_{t_\delta-\tau}^{\varepsilon,N},\mu_{t-\tau}^{0,i})^2\Big\} \qquad (6.66)\\
&\leq C|X_{t_\delta}^{\varepsilon,i,N}-X_t^{0,i}|^2\Big\{1+|X_{t_\delta}^{\varepsilon,i,N}|^4+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{q_3}+\mathbb{W}_2(\mu_{t_\delta}^{\varepsilon,N},\delta_0)^4+\mathbb{W}_2(\mu_{t_\delta-\tau}^{\varepsilon,N},\delta_0)^4\Big\}\\
&\quad+C|X_{t_\delta-\tau}^{\varepsilon,i,N}-X_{t-\tau}^{0,i}|^2\Big\{1+|X_{t_\delta}^{\varepsilon,i,N}|^8+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{2q_3}+\mathbb{W}_2(\mu_{t_\delta}^{\varepsilon,N},\delta_0)^8+\mathbb{W}_2(\mu_{t_\delta-\tau}^{\varepsilon,N},\delta_0)^8\Big\}\\
&\quad+C\mathbb{W}_2(\mu_{t_\delta}^{\varepsilon,N},\mu_t^{0,i})^2\Big\{1+|X_{t_\delta}^{\varepsilon,i,N}|^4+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{q_3}+\mathbb{W}_2(\mu_{t_\delta}^{\varepsilon,N},\delta_0)^4+\mathbb{W}_2(\mu_{t_\delta-\tau}^{\varepsilon,N},\delta_0)^4\Big\}\\
&\quad+C\mathbb{W}_2(\mu_{t_\delta-\tau}^{\varepsilon,N},\mu_{t-\tau}^{0,i})^2\Big\{1+|X_{t_\delta}^{\varepsilon,i,N}|^4+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{q_3}+\mathbb{W}_2(\mu_{t_\delta}^{\varepsilon,N},\delta_0)^4+\mathbb{W}_2(\mu_{t_\delta-\tau}^{\varepsilon,N},\delta_0)^4\Big\}\\
&=:\sum_{k=1}^4\Sigma_k,
\end{aligned}
$$

where $q_3=4((r_2+1)\vee(r_3+1))\vee(2r_4)$.

For any $\rho>0$ and $i\in\mathcal{S}_N$, in view of the Chebyshev inequality and (6.57), we arrive at

$$
\begin{aligned}
&\mathbb{P}\Big(\int_0^T\|\Sigma_1\|\mathrm{d}t\geq\rho\Big)\\
&\quad\leq\mathbb{P}\Big(C\int_0^T|X_{t_\delta}^{\varepsilon,i,N}-X_t^{0,i}|^2\Big\{1+|X_{t_\delta}^{\varepsilon,i,N}|^4+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{q_3}+\mathbb{W}_2(\mu_{t_\delta}^{\varepsilon,N},\delta_0)^4\\
&\qquad+\mathbb{W}_2(\mu_{t_\delta-\tau}^{\varepsilon,N},\delta_0)^4\Big\}\mathrm{d}t\geq\rho\Big)\\
&\quad\leq\frac{C}{\rho}\int_0^T\Big(\mathbb{E}\Big\{1+|X_{t_\delta}^{\varepsilon,i,N}|^8+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{2q_3}+\mathbb{W}_2(\mu_{t_\delta}^{\varepsilon,N},\delta_0)^8\\
&\qquad+\mathbb{W}_2(\mu_{t_\delta-\tau}^{\varepsilon,N},\delta_0)^8\Big\}\Big)^{\frac12}\Big(\mathbb{E}|X_{t_\delta}^{\varepsilon,i,N}-X_t^{0,i}|^4\Big)^{\frac12}\mathrm{d}t \qquad (6.67)\\
&\quad\leq\frac{C}{\rho}\int_0^T\Big(1+\mathbb{E}|X_{t_\delta}^{\varepsilon,i,N}|^8+\mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{2q_3}+\frac1N\sum_{j=1}^N\mathbb{E}|X_{t_\delta}^{\varepsilon,j,N}|^8\\
&\qquad+\frac1N\sum_{j=1}^N\mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,j,N}|^8\Big)^{\frac12}\Big(\mathbb{E}|X_{t_\delta}^{\varepsilon,i,N}-X_t^{0,i}|^4\Big)^{\frac12}\mathrm{d}t\to0
\end{aligned}
$$

as $\varepsilon \to 0$, $\delta \to 0$ and $N \to \infty$. Using the same idea like in the above, we get

$$\mathbb{P}\Big(\int_0^T \|\Sigma_2\| \mathrm{d}t \geq \rho \Big) \to 0, \tag{6.68}$$

as $\varepsilon \to 0$, $\delta \to 0$ and $N \to \infty$. At the same time, it follows from the Hölder inequality that

$$\begin{aligned}
&\mathbb{E}\|\Sigma_3\| \\
&\quad \leq C\Big(\mathbb{E}\mathbb{W}_2(\mu^{\varepsilon,N}_{t_\delta}, \mu^{0,i}_t)^4\Big)^{\frac{1}{2}}\Big(1 + \mathbb{E}|X^{\varepsilon,i,N}_{t_\delta}|^8 + \mathbb{E}|X^{\varepsilon,i,N}_{t_\delta-\tau}|^{2q_3} + \mathbb{E}\mathbb{W}_2(\mu^{\varepsilon,N}_{t_\delta}, \delta_0)^8 \\
&\qquad + \mathbb{E}\mathbb{W}_2(\mu^{\varepsilon,N}_{t_\delta-\tau}, \delta_0)^8\Big)^{\frac{1}{2}} \\
&\quad \leq C\Big(\mathbb{E}\mathbb{W}_4(\mu^{\varepsilon,N}_{t_\delta}, \tilde{\mu}^{\varepsilon,N}_{t_\delta})^4 + \mathbb{E}\mathbb{W}_4(\tilde{\mu}^{\varepsilon,N}_{t_\delta}, \mu^{\varepsilon}_{t_\delta})^4 + \mathbb{E}\mathbb{W}_2(\mu^{\varepsilon}_{t_\delta}, \mu^{0,i}_t)^4\Big)^{\frac{1}{2}}\Big(1 + \mathbb{E}|X^{\varepsilon,i,N}_{t_\delta}|^8 \\
&\qquad + \mathbb{E}|X^{\varepsilon,i,N}_{t_\delta-\tau}|^{2q_3} + \frac{1}{N}\sum_{j=1}^N \mathbb{E}|X^{\varepsilon,j,N}_{t_\delta}|^8 + \frac{1}{N}\sum_{j=1}^N \mathbb{E}|X^{\varepsilon,j,N}_{t_\delta-\tau}|^8\Big)^{\frac{1}{2}} \\
&\quad \leq C\Big(\mathbb{E}|Z^{i,N}_{t_\delta}|^4 + C_N + \mathbb{E}|X^{\varepsilon,i}_{t_\delta} - X^{0,i}_t|^4\Big)^{\frac{1}{2}}\Big(1 + \mathbb{E}|X^{\varepsilon,i,N}_{t_\delta}|^8 + \mathbb{E}|X^{\varepsilon,i,N}_{t_\delta-\tau}|^{2q_3} \\
&\qquad + \frac{1}{N}\sum_{j=1}^N \mathbb{E}|X^{\varepsilon,j,N}_{t_\delta}|^8 + \frac{1}{N}\sum_{j=1}^N \mathbb{E}|X^{\varepsilon,j,N}_{t_\delta-\tau}|^8\Big)^{\frac{1}{2}} \\
&\quad \leq C\Big(C_N + C_N^{\frac{1}{2}} + \delta(\delta^{p-1} + \varepsilon^p) + \varepsilon^p\Big)^{\frac{1}{2}} \to 0,
\end{aligned} \tag{6.69}$$

as $\varepsilon \to 0$, $\delta \to 0$ and $N \to \infty$. Similarly, one has

$$\mathbb{E}\|\Sigma_3\| \to 0, \tag{6.70}$$

as $\varepsilon \to 0$, $\delta \to 0$ and $N \to \infty$. Consequently, from (6.66)-(6.69), we get

$$\int_0^T |G_1| \mathrm{d}t \to 0, \quad \text{in probability}, \tag{6.71}$$

when $\varepsilon \to 0$, $\delta \to 0$ and $N \to \infty$.

For the second term G_2, following a similar line of argument as (6.71), we get

$$\int_0^T |G_2| \mathrm{d}t \to 0, \quad \text{in probability}, \tag{6.72}$$

when $\varepsilon \to 0$, $\delta \to 0$ and $N \to \infty$.

