PhiBE: A PDE-based Bellman Equation for Continuous Time Policy Evaluation

Yuhua Zhu*

Abstract

In this paper, we address the problem of continuous-time reinforcement learning in scenarios where the dynamics follow a stochastic differential equation. When the underlying dynamics remain unknown and we have access only to discrete-time information, how can we effectively conduct policy evaluation? We first highlight that the commonly used Bellman equation (BE) is not always a reliable approximation to the true value function. We then introduce a new bellman equation, PhiBE, which integrates the discrete-time information into a PDE formulation. The new bellman equation offers a more accurate approximation to the true value function, especially in scenarios where the underlying dynamics change slowly. Moreover, we extend PhiBE to higher orders, providing increasingly accurate approximations. We conduct the error analysis for both BE and PhiBE with explicit dependence on the discounted coefficient, the reward and the dynamics. Additionally, we present a model-free algorithm to solve PhiBE when only discrete-time trajectory data is available. Numerical experiments are provided to validate the theoretical guarantees we propose.

1 Introduction

Reinforcement learning (RL) [21] has achieved significant success in applications in discrete-time decision-making process. Remarkable milestones include its applications in Atari Games [14], AlphaGO [19], and ChatGPT [27, 16], demonstrating capabilities similar to human intelligence. In all these applications, there is no concept of time, where state transitions occur only after actions are taken. However, in most applications in the physical world, such as autonomous driving [3, 13] and robotics [11], state changes continuously over time regardless of whether actions are discrete or not. On the other hand, the data are always collected in discrete time. The mismatch between the continuous-time dynamics and discrete-time data makes continuous-time RL more challenging. This paper directs its focus toward addressing continuous-time RL problems that can be equivalently viewed as a stochastic optimal control problem with unknown dynamics [24, 6, 1, 9, **?**]. Since one can divide the RL problem into policy evaluation

*Department of Mathematics and Halicioğlu Data Science Institute, University of California, San Diego, La Jolla, California, U.S.A; e-mail: yuz244@ucsd.edu

and policy update [22, 12, 23, 9], we first focus on the continuous-time policy evaluation (PE) problem in this paper.

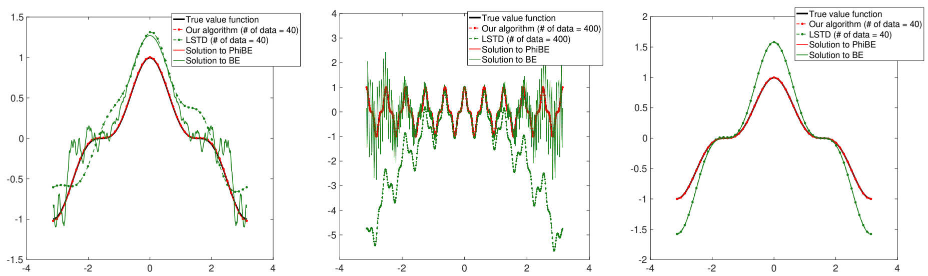
Given discrete-time trajectory data generated from the underlying dynamics, a common approach to address the continuous-time PE problem involves discretizing time and treating it as a Markov reward process. This method yields an approximated value function satisfying a Bellman equation, thereby one can use RL algorithms such as Temporal difference[21], gradient TD[18], Least square TD [4] to solve the Bellman equation. However, this paper shows that the Bellman equation is not always a good tool for approximating the continuous-time value function. We show that the solution to the Bellman equation is sensitive to time discretization, the change rate