For the third term G_3, by Assumption 6.2, (6.59) and (6.64), one has

$$G_3 \leq C\Big\{|X^{\varepsilon,i,N}_{t_\delta} - X^{0,i}_t|^2 + (1 + |X^{\varepsilon,i,N}_{t_\delta-\tau}|^{2r_2} + |X^{0,i}_{t-\tau}|^{2r_2})|X^{\varepsilon,i,N}_{t_\delta-\tau} - X^{0,i}_{t-\tau}|^2$$

$$+\mathbb{W}_2(\mu^{\varepsilon,N}_{t_\delta},\mu^{0,i}_t)^2+\mathbb{W}_2(\mu^{\varepsilon,N}_{t_\delta-\tau},\mu^{0,i}_{t-\tau})^2\Big\}. \tag{6.73}$$

On the other hand, thanks to (6.57) and (6.73), it follows that

$$\begin{aligned}
&\mathbb{P}\Big(\int_0^T G_3\mathrm{d}t\geq\epsilon\Big)\\
&\leq\frac{C}{\epsilon}\int_0^T\Big\{\mathbb{E}|X^{\varepsilon,i,N}_{t_\delta}-X^{0,i}_t|^2+\mathbb{E}(1+|X^{\varepsilon,i,N}_{t_\delta-\tau}|^{2r_2}+|X^{0,i}_{t-\tau}|^{2r_2})|X^{\varepsilon,i,N}_{t_\delta-\tau}-X^{0,i}_{t-\tau}|^2\\
&\quad+\frac{1}{N}\sum_{j=1}^N\mathbb{E}|X^{\varepsilon,j,N}_{t_\delta}-X^{0,j}_t|^2+\frac{1}{N}\sum_{j=1}^N\mathbb{E}|X^{\varepsilon,j,N}_{t_\delta-\tau}-X^{0,j}_{t-\tau}|^2\Big\}\mathrm{d}t\to 0,
\end{aligned}$$

as $\varepsilon\to 0$, $\delta\to 0$ and $N\to\infty$. Hence,

$$\int_0^T G_3\mathrm{d}t\to 0,\quad\text{in probability}, \tag{6.74}$$

as $\varepsilon\to 0$, $\delta\to 0$ and $N\to\infty$. As a consequence, (6.65) follows from (6.71), (6.72) and (6.74). What's more, for any $\rho>0$ and $\epsilon>0$, owing to (6.65), one gets

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\int_0^T(\Upsilon(X^{\varepsilon,i,N}_{t_\delta},X^{\varepsilon,i,N}_{t_\delta-\tau},\theta_0)-\Upsilon(X^{0,i}_t,X^{0,i}_{t-\tau},\theta_0))\mathrm{d}W^i_t\Big|\geq\rho\Big)\\
&\qquad\leq\mathbb{P}\Big(\int_0^T\|\Upsilon(X^{\varepsilon,i,N}_{t_\delta},X^{\varepsilon,i,N}_{t_\delta-\tau},\theta_0)-\Upsilon(X^{0,i}_t,X^{0,i}_{t-\tau},\theta_0)\|^2\mathrm{d}t\geq\rho^2\epsilon\Big)+\epsilon,
\end{aligned}$$

which, together with the arbitrariness of ϵ and (6.65), implies that (6.63) holds. And by a simple calculation, one gets

$$\begin{aligned}
&\varepsilon^{-1}(\nabla_\theta\Phi^{i,N}_{n,\varepsilon})(\theta_0)\\
&=-2\sum_{k=1}^n(\nabla_\theta b)^*(X^{\varepsilon,i,N}_{(k-1)\delta},X^{\varepsilon,i,N}_{(k-1)\delta-\tau},\mu^{\varepsilon,N}_{(k-1)\delta},\mu^{\varepsilon,N}_{(k-1)\delta-\tau},\theta_0)\\
&\quad\times\Lambda^{-1}(X^{\varepsilon,i,N}_{(k-1)\delta},X^{\varepsilon,i,N}_{(k-1)\delta-\tau})\sigma(X^{\varepsilon,i,N}_{(k-1)\delta},X^{\varepsilon,i,N}_{(k-1)\delta-\tau},\mu^{\varepsilon,N}_{(k-1)\delta},\mu^{\varepsilon,N}_{(k-1)\delta-\tau})\delta W^i_k\\
&=-2\int_0^T\Upsilon(X^{\varepsilon,i,N}_{t_\delta},X^{\varepsilon,i,N}_{t_\delta-\tau},\theta_0)\mathrm{d}W^i_t\to-2\int_0^T\Upsilon(X^{0,i}_t,X^{0,i}_{t-\tau},\theta_0)\mathrm{d}W^i_t,\qquad\mathbb{P}-\text{a.s.},
\end{aligned}$$

whenever $\varepsilon\to 0$ and $n,N\to\infty$. □

Lemma 6.14. *Under the assumptions of Theorem 6.3,*

$$(\nabla^{(2)}_\theta\Phi^{i,N}_{n,\varepsilon})(\theta)\to\overline{K}(\theta):=K(\theta)+2I(\theta),\qquad\mathbb{P}-a.s., \tag{6.75}$$

as $\varepsilon\to 0,\ \delta\to 0$ *and* $N\to\infty$, *where* $I(\theta)$ *is defined in* (6.14), *and*

$$\begin{aligned}
K(\theta):&=-2\int_0^T(\nabla^{(2)}_\theta b^*)(X^{0,i}_t,X^{0,i}_{t-\tau},\mu^{0,i}_t,\mu^{0,i}_{t-\tau})\\
&\qquad\circ\Big\{\Lambda^{-1}(X^{0,i}_t,X^{0,i}_{t-\tau})B(X^{0,i}_t,X^{0,i}_{t-\tau},\theta_0,\theta)\Big\}\mathrm{d}t.
\end{aligned}$$

Proof. We first calculate that

$$
\begin{aligned}
&(\nabla_\theta^{(2)}\Phi_{n,\varepsilon}^{i,N})(\theta)\\
&\quad=(\nabla_\theta(\nabla_\theta\Phi_{n,\varepsilon}^{i,N}))(\theta)\\
&\quad=-2\sum_{k=1}^{n}(\nabla_\theta^{(2)}b)^*(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\mu_{(k-1)\delta}^{\varepsilon,N},\mu_{(k-1)\delta-\tau}^{\varepsilon,N},\theta)\\
&\qquad\quad\circ\Big\{\Lambda^{-1}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})P_k^{\varepsilon,i,N}(\theta)\Big\}\\
&\qquad-2\sum_{k=1}^{n}(\nabla_\theta b)^*(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\mu_{(k-1)\delta}^{\varepsilon,N},\mu_{(k-1)\delta-\tau}^{\varepsilon,N},\theta)\\
&\qquad\quad\times\Lambda^{-1}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})(\nabla_\theta P_k^{\varepsilon,i,N})(\theta)\\
&\quad=-2\sum_{k=1}^{n}(\nabla_\theta^{(2)}b)^*(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\mu_{(k-1)\delta}^{\varepsilon,N},\mu_{(k-1)\delta-\tau}^{\varepsilon,N},\theta)\\
&\qquad\quad\circ\Big\{\Lambda^{-1}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})P_k^{\varepsilon,i,N}(\theta_0)\Big\}\\
&\qquad-2\delta\sum_{k=1}^{n}\Big\{(\nabla_\theta^{(2)}b)^*(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\mu_{(k-1)\delta}^{\varepsilon,N},\mu_{(k-1)\delta-\tau}^{\varepsilon,N},\theta)\\
&\qquad\quad\circ\Big\{\Lambda^{-1}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})B(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta)\Big\}\\
&\qquad-(\nabla_\theta b)^*(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\mu_{(k-1)\delta}^{\varepsilon,N},\mu_{(k-1)\delta-\tau}^{\varepsilon,N},\theta)\\
&\qquad\quad\times\Lambda^{-1}(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N})(\nabla_\theta b)(X_{(k-1)\delta}^{\varepsilon,i,N},X_{(k-1)\delta-\tau}^{\varepsilon,i,N},\mu_{(k-1)\delta}^{\varepsilon,N},\mu_{(k-1)\delta-\tau}^{\varepsilon,N},\theta)\Big\}\\
&\quad=:\Pi_1+\Pi_2.
\end{aligned}
$$

For any $x,y\in\mathbb{R}^d$ and $\mu,\nu\in\mathscr{P}_p(\mathbb{R}^d)$, notice from Assumption 6.5 that

$$
\begin{aligned}
&\sup_{\theta\in\overline{\Theta}}\|(\nabla_\theta^{(2)}b^*)(x,y,\mu,\nu,\theta)\|\\
&\qquad\le C\{1+|x|+|y|+|y|^{r_5+1}+\mathbb{W}_2(\mu,\delta_0)+\mathbb{W}_2(\nu,\delta_0)\}.
\end{aligned}
\tag{6.76}
$$

Set $q_4=(r_3+1)\vee(r_5+1)$. For the first term Π_1, by (6.51) and (6.76), one arrives at

$$
\begin{aligned}
&\mathbb{E}|\Pi_1|\\
&\quad\le 2\varepsilon\Big(\mathbb{E}\int_0^T\|(\nabla_\theta^{(2)}b)^*(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\mu_{t_\delta}^{\varepsilon,N},\mu_{t_\delta-\tau}^{\varepsilon,N},\theta)\|^2\|\Lambda^{-1}(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N})\\
&\qquad\times\sigma(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\mu_{t_\delta}^{\varepsilon,N},\mu_{t_\delta-\tau}^{\varepsilon,N})\|^2\mathrm{d}t\Big)^{1/2}\\
&\quad\le C\varepsilon\Big(\mathbb{E}\int_0^T\Big(1+|X_{t_\delta}^{\varepsilon,i,N}|^2+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^2+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{2q_4}+\frac{1}{N}\sum_{j=1}^{N}|X_{t_\delta}^{\varepsilon,j,N}|^2
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \frac{1}{N}\sum_{j=1}^{N}|X_{t_\delta-\tau}^{\varepsilon,j,N}|^2\Big)^3\mathrm{d}t\Big)^{1/2}\\
&\leq C\varepsilon\Big(\int_0^T\Big(1+\mathbb{E}|X_{t_\delta}^{\varepsilon,i,N}|^6+\mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{6q_4}+\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}|X_{t_\delta}^{\varepsilon,j,N}|^6\\
&\quad + \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,j,N}|^6\Big)\mathrm{d}t\Big)^{1/2}\\
&\leq C\varepsilon\to 0,\quad \text{as}\ \ \varepsilon\to 0,\ \ \delta\to 0 \text{ and } N\to\infty.
\end{aligned}
$$

For the second term Π_2, we infer that

$$
\begin{aligned}
\Pi_2 = &-2\int_0^T(\nabla_\theta^{(2)}b)^*(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\mu_{t_\delta}^{\varepsilon,N},\mu_{t_\delta-\tau}^{\varepsilon,N},\theta)\circ(\Lambda^{-1}(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N})\\
&\quad\times B(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta))\mathrm{d}t\\
&+2\int_0^T(\nabla_\theta b)^*(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\mu_{t_\delta}^{\varepsilon,N},\mu_{t_\delta-\tau}^{\varepsilon,N},\theta)\Lambda^{-1}(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N})\\
&\quad\times(\nabla_\theta b)(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\mu_{t_\delta}^{\varepsilon,N},\mu_{t_\delta-\tau}^{\varepsilon,N},\theta)\mathrm{d}t\\
=:&\ H_1+H_2.
\end{aligned}
$$

Taking into consideration Lemma 6.11 and Assumption 6.5 yields that

$$
\begin{aligned}
&H_1-K(\theta)\\
&=-2\int_0^T\Big((\nabla_\theta^{(2)}b)^*(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\mu_{t_\delta}^{\varepsilon,N},\mu_{t_\delta-\tau}^{\varepsilon,N},\theta)\circ(\Lambda^{-1}(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N})\\
&\quad\times B(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta))-(\nabla_\theta^{(2)}b^*)(X_t^{0,i},X_{t-\tau}^{0,i},\mu_t^{0,i},\mu_{t-\tau}^{0,i},\theta)\\
&\quad\circ(\Lambda^{-1}(X_t^{0,i},X_{t-\tau}^{0,i})B(X_t^{0,i},X_{t-\tau}^{0,i},\theta_0,\theta))\Big)\mathrm{d}t\\
&=-2\int_0^T\Big(((\nabla_\theta^{(2)}b)^*(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\mu_{t_\delta}^{\varepsilon,N},\mu_{t_\delta-\tau}^{\varepsilon,N},\theta)\\
&\quad-(\nabla_\theta^{(2)}b^*)(X_t^{0,i},X_{t-\tau}^{0,i},\mu_t^{0,i},\mu_{t-\tau}^{0,i},\theta))\circ(\Lambda^{-1}(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N})B(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta))\\
&\quad+(\nabla_\theta^{(2)}b^*)(X_t^{0,i},X_{t-\tau}^{0,i},\mu_t^{0,i},\mu_{t-\tau}^{0,i},\theta)\circ(\Lambda^{-1}(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N})-\Lambda^{-1}(X_t^{0,i},X_{t-\tau}^{0,i}))\\
&\quad\times B(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta))+(\nabla_\theta^{(2)}b^*)(X_t^{0,i},X_{t-\tau}^{0,i},\mu_t^{0,i},\mu_{t-\tau}^{0,i},\theta)\\
&\quad\circ(\Lambda^{-1}(X_t^{0,i},X_{t-\tau}^{0,i})(B(X_{t_\delta}^{\varepsilon,i,N},X_{t_\delta-\tau}^{\varepsilon,i,N},\theta_0,\theta)-B(X_t^{0,i},X_{t-\tau}^{0,i},\theta_0,\theta))\Big)\mathrm{d}t\\
&=:\sum_{i=1}^{3}M_3.
\end{aligned}
$$

For the term M_1, thanks to Assumptions 6.5, (6.22) and (6.48), it follows from the

Hölder inequality that

$$
\begin{aligned}
&\mathbb{E}|M_1| \\
&\le C\int_0^T \Big(\mathbb{E}|X_{t_\delta}^{\varepsilon,i,N}-X_t^{0,i}|^2+\mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,i,N}-X_{t-\tau}^{0,i}|^2(1+|X_{t-\tau}^{0,i}|^{2r_5}+|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{2r_5}) \\
&\quad+\frac{1}{N}\sum_{j=1}^N \mathbb{E}|X_{t_\delta}^{\varepsilon,j,N}-X_t^{0,j}|^2+\frac{1}{N}\sum_{j=1}^N \mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,j,N}-X_{t-\tau}^{0,j}|^2\Big)^{1/2} \\
&\quad\times\Big(1+\mathbb{E}|X_{t_\delta}^{\varepsilon,i,N}|^4+\mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,i,N}|^4+\mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{4q_2}+\frac{1}{N}\sum_{j=1}^N \mathbb{E}|X_{t_\delta}^{\varepsilon,j,N}|^4 \\
&\quad+\frac{1}{N}\sum_{j=1}^N \mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,j,N}|^4\Big)^{1/2}\mathrm{d}t \\
&\le C\int_0^T \Big(\mathbb{E}|X_{t_\delta}^{\varepsilon,i,N}-X_t^{0,i}|^2+(\mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,i,N}-X_{t-\tau}^{0,i}|^4)^{\frac{1}{2}}(1+\mathbb{E}|X_{t-\tau}^{0,i}|^{4r_5} \\
&\quad+\mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,i,N}|^{4r_5})^{\frac{1}{2}}+\frac{1}{N}\sum_{j=1}^N \mathbb{E}|X_{t_\delta}^{\varepsilon,j,N}-X_t^{0,j}|^2+\frac{1}{N}\sum_{j=1}^N \mathbb{E}|X_{t_\delta-\tau}^{\varepsilon,j,N}-X_{t-\tau}^{0,j}|^2\Big)^{1/2}\mathrm{d}t,
\end{aligned}
$$

where, in the second step we have used the result in Lemma 6.6. Then, according to (6.57), one has

$$\mathbb{E}|M_1|\to 0 \quad \text{as } \ \varepsilon\to 0, \ \ \delta\to 0, \ \ N\to\infty.$$

By Assumptions 6.3, (6.48), (6.52) and (6.76) and carrying out similar arguments, one has

$$\mathbb{E}|M_2|\to 0 \quad \text{as } \ \varepsilon\to 0, \ \ \delta\to 0, \ \ N\to\infty$$

and

$$\mathbb{E}|M_3|\to 0 \quad \text{as } \ \varepsilon\to 0, \ \ \delta\to 0, \ \ N\to\infty.$$

As a result, we conclude that

$$H_1\to K(\theta) \ \ \mathbb{P}-\text{a.s.} \ \text{ as } \ \varepsilon\to 0, \ \ \delta\to 0, \ \ N\to\infty. \tag{6.77}$$

Again, carrying out analogous arguments to derive (6.77), we obtain

$$H_2\to 2I(\theta) \quad \mathbb{P}-\text{a.s.} \ \text{ as } \ \varepsilon\to 0, \ \ \delta\to 0, \ \ N\to\infty. \tag{6.78}$$

Therefore, the desired assertion is completed by (6.77) and (6.78) immediately. □

Now we start to finish the argument of Theorem 6.3 on the basis of the previous lemmas.

Proof of Theorem 6.3. According to the result of Theorem 6.2, there exists a sequence $\eta_{n,\varepsilon}^{i,N} \to 0$ as $N, n \to \infty$ and $\varepsilon \to 0$ such that $\hat{\theta}_{n,\varepsilon}^{i,N} \in B_{\eta_{n,\varepsilon}^{i,N}}(\theta_0) \subset \Theta$, $\mathbb{P}$-a.s., that is to say,

$$\mathbb{P}\Big(\hat{\theta}_{n,\varepsilon}^{i,N} \in B_{\eta_{n,\varepsilon}^{i,N}}(\theta_0)\Big) \to 1, \quad \text{as } n, N \to \infty, \quad \varepsilon \to 0. \tag{6.79}$$

Then, it is easy to see that

$$(\nabla_\theta \Phi_{n,\varepsilon}^{i,N})(\hat{\theta}_{n,\varepsilon}^{i,N}) = (\nabla_\theta \Phi_{n,\varepsilon}^{i,N})(\theta_0) + F_{n,\varepsilon}^{i,N}(\hat{\theta}_{n,\varepsilon}^{i,N} - \theta_0), \quad \hat{\theta}_{n,\varepsilon}^{i,N} \in B_{\eta_{n,\varepsilon}^{i,N}}(\theta_0) \tag{6.80}$$

with

$$F_{n,\varepsilon}^{i,N} := \int_0^1 (\nabla_\theta^{(2)} \Phi_{n,\varepsilon}^{i,N})(\theta_0 + v(\hat{\theta}_{n,\varepsilon}^{i,N} - \theta_0))\mathrm{d}v, \quad \hat{\theta}_{n,\varepsilon}^{i,N} \in B_{\eta_{n,\varepsilon}^{i,N}}(\theta_0),$$

owing to the Taylor expansion. In what follows we intend to deduce that

$$F_{n,\varepsilon}^{i,N} \to \overline{K}(\theta_0) \quad \mathbb{P}-\text{a.s.} \tag{6.81}$$

as $n, N \to \infty$ and $\varepsilon \to 0$. Indeed, for $\hat{\theta}_{n,\varepsilon}^{i,N} \in B_{\eta_{n,\varepsilon}^{i,N}}(\theta_0)$,

$$\begin{aligned}
&\|F_{n,\varepsilon}^{i,N} - \overline{K}(\theta_0)\| \\
&\le \|F_{n,\varepsilon}^{i,N} - (\nabla_\theta^{(2)} \Phi_{n,\varepsilon}^{i,N})(\theta_0)\| + \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon}^{i,N})(\theta_0) - \overline{K}(\theta_0)\| \\
&\le \int_0^1 \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon}^{i,N})(\theta_0 + v(\hat{\theta}_{n,\varepsilon}^{i,N} - \theta_0)) - (\nabla_\theta^{(2)} \Phi_{n,\varepsilon}^{i,N})(\theta_0)\|\mathrm{d}v \\
&\quad + \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon}^{i,N})(\theta_0) - \overline{K}(\theta_0)\| \\
&\le \sup_{\theta \in B_{\eta_{n,\varepsilon}}(\theta_0)} \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon}^{i,N})(\theta) - \overline{K}(\theta)\| + \sup_{\theta \in B_{\eta_{n,\varepsilon}}(\theta_0)} \|\overline{K}(\theta) - \overline{K}(\theta_0)\| \\
&\quad + 2\|(\nabla;_\theta^{(2)} \Phi_{n,\varepsilon}^{i,N})(\theta_0) - \overline{K}(\theta_0)\|,
\end{aligned}$$

where $\overline{K}(\cdot)$ is shown in (6.75). This, together with Lemma 6.14 and the continuity of $\overline{K}(\cdot)$, yields that (6.81) holds. Next we show the asymptotic distribution of $\hat{\theta}_{n,\varepsilon}^{i,N}$. Let

$$\mathcal{F}_{n,\varepsilon}^{i,N} = \{F_{n,\varepsilon}^{i,N} \text{ is invertible }, \hat{\theta}_{n,\varepsilon}^{i,N} \in B_{\eta_{n,\varepsilon}^{i,N}}(\theta_0)\}.$$

By Lemma 6.14, one gets, for some positive constant α,

$$\mathbb{P}\Big(\sup_{\theta \in B_{\eta_{n,\varepsilon}^{i,N}}(\theta_0)} \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon}^{i,N})(\theta) - \overline{K}(\theta_0)\| \le \frac{\alpha}{2}\Big) \to 1 \tag{6.82}$$

as $n, N \to \infty$ and $\varepsilon \to 0$. What's more, by following the line of [LSS13, Theorem 2.2], we can deduce that $F_{n,\varepsilon}^{i,N}$ is invertible on the set

$$\Gamma_{n,\varepsilon}^{i,N} := \Big\{\sup_{\theta \in B_{\eta_{n,\varepsilon}^{i,N}}(\theta_0)} \|(\nabla_\theta^{(2)} \Phi_{n,\varepsilon}^{i,N})(\theta) - \overline{K}(\theta_0)\| \le \frac{\alpha}{2}, \ \hat{\theta}_{n,\varepsilon}^{i,N} \in B_{\eta_{n,\varepsilon}^{i,N}}(\theta_0)\Big\}.$$

Clearly,

$$\begin{aligned}1 \geq \mathbb{P}(\Gamma^{i,N}_{n,\varepsilon}) \geq &\mathbb{P}\Big(\sup_{\theta\in B_{\eta_{n,\varepsilon}}(\theta_0)} \|(\nabla^{(2)}_\theta \Phi^{i,N}_{n,\varepsilon})(\theta) - K_0(\theta_0)\| \leq \frac{\alpha}{2}\Big)\\ &+ \mathbb{P}\Big(\hat{\theta}^{i,N}_{n,\varepsilon} \in B_{\eta_{n,\varepsilon}}(\theta_0)\Big) - 1. \end{aligned}\tag{6.83}$$

Thus, taking advantage of (6.82), (6.79) as well as (6.83), we deduce

$$\mathbb{P}(\mathcal{F}^{i,N}_{n,\varepsilon}) \geq \mathbb{P}(\Gamma^{i,N}_{n,\varepsilon}) \to 1 \quad \text{as} \quad n, N \to \infty, \quad \varepsilon \to 0. \tag{6.84}$$

Let

$$U^{i,N}_{n,\varepsilon} = F^{i,N}_{n,\varepsilon}\mathbf{1}_{\mathcal{F}^{i,N}_{n,\varepsilon}} + I_p \mathbf{1}_{(\mathcal{F}^{i,N}_{n,\varepsilon})^c} \ ,$$

where I_p is a $p \times p$ identity matrix. It follows from (6.80) that

$$\begin{aligned}&\varepsilon^{-1}(\hat{\theta}^{i,N}_{n,\varepsilon} - \theta_0)\\ &\quad= (\varepsilon^{-1}(\hat{\theta}^{i,N}_{n,\varepsilon} - \theta_0))\mathbf{1}_{\mathcal{F}^{i,N}_{n,\varepsilon}} + (\varepsilon^{-1}(\hat{\theta}^{i,N}_{n,\varepsilon} - \theta_0))\mathbf{1}_{(\mathcal{F}^{i,N}_{n,\varepsilon})^c}\\ &\quad= (U^{i,N}_{n,\varepsilon})^{-1} F^{i,N}_{n,\varepsilon}(\varepsilon^{-1}(\hat{\theta}^{i,N}_{n,\varepsilon} - \theta_0))\mathbf{1}_{\mathcal{F}^{i,N}_{n,\varepsilon}} + (\varepsilon^{-1}(\hat{\theta}^{i,N}_{n,\varepsilon} - \theta_0))\mathbf{1}_{(\mathcal{F}^{i,N}_{n,\varepsilon})^c}\\ &\quad= \varepsilon^{-1}(U^{i,N}_{n,\varepsilon})^{-1}\{(\nabla_\theta \Phi^{i,N}_{n,\varepsilon})(\hat{\theta}^{i,N}_{n,\varepsilon}) - (\nabla_\theta \Phi^{i,N}_{n,\varepsilon})(\theta_0)\}\mathbf{1}_{\mathcal{F}^{i,N}_{n,\varepsilon}}\\ &\qquad+ (\varepsilon^{-1}(\hat{\theta}^{i,N}_{n,\varepsilon} - \theta_0))\mathbf{1}_{(\mathcal{F}^{i,N}_{n,\varepsilon})^c}\\ &\quad= -\varepsilon^{-1}(U^{i,N}_{n,\varepsilon})^{-1}(\nabla_\theta \Phi^{i,N}_{n,\varepsilon})(\theta_0)\mathbf{1}_{\mathcal{F}^{i,N}_{n,\varepsilon}} + (\varepsilon^{-1}(\hat{\theta}^{i,N}_{n,\varepsilon} - \theta_0))\mathbf{1}_{(\mathcal{F}^{i,N}_{n,\varepsilon})^c}\\ &\quad\to I^{-1}(\theta_0)\int_0^T \Upsilon(X^{0,i}_t, X^{0,i}_{t-\tau}, \theta_0)\mathrm{dW}^i_t \qquad \text{as} \ \ n, N \to \infty, \quad \varepsilon \to 0,\end{aligned}$$

where, in the fourth step we have used the Fermat lemma and dropped the term $(\nabla_\theta \Phi^{i,N}_{n,\varepsilon})(\hat{\theta}^{i,N}_{n,\varepsilon})$, and in the last step we have utilized Lemma 6.6, (6.75), (6.81), and (6.84). The desired conclusion is obtained. □

Bibliography

[AHPP23] Chiara Amorino, Akram Heidari, Vytautė Pilipauskaitė, and Mark Podolskij. Parameter estimation of discretely observed interacting particle systems. *Stochastic Process. Appl.*, 163:350–386, 2023.

[BBAKP18] Mátyás Barczy, Mohamed Ben Alaya, Ahmed Kebaier, and Gyula Pap. Asymptotic properties of maximum likelihood estimator for the growth rate for a jump-type CIR process based on continuous time observations. *Stochastic Process. Appl.*, 128(4):1135–1164, 2018.

[BH22] Jianhai Bao and Xing Huang. Approximations of McKean-Vlasov stochastic differential equations with irregular coefficients. *J. Theoret. Probab.*, 35(2):1187–1215, 2022.

[BHY19] Jianhai Bao, Xing Huang, and Chenggui Yuan. Convergence Rate of Euler-Maruyama Scheme for SDEs with Hölder-Dini Continuous Drifts. *J. Theor. Probab.*, 32(2):848–871, 2019.

[BLM17] Rainer Buckdahn, Juan Li, and Jin Ma. A mean-field stochastic control problem with partial observations. *Ann. Appl. Probab.*, 27(5):3201–3245, 2017.

[BP16] Mátyás Barczy and Gyula Pap. Asymptotic properties of maximum-likelihood estimators for Heston models based on continuous time observations. *Statistics*, 50(2):389–417, 2016.

[BS18] Jianhai Bao and Jinghai Shao. Weak convergence of path-dependent sdes with irregular coefficients, 2018.

[BY13] Jianhai Bao and Chenggui Yuan. Convergence rate of EM scheme for SDDEs. *Proc. Amer. Math. Soc.*, 141(9):3231–3243, 2013.

[CD18] René Carmona and François Delarue. *Probabilistic theory of mean field games with applications. I*, volume 83 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 2018.

[CIlL92] Michael G. Crandall, Hitoshi Ishii, and Pierre louis Lions. Use's guide to viscosity solutions of second order partial differential equations. *Bulletin of the A.M.S.*, 27(1):1–67, 1992.

[CIR85] John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross. An Intertemporal General Equilibrium Model of Asset Prices. *Econometrica*, 53(2):363–384, 1985.

[CIR05] John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross. A Theory of the term structure of interest rates. *Econometrica*, 53(2):129–164, 2005.

[CK17] Francesca Collet and Richard C. Kraaij. Dynamical moderate deviations for the Curie–Weiss model. *Stochastic Process. Appl.*, 127(9):2900–2925, 2017.

[CL71] Michael G. Crandall and Thomas M. Liggett. Generation of Semi-groups of nonlinear transformations on general Banach spaces. *Amer. J. Math.*, 93(2):265–298, 1971.

[CL83] Michael G. Crandall and Pierre-Louis Lions. Viscosity solutions of hamilton-jacobi equations. *Trans. Amer. Math. Soc.*, 277(1):1–42, 1983.

[dA97] Alejandro D. de Acosta. Exponential tightness and projective systems in large deviation theory. In *Festschrift for Lucien Le Cam*, pages 143–156. Springer, New York, 1997.

[DCK24] Serena Della Corte and Richard C. Kraaij. Large deviations for Markov processes with switching and homogenisation via Hamilton-Jacobi-Bellman equations. *Stochastic Process. Appl.*, 170:104301, 2024.

[DE97] Paul Dupuis and Richard S. Ellis. *A weak convergence approach to the theory of large deviations*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1997.

[Dei92] Klaus Deimling. *Multivalued differential equations*. Walter de Gruyter & Co., Berlin, 1992.

[DFL11] Xiaoxue Deng, Jin Feng, and Yong Liu. A singular 1-D Hamilton-Jacobi equation, with application to large deviation of diffusions. *Commun. Math. Sci.*, 9(1):289–300, 2011.

[DGCL18] Maud Delattre, Valentine Genon-Catalot, and Catherine Larédo. Parametric inference for discrete observations of diffusion processes with mixed effects. *Stochastic Process. Appl.*, 128(6):1929–1957, 2018.

[dH08] Frank den Hollander. *Large Deviations*. American Mathematical Society, New York, 2008.

[dRST19] Gonçalo dos Reis, William Salkeld, and Julian Tugaut. Freidlin-Wentzell LDP in path space for McKean-Vlasov equations and the functional iterated logarithm law. *Ann. Appl. Probab.*, 29(3):1487–1540, 2019.

[DV75] Monroe D. Donsker and S. R. S. Varadhan. On a variational formula for the principal eigenvalue for operators with maximum principle. *Proc.Nat.Acad.Sci.USA*, 72(3):780–783, 1975.

[DZ98] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*, volume 38. Springer Berlin, Heidelberg, 1998.

[EGZ19] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Quantitative Harris-type theorems for diffusions and McKean-Vlasov processes. *Trans. Amer. Math. Soc.*, 371(10):7135–7173, 2019.

[EI85] L. C. Evans and H. Ishii. A PDE approach to some asymptotic problems concerning random differential equations with small noise intensities. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 2(1):1–20, 1985.

[Ein05] A. Einstein. Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Annalen der Physik*, 322(8):549–560, 1905.

[Elw82] K. David Elworthy. *Stochastic differential equations on manifolds*, volume 70 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge-New York, 1982.

[Eng89] Ryszard Engelking. *General topology*, volume 6 of *Sigma Series in Pure Mathematics*. Heldermann Verlag, Berlin, second edition, 1989.

[FFF10] Jin Feng, Martin Forde, and Jean-Pierre Fouque. Short-maturity asymptotics for a fast mean-reverting heston stochastic volatility model. *SIAM J. Finac. Math.*, 1(1):126–141, 2010.

[FFK12] Jin Feng, Jean-Pierre Fouque, and Rohini Kumar. Small-time asymptotics for fast mean-reverting stochastic volatility models. *Ann. Appl. Probab.*, 22(4):1541–1575, 2012.

[FG15] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738, 2015.

[FK06] Jin Feng and Thomas G. Kurtz. *Large Deviations for Stochastic Processes*. Springer Berlin, Heidelberg, 2006.

[Fle78] Wendell H. Fleming. Exit probabilities and optimal stochastic control. *Appl. Math. Optim*, 4(1):329–346, 1978.

[Fle85] Wendell H. Fleming. A stochastic control approach to some large deviations problems. In *Recent mathematical methods in dynamic programming (Rome, 1984)*, volume 1119, pages 52–66. Springer, Berlin, 1985.

[Fri75] Avner Friedman. *Stochastic differential equations and applications.* Vol 1. Academic Press, New York, 1975.

[FZ89] Daniéle Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20(4):547–557, 1989.

[GM01] Istvan Gyongy and Teresa Martinez. On Stochastic Differential Equations with Locally Unbounded Drift. *Czechoslovak Math. J.*, 51(4):763–783, 2001.

[HKX23] Yanyan Hu, Richard C. Kraaij, and Fubao Xi. Large deviations for Cox-Ingersoll-Ross processes with state-dependent fast switching, 2023.

[HKX24] Yanyan Hu, Richard C. Kraaij, and Fubao Xi. Large deviations for slow-fast processes on connected complete Riemannian manifolds. *Stochastic Process. Appl.*, 2024.

[HMS16] Gang Huang, Michel Mandjes, and Peter Spreij. Large deviations for Markov-modulated diffusion processes with rapid switching. *Stochastic Process. Appl.*, 126(6):1785–1818, 2016.

[HRW19] Xing Huang, Michael Röckner, and Feng-Yu Wang. Nonlinear Fokker-Planck equations for probability measures on path space and path-distribution dependent SDEs. *Discrete Contin. Dyn. Syst.*, 39(6):3017–3035, 2019.

[Hsu02] Elton P. Hsu. *Stochastic analysis on manifolds*, volume 38 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, 2002.

[Hua19] Xing Huang. Exponential convergence for functional SDEs with Hölder continuous drift. *Stoch. Anal. Appl.*, 37(6):977–990, 2019.

[Hua21] Xing Huang. Path-distribution dependent SDEs with singular coefficients. *Electron. J. Probab.*, 26:1–21, 2021.

[HXZ24] Yanyan Hu, Fubao Xi, and Min Zhu. Asymptotic properties for the parameter estimation instochastic (functional) differential equations with Hölder drift. *Stochastics*, 96(1):766–798, 2024.

[HY21a] Xing Huang and Fenfen Yang. Distribution-dependent SDEs with Hölder continuous drift and α-stable noise. *Numer Algor*, 86(2):813–831, 2021.

[HY21b] Xing Huang and Chenggui Yuan. Comparison theorem for distribution-dependent neutral SFDEs. *J. Evol. Equ.*, 21(1):653–670, 2021.

[HZ19] Xing Huang and Shao-Qin Zhang. Mild solutions and Harnack inequality for functional stochastic partial differential equations with Dini drift. *J. Theor. Probab.*, 32:303–329, 3 2019.

[Kas88] Raphael A. Kasonga. The consistency of a nonlinear least squares estimator from diffusion processes. *Stochastic Process. Appl.*, 30(2):263–275, 1988.

[Kes97] Mathieu Kessler. Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist.*, 24(2):211–229, 1997.

[Kli82] Wilhelm Klingenberg. *Riemannian geometry*, volume 1 of *de Gruyter Studies in Mathematics.* Walter de Gruyter & Co., Berlin-New York, 1982.

[KR05] Nicolai V. Krylov and Michael Röckner. Strong solutions of stochastic equations with singular time dependent drift. *Probab. Theory Relat. Fields*, 131(2):154–196, 2005.

[Kra16] Richard C. Kraaij. *Semigroup methods for large deviations of Markov processes.* PhD thesis, Delft University of Technology, 2016.

[Kra20] Richard C. Kraaij. The exponential resolvent of a Markov process and large deviations for Markov processes via Hamilton-Jacobi equations. *Electron. J. Probab.*, 25:1–39, 2020.

[Kra22] Richard C. Kraaij. A general convergence result for viscosity solutions of hamilton-jacobi equations and non-linear semigroups. *J. Funct. Anal.*, 282(1):1–55, 2022.

[KRV19] Richard C. Kraaij, Frank Redig, and Rik Versendaal. Classical large deviation theorems on complete riemannian manifolds. *Stochastic Processes and their Applications*, 129(11):4294–4334, 2019.

[KS91] Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus.* Second Edition, Springer, New York, 1991.

[KS10] Uwe Küchler and Michael Sørensen. A simple estimator for discrete-time samples from affine stochastic delay differential equations. *Stat. Inference Stoch. Process.*, 13(2):125–132, 2010.

[KS20] Richard C. Kraaij and Mikola C. Schlottke. A large deviation principle for markovian slow-fast systems, 2020.

[KS21] Richard C. Kraaij and Mikola C. Schlottke. Comparison principle for hamilton-jacobi-bellman equations via a bootstrapping procedure. *Nonlinear Differ. Equ. Appl.*, 28(22):1–45, 2021.

[Kut04] Yury A. Kutoyants. *Statistical inference for ergodic diffusion processes.* Springer Series in Statistics. Springer-Verlag London, Ltd., London, 2004.

[LB76] Alain Le Breton. On continuous and discrete sampling for parameter estimation in diffusion type processes. *Math. Programming Stud.*, 5(5):124–144, 1976.

[Lee97] John M. Lee. *Riemannian manifolds*, volume 176 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1997. An introduction to curvature.

[Lee03] John M. Lee. *Introduction to smooth manifolds*, volume 218 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 2003.

[LMS17] Hongwei Long, Chunhua Ma, and Yasutaka Shimizu. Least squares estimators for stochastic differential equations driven by small Lévy noises. *Stoch. Proc. Appli.*, 127(5):1475–1495, 2017.

[Lon09] Hongwei Long. Least squares estimator for discretely observed Ornstein-Uhlenbeck processes with small Lévy noises. *Statist. Probab. Lett.*, 79(19):2076–2085, 2009.

[Lon10] Hongwei Long. Parameter estimation for a class of stochastic differential equations driven by small stable noises from discrete observations. *Acta Math. Sci. Ser. B (Engl. Ed.)*, 30(3):645–663, 2010.

[LS01] Robert S. Liptser and Albert N. Shiryaev. *Statistics of random processes. II*, volume 6 of *Applications of Mathematics (New York).* Springer-Verlag, Berlin, expanded edition, 2001.

[LSS13] Hongwei Long, Yasutaka Shimizu, and Wei Sun. Least squares estimators for discretely observed stochastic processes driven by small Lévy noises. *J. Multivariate Anal.*, 116:422–439, 2013.

[Ma10] Chunhua Ma. A note on "Least squares estimator for discretely observed Ornstein-Uhlenbeck processes with small Lévy noises". *Statist. Probab. Lett.*, 80(19-20):1528–1531, 2010.

[McK66] Jr. McKean, Henry P. A class of Markov processes associated with nonlinear parabolic equations. *Proc. Nat. Acad. Sci. U.S.A.*, 56, 1966.

[McK67] Jr. McKean, Henry P. Propagation of chaos for a class of non-linear parabolic equations. pages 41–57. Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, Arlington, VA, 1967.

[MYZ99] William M. McEneaney, G. George Yin, and Qing Zhang, editors. *Stochastic analysis, control, optimization and applications.* Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 1999.

[NBK$^+$20] Neelima, Sani Biswas, Chaman Kumar, Gonçalo dos Reis, and Christoph Reisinger. Well-posedness and tamed Euler schemes for Mckean-Vlasov equations driven by Lévy noise, 2020.

[PR83] B. L. S. Prakasa Rao. Asymptotic theory for nonlinear least squares estimator for diffusion processes. *Math. Operationsforsch. Statist. Ser. Statist.*, 14(2):195–209, 1983.

[PS24] Mark A. Peletier and Mikola C. Schlottke. Large-deviation principles of switching Markov processes via Hamilton-Jacobi equations. *Electron. J. Probab.*, 29:Paper No. 88, 39, 2024.

[Puh94] A. Puhalskii. The method of stochastic exponentials for large deviations. *Stochastic Process. Appl.*, 54(1):45–70, 1994.

[PY19] Yurong Pan and Litan Yan. The least squares estimation for the α-stable Ornstein-Uhlenbeck process with constant drift. *Methodol. Comput. Appl. Probab.*, 21(4):1165–1182, 2019.

[Rei05] Markus Reiss. Adaptive estimation for affine stochastic delay differential equations. *Bernoulli*, 11(1):67–102, 2005.

[RW19a] Panpan Ren and Feng-Yu Wang. Bismut formula for Lions derivative of distribution dependent SDEs and applications. *J. Differential Equations*, 267(8):4745–4777, 2019.

[RW19b] Panpan Ren and Jiang-Lun Wu. Least Squares estimator for path-dependent McKean-Vlasov SDEs via disctete-time observations. *Acta Math. Sci. Ser. B Engl. Ed.*, 3:691–716, 2019.

[RW21] Panpan Ren and Jiang-Lun Wu. Least squares estimation for path-distribution dependent stochastic differential equations. *Appl. Math. Comput.*, 410:126457, 2021.

[RZ04] Michael Röckner and Tusheng Zhang. Sample path large deviations for diffusion processes on configuration spaces over a Riemannian manifold. *Publ. Res. Inst. Math. Sci.*, 40(2):385–427, 2004.

[RZ21] Michael Röckner and Xicheng Zhang. Well-posedness of distribution dependent SDEs with singular drifts. *Bernoulli*, 27(2):1131–1158, 2021.

[Sha18] Jinghai Shao. Invariant measures and Euler–Maruyama's approximations of state-dependent regime-switching diffusions. *SIAM J. Control Optim.*, 56:3215–3238, 2018.

[vdV98] Aad van der Vaart. *Asymptotic statistics.* Cambridge University Press, Cambridge, 1998.

[Ver21] Rik Versendaal. Large deviations for Brownian motion in evolving Riemannian manifolds. *Markov Process. Related Fields*, 27(3):381–412, 2021.

[Vil09] Cédric Villani. *Optimal transport, Old and new.* Springer-Verlag, Berlin, 2009.

[Wan14] Feng-Yu Wang. *Analysis for diffusion processes on Riemannian manifolds.* World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2014.

[Wan16] Feng-Yu Wang. Gradient estimates and applications for SDEs in Hilbert space with multiplicative noise and Dini continuous drift. *J. Differential Equations*, 260(3):2792–2829, 2016.

[Wan18] Feng-Yu Wang. Distribution dependent SDEs for Landau type equations. *Stochastic Process. Appl.*, 128(2):595–621, 2018.

[Whi02] Ward Whitt. *Stochastic-process limits.* Springer Series in Operations Research. Springer-Verlag, New York, 2002. An introduction to stochastic-process limits and their application to queues.

[WS16] Chao Wei and Huisheng Shu. Maximum likelihood estimation for the drift parameter in diffusion processes. *Stochastics*, 88(5):699–710, 2016.

[WWMX16] Jianghui Wen, Xiangjun Wang, Shuhua Mao, and Xinping Xiao. Maximum likelihood estimation of McKean-Vlasov stochastic differential equation and its application. *Appl. Math. Comput.*, 274:237–246, 2016.

[XZ17] Fubao Xi and Chao Zhu. On Feller and Strong Feller properties and exponential ergodicity of regime-switching Jump diffusion processes with countable regimes. *SIAM J. Control Optim.*, 55(3):1789–1818, 2017.

[Yos92] Nakahiro Yoshida. Estimation for diffusion processes from discrete observation. *J. Multivariate Anal.*, 41(2):220–242, 1992.

[ZH22] Min Zhu and Yanyan Hu. Least squares estimation for delay McKean–Vlasov stochastic differential equations and interacting particle systems. *Commun. Math. Sci.*, 20(1):265–296, 2022.

[Zha05] Xicheng Zhang. Strong solutions of SDEs with singular drift and Sobolev diffusion coefficients. *Stochastic Process. Appl.*, 115(11):1805–1818, 2005.

[Zha11] Xicheng Zhang. Stochastic homeomorphism flows of SDEs with singular drifts and Sobolev diffusion coefficients. *Electron. J. Probab.*, 16:1096–1116, 2011.

[Zha16] Xicheng Zhang. Stochastic differential equations with Sobolev diffusion and singular drift and applications. *Ann. of Appl. Probab.*, 26(5):2697–2732, 2016.

Summary

In this thesis, we study large deviations and parameter estimations for small noise diffusion processes. In Chapter 1, we start with the classical limit theorems to intuitively introduce large deviations and parameter estimations, which provide for further developments in the thesis.

The first part, consisting of Chapters 2 - 4, is on large deviations. In Chapter 2, we begin with the simple stochastic differential equation to explain the idea behind the proof of the nonlinear semigroup method, which is used to prove large deviations in Chapters 3 and 4. In the process, viscosity solutions and the Hamilton-Jacobi-Bellman equations are introduced.

Chapter 3 is concerned with the Cox–Ingersoll–Ross process

$$\begin{cases} \mathrm{d}X_n^\varepsilon(t) = \eta(\mu(\Lambda_n^\varepsilon(t)) - X_n^\varepsilon(t))\mathrm{d}t + n^{-\frac{1}{2}}\theta\sqrt{X_n^\varepsilon(t)}\mathrm{d}W(t), \\ (X_n^\varepsilon(0), \Lambda_n^\varepsilon(0)) = (x_0, k_0) \in E \times S, \end{cases}$$

where the fast process $\Lambda_n^\varepsilon(t)$ is a jumping process on finite sets S satisfying

$$\mathbb{P}(\Lambda_n^\varepsilon(t+\triangle) = j \mid \Lambda_n^\varepsilon(t) = i, X_n^\varepsilon(t) = x) = \begin{cases} \frac{1}{\varepsilon} q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j \neq i, \\ 1 + \frac{1}{\varepsilon} q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j = i, \end{cases}$$

for $\Delta > 0$, $i, j \in S$, $x \in E = (0, \infty)$, where $\varepsilon > 0$ is a small perturbation. Then, under suitable conditions the large deviation principle with speed n holds for the slow process $X_n^\varepsilon(t)$ on the Skorokhod space $\mathcal{D}_E(\mathbb{R}^+)$ with a good rate function I having action-integral representation,

$$I(\gamma) = \begin{cases} I_0(\gamma(0)) + \int_0^\infty \mathcal{L}\left(\gamma(s), \dot{\gamma}(s)\right) \mathrm{d}s, & \text{if } \gamma \in \mathcal{AC}(E), \\ \infty, & \text{otherwise} \end{cases}$$

with $\mathcal{L}(x, v) = \sup_{p \in \mathbb{R}}\{\langle p, v\rangle - \mathcal{H}(x, p)\}$ which is the Legendre dual of $\mathcal{H}$ given by

$$\mathcal{H}(x, \partial_x f(x)) = \sup_{\pi \in \mathcal{P}(S)} \left\{ \int B_{x, \partial_x f(x)}(z)\pi(\mathrm{d}z) - \mathcal{I}(x, \pi) \right\},$$

where

$$B_{x,\partial_x f(x)}(i) = \eta(\mu(i) - x)\partial_x f(x) + \frac{1}{2}\theta^2 x(\partial_x f(x))^2$$

is coming from the slow process $X_n(t)$ and Donsker–Varadhan function

$$\mathcal{I}(x, \pi) = -\inf_{g>0} \int \frac{R_x g(z)}{g(z)} \pi(\mathrm{d}z),$$

where R_x is the generator corresponding to the fast process $\Lambda_n^\varepsilon(t)$ defined by

$$R_x g(z) = \sum_{j \in S} q_{zj}(x) \left(g(j) - g(z)\right).$$

During the proof of the above result, we obtain operator convergence, exponential tightness, comparison principle, and action-integral representation. As the slow process is a Cox-Ingersoll-Ross process that is singular at the point 0, deriving a non-compact space leads to obstacles in proving the exponential tightness and the comparison principle. To solve it, we find a good containment function

$$\Upsilon(x) = -\log(x) + \log(1 + \frac{1}{2}x^2) - \log\sqrt{2},$$

which has the same function as the usual Lyapunov function used to get exponential tightness. Subsequently, we prove the comparison principle using Riemannian distance

$$d(x, y) = |\sqrt{x} - \sqrt{y}|, \quad x, y \in E \tag{85}$$

instead of the normal distance $d(x, y) = |x - y|$, x, $y \in E$.

In Chapter 4, we study general slow-fast systems on Riemannian manifolds. This leads to the analysis of a rather complex Hamiltonian

$$H_{f,\phi}(x, i) = b(x, i)\mathrm{d}f(x) + \frac{1}{2}|\mathrm{d}f(x)|^2 + \sum_{j \in S} q_{ij}(x)(e^{\phi(x,j)-\phi(x,i)} - 1),$$

where $f \in C^2(M)$ and $\phi \in C^2(M \times S)$. In this setting, the approximate Lyapunov function is $\Upsilon(x) = \frac{1}{2}\log(1 + f^2(x))$ for getting the exponential tightness and the comparison principle. For the proof of the exponential tightness, our condition is linear growth which is weaker than the conditions in the existing literature. For the proof of the comparison principle, the extra difficulties come from three reasons: first, compared with the CIR process, the target process is a more general Stochastic differential equation; second, the distance function d is not smooth on the Riemannian manifold; third, the drift coefficient is a locally one-sided Lipschitz condition. Finally, for the proof of action-integral representation, it suffices to prove the existence of a viscosity solution, in which we need to find a global curve. To get it, we introduce a new approach using a local analysis that allows them to transfer the problem to Euclidean space, before having to patch things together.

Next, we turn to the second part of the thesis. In this part, the small noise diffusion processes with unknown parameters in the drift coefficient are studied. We solve how to get consistency and asymptotic normality of the estimator by the least squares method in Chapters 4 and 5.

More precisely, in Chapter 5, we consider a multidimensional stochastic differential equation

$$\mathrm{d}X^\varepsilon(t) = b(X^\varepsilon(t),\theta)\mathrm{d}t + \varepsilon\sigma(X^\varepsilon(t))\mathrm{d}W(t), \tag{86}$$

where the drift b is bounded and Hölder continuous. When $\varepsilon \to 0$, we have

$$\mathrm{d}X^0(t) = b(X^0(t),\theta_0)\mathrm{d}t.$$

In this case, we first constructed the estimator $\hat{\theta}_{n,\varepsilon} = \arg\min_{\theta\in\Theta}\Psi_{n,\varepsilon}(\theta)$, where $\Psi_{n,\varepsilon}(\theta)$ is a contrast function utilizing the Euler-Maruyama (EM) scheme and the theory of least squares. The main results are obtained with high frequency $(n \to \infty)$ and small dispersion $(\varepsilon \to 0)$. The first main result is consistency, $\hat{\theta}_{n,\varepsilon} \to \theta_0$ in probability; the second main result is asymptotic normality property,

$$\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0) \to I^{-1}(\theta_0)\int_0^T \Upsilon(X^0(t),\theta_0)\mathrm{d}B(t) \quad \text{in probability,}$$

where for any $x \in \mathbb{R}^d$ and $\theta \in \overline{\Theta}$,

$$\Upsilon(x,\theta_0) := (\nabla_\theta b)^*(x,\theta_0)\hat{\sigma}(x)\sigma(x)$$

with

$$\hat{\sigma}(x) := (\sigma\sigma^*)^{-1}(x)$$

and

$$I(\theta) := \int_0^T (\nabla_\theta b)^*(X^0(t),\theta)\hat{\sigma}(X^0(t))(\nabla_\theta b)(X^0(t),\theta)\mathrm{d}t.$$

The hard part in proving these results is to use the Zvonkin transform to handle the Hölder drift. The idea of the Zvonkin transformation is to construct a one-to-one transformation that allows us to transition from a diffusion process with a non-zero drift coefficient to a process without drift. Then we further disturb the SDE in (86), and we obtain a stochastic functional differential equation (SFDE)

$$\mathrm{d}X^\varepsilon(t) = b(X^\varepsilon(t),\theta)\mathrm{d}t + Z(X_t^\varepsilon)\mathrm{d}t + \varepsilon\sigma(X^\varepsilon(t))\mathrm{d}W(t).$$

The perturbation function Z depends on the history state and describes a delayed feedback loop that has a weak impact on the unperturbed dynamics. We follow the framework of the least squares method to demonstrate both consistency and asymptotic normality, the difference point is that we adopt the truncated EM instead of the EM scheme to discrete the SFDE.

Finally, in Chapter 6, we further investigate parameter estimations for more complex systems, McKean-Vlasov SDEs with the point delay. The evolution of these

equations depends not only on the state of the microscopic particles but also on the distribution of the macroscopic particles. Compared with the conditions set in the existing parameter estimation literature, we obtain the asymptotic properties of the least square estimator under the weaker condition: the drift and diffusion coefficients both satisfied the superlinear growth instead of the Lipschitz condition. To get it, we approximate the McKean-Vlasov SDEs with point delay via weakly interacting particle systems. As a key step in the proof, the propagation of chaos and the convergence of the EM scheme associated with the consequent weakly interacting particle systems are obtained.

Samenvatting

In dit proefschrift bestuderen wij grote afwijkingen en parameterschattingen voor diffusie processen met een kleine noise factor. In hoofdstuk 1 beginnen we met de klassieke limietstelling en introduceren we op intuïtieve wijze grote afwijkingen en parameterschattingen, die de basis vormen voor de rest van het proefschrift.

Het eerste deel bestaat uit de hoofdstukken 2, 3 en 4, en gaat over grote afwijkingen. In hoofdstuk 2 beginnen we met eenvoudige stochastische differentiaalvergelijkingen om de ideeën uit te leggen die ten grondslag liggen aan de niet-lineaire semigroepbenadering die in gebruikt om grote afwijkingen te de hoofdstukken 3 en 4 worden bewijzen. Tegelijkertijd worden ook viscositeitsoplossing en de Hamilton-Jacob-Bellman-vergelijking beschreven.

Hoofdstuk 3 behandelt het Cox-Ingersoll-Ross proces

$$\begin{cases} \mathrm{d}X_n^\varepsilon(t) = \eta(\mu(\Lambda_n^\varepsilon(t)) - X_n^\varepsilon(t))\mathrm{d}t + n^{-\frac{1}{2}}\theta\sqrt{X_n^\varepsilon(t)}\mathrm{d}W(t), \\ (X_n^\varepsilon(0), \Lambda_n^\varepsilon(0)) = (x_0, k_0) \in E \times S, \end{cases}$$

waar het snelle proces $\Lambda_n^\varepsilon(t)$ het sprongproces is op de eindige verzameling S voldoet aan

$$\mathbb{P}(\Lambda_n^\varepsilon(t+\triangle) = j \mid \Lambda_n^\varepsilon(t) = i, X_n^\varepsilon(t) = x) = \begin{cases} \frac{1}{\varepsilon} q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j \neq i, \\ 1 + \frac{1}{\varepsilon} q_{ij}(x)\triangle + \circ(\triangle), & \text{if } j = i, \end{cases}$$

voor $\Delta > 0$, $i, j \in S$, $x \in E = (0, \infty)$, waar $\varepsilon > 0$ een kleine verstoring is. Onder de juiste omstandigheden geldt dan het principe van de grote afwijking van de snelheid n voor het langzame proces $X_n^\varepsilon(t)$ op Skorokhod-Ruimten $\mathcal{D}_E(\mathbb{R}^+)$ met een goede snelheidsfunctie I met een geïntegreerde weergave van actie

$$I(\gamma) = \begin{cases} I_0(\gamma(0)) + \int_0^\infty \mathcal{L}(\gamma(s), \dot{\gamma}(s))\,\mathrm{d}s, & \text{if } \gamma \in \mathcal{AC}(E), \\ \infty, & \text{anders} \end{cases}$$

met $\mathcal{L}(x, v) = \sup_{p \in \mathbb{R}}\{\langle p, v\rangle - \mathcal{H}(x, p)\}$ die de Legendre duale van $\mathcal{H}$ is, gegeven door

$$\mathcal{H}(x, \partial_x f(x)) = \sup_{\pi \in \mathcal{P}(S)} \left\{ \int B_{x, \partial_x f(x)}(z)\pi(\mathrm{d}z) - \mathcal{I}(x, \pi) \right\},$$

waar

$$B_{x,\partial_x f(x)}(i) = \eta(\mu(i) - x)\partial_x f(x) + \frac{1}{2}\theta^2 x(\partial_x f(x))^2$$

komt van het langzame proces $X_n(t)$ en de Donsker-Varadhan functie

$$\mathcal{I}(x,\pi) = -\inf_{g>0} \int_E \frac{R_x g(z)}{g(z)} \pi(\mathrm{d}z),$$

waar R_x de generator is die overeenkomt met het snelle proces $\Lambda_n^\varepsilon(t)$ gedefinieerd door

$$R_x g(z) = \sum_{j\in S} q_{zj}(x)\left(g(j) - g(z)\right).$$

Bij het aantonen van de bovenstaande resultaten krijgen we operator convergentie, exponentiele compactheid, een vergelijkingsprincipe en action-integral representatie. Aangezien langzame processen een Cox-Ingersoll-Ross proces is dat singulier is in 0, het afleiden van een niet-compacte ruimte leidt tot obstakels om de exponentiële tightness en het vergelijkings principe te bewijzen. Vervolgens bewijzen wij we het vergelijkingsprincipe met behulp van de Riemannse manifolds

$$d(x,y) = |\sqrt{x} - \sqrt{y}|, \quad x, y \in E \tag{87}$$

in plaats van de normale afstand $d(x,y) = |x-y|$, x, $y \in E$.

In Hoofdstuk 4 bestuderen we algemene langzaam-snelle systemen op Riemannse manifolds. Dit geeft dan aanleiding tot de analyse van een vrij complexe hamiltoniaanse functie

$$H_{f,\phi}(x,i) = b(x,i)\mathrm{d}f(x) + \frac{1}{2}|\mathrm{d}f(x)|^2 + \sum_{j\in S} q_{ij}(x)[e^{\phi(x,j)-\phi(x,i)} - 1],$$

waar $f \in C^2(M)$ en $\phi \in C^2(M \times S)$. In dit geval is de approximatieve Lyapunov-functie $\Upsilon(x) = \frac{1}{2}\log(1 + f^2(x))$, gebruikt om exonentïele tightness en het vergelijkingsprincipe te verkrijgen. Voor het bewijs van exonentïele tightness gebruiken we lineaire groei, een conditie lineaire groei voor, die zwakker is dan de condities in de bestaande literatuur. Voor het bewijs van het vergelijkingsprincipe doen zich drie extra moeilijkheden voor: ten eerste is het doelproces een meer algemene stochastische differentiaalvergelijking dan het CIR-proces; ten tweede is de afstandsfunctie d niet glad op de Riemann-varieteit; ten derde, de drift coefficient is lokaal eenzijdige Lipschitz voorwaarde. Tenslotte is het voor het bewijs van de action-integral representatie voldoende om het bestaan van een viscosity oplossing te bewijzen, waarin we een globale curve moeten vinden. Om een globale curve te vinden introduceren we een nieuwe methode, die gebruik maakt van lokale analyse, waardoor het probleem kan worden overgebracht naar de Euclidische ruimte, waarna alles samen kan worden gebracht.

Vervolgens gaan we naar het tweede deel van dit proefschrift. In dit deel worden small noise diffusie processen met onbekende parameters in de drift coefficient bestudeerd.

In de hoofdstukken 4 en 5 hebben we met behulp van de kleinste kwadraten methode onderzocht hoe we de consistency en asymptotische normaliteit van de schatter kunnen krijgen.

Preciezer, in hoofdstuk 5 bekijken we een multidimensionele stochastische differentiaalvergelijking

$$\mathrm{d}X^\varepsilon(t) = b(X^\varepsilon(t),\theta)\mathrm{d}t + \varepsilon\sigma(X^\varepsilon(t))\mathrm{d}W(t), \tag{88}$$

waar de drift b begrensd en Hölder continu is. Als $\varepsilon \to 0$ gaat, hebben we dat

$$\mathrm{d}X^0(t) = b(X^0(t),\theta_0)\mathrm{d}t.$$

In dit geval construeren we eerst de schatter $\hat{\theta}_{n,\varepsilon} = \arg\min_{\theta\in\Theta}\Psi_{n,\varepsilon}(\theta)$, waar $\Psi_{n,\varepsilon}(\theta)$ is een contrastfunctie die gebruik maakt van het Euler-Maruyama (EM) scheme en de theorie van de kleinste kwadraten. De belangrijkste resultaten zijn verkregen met hoge frequenties ($n \to \infty$) en kleine dispersie ($\varepsilon \to 0$). Het eerste hoofdresultaat resultaat is consistency, $\hat{\theta}_{n,\varepsilon} \to \theta_0$ in kans; het tweede hoofdresultaat resultaat is asymptotische normaliteit,

$$\varepsilon^{-1}(\hat{\theta}_{n,\varepsilon} - \theta_0) \to I^{-1}(\theta_0)\int_0^T \Upsilon(X^0(t),\theta_0)\mathrm{d}B(t) \quad \text{in probability,}$$

waar voor elke $x \in \mathbb{R}^d$ en $\theta \in \overline{\Theta}$,

$$\Upsilon(x,\theta_0) := (\nabla_\theta b)^*(x,\theta_0)\hat{\sigma}(x)\sigma(x)$$

met

$$\hat{\sigma}(x) := (\sigma\sigma^*)^{-1}(x)$$

en

$$I(\theta) := \int_0^T (\nabla_\theta b)^*(X^0(t),\theta)\hat{\sigma}(X^0(t))(\nabla_\theta b)(X^0(t),\theta)\mathrm{d}t.$$

Het moeilijke deel om deze resultaten te bewijzen is het gebruik van Zvonkin transformaties om Hölder drift te verwerken. Het idee van de zvonkin-transformatie is om een injectieve transformatie te construeren die ons in staat stelt om over te gaan van een diffusieproces met een niet-nuldrift coefficient naar een proces zonder drift. De SDE in (88) wordt dan verder verstoord om een Stochastische functionele differentiaalvergelijking (SFDE) te verkrijgen

$$\mathrm{d}X^\varepsilon(t) = b(X^\varepsilon(t),\theta)\mathrm{d}t + Z(X_t^\varepsilon)\mathrm{d}t + \varepsilon\sigma(X^\varepsilon(t))\mathrm{d}W(t).$$

De perturbatiefunctie Z is afhankelijk van historische toestanden en beschrijft een terugkoppelingslus met vertraging die een zwakke invloed heeft op de dynamica van de onverstoorde dynamica. We volgen het raamwerk van de kleinste kwadraten methode om consistentie en asymptotische normaliteit aan te tonen, met het verschil dat we afgekorte EM gebruiken in plaats van EM om de SFDE te discretiseren.

Tenslotte wordt in hoofdstuk 6 nader ingegaan op de parameterschatting van complexere systemen, namelijk Mckean-Vlasov SDEs met puntvertraging. De evolutie van deze vergelijkingen hangt niet alleen af van de toestand van de microscopische deeltjes, maar ook van de verdeling van de macroscopische deeltjes. In vergelijking met de voorwaarden die in de literatuur voor parameter schatting, krijgen we de asymptotische eigenschappen van de kleinste kwadraten schatter onder zwakkere schatter: zowel de drift coefficient en de diffusie coefficient voldoen aan de hyperlineaire groei, niet aan de Lipschitz-voorwaarde. Om het te verkrijgen benaderen we de McKean-Vlasov SDEs met puntvertraging via het weakly interacting particle systems. Als een cruciale stap in het bewijs, wordt de convergentie van het EM algorithme ven de verspreiding van chaos en het daaruit voortvloeiende weakly interacting particle systems verkregen.

Acknowledgements

I would like to express my sincere gratitude to everyone who helped me throughout my PhD journey.

First and foremost, I want to thank you, Richard and Fubao, for your incredible guidance. Richard, my deepest gratitude goes to you for introducing me to the fascinating realm of large deviations. Your passion for research has profoundly influenced me, and our weekly discussions were always the highlight of my week. I learned many invaluable lessons, especially the importance of combining intuitive analysis with rigorous proof, writing with the reader's perspective in mind, and fostering independence in research. Fubao, I am deeply grateful to you for taking care of both my academic and personal life. Your diligent and practical academic attitude has been a constant source of inspiration. You always identified my weaknesses in research and provided patient guidance, which has helped me grow immensely. Thank you as well for introducing me to Richard, and enabling me to further my studies.

Frank, I also want to express my heartfelt thanks to you. Your talk, "How to be a good PhD Student", left a lasting impression on me and inspired me to carefully shape my own PhD trajectory. I deeply appreciated the seminar on Riemannian manifolds you organized with Sonja, your feedback during my yearly progress meetings, and your contributions to the applied probability group, which fostered a supportive and stimulating environment.

Jianhai, I am so grateful to you for your ongoing support even after I completed my master's degree. The lessons you taught me significantly enriched my understanding of key topics and were invaluable for my research. Min, I also want to thank you for your collaboration. Discussing problems with you has been a source of inspiration, and the research and learning experiences you shared have benefited me immensely.

Rik, thank you for the insightful discussions about Riemannian manifolds, which greatly advanced my work. Cor, I am also deeply grateful to you for teaching me how to navigate MathSciNet and encouraging me to pursue my dreams.

To my other doctoral committee members, Antonis Papapantoleon, Frank van der Meulen, Geurt Jongbloed, and Michel Mandjes, I would like to thank you for your time and effort in reviewing my thesis and for being part of this significant step in my academic journey.

To my fellow PhD colleagues: Ardjen, Berend, Chenxi, Chenguang, Dan, Francesco, Hidde, Huijie, Jan-Tino, Jasper, Jinzhi, Jonathan, Jun, Marc, Roy, Serena, Shixiang, Siddig, Stefan, Vicente, Wenda, Xiaocui, Yafei, Yago, Yidan, Yue, Yuhang, and Zuozheng, I am so grateful to you for the memorable PhD events we shared. Talking with you about our challenges provided a sense of relief and reminded me that I was not alone. I will always cherish the enjoyable conversations, coffee breaks, and lunches we had together.

A big thanks to you, Delft Institute of Applied Mathematics, Delft University of Technology, and School of Mathematics and Statistics, Beijing Institute of Technology, for providing such excellent research environments. Thank you also to the China Scholarship Council for financially supporting my life in the Netherlands.

Last but not least, I want to express my deepest gratitude to you, my beloved family, for being my warmest harbor. Your unwavering support and concern, always given without complaint, have been a source of strength throughout this journey.

Curriculum Vitae

Yanyan Hu was born on January 22nd, 1993, in Shandong, China. She earned her Bachelor's Degree in Mathematics from Shandong University of Technology in 2017. Subsequently, she pursued further studies in Mathematics at Central South University under the guidance of Prof. Dr. J. Bao, completing her Master's degree in June 2020. In September 2020, she commenced her PhD research under the supervision of Prof. Dr. F. Xi at Beijing Institute of Technology. Later, in September 2022, she successfully obtained research funding from the China Scholarship Council for a joint PhD project with Delft University of Technology, supervised by Dr. R.C. Kraaij. This dissertation presents the results of her PhD research.

Publications

Published

[ZH22] Min Zhu and Yanyan Hu. Least squares estimation for delay McKean–Vlasov stochastic differential equations and interacting particle systems. *Commun. Math. Sci.*, 20(1):265–296, 2022

[HXZ24] Yanyan Hu, Fubao Xi, and Min Zhu. Asymptotic properties for the parameter estimation instochastic (functional) differential equations with Hölder drift. *Stochastics*, 96(1):766–798, 2024

[HKX24] Yanyan Hu, Richard C. Kraaij, and Fubao Xi. Large deviations for slow-fast processes on connected complete Riemannian manifolds. *Stochastic Process. Appl.*, 2024

Submitted

[HKX23] Yanyan Hu, Richard C. Kraaij, and Fubao Xi. Large deviations for Cox-Ingersoll-Ross processes with state-dependent fast switching, 2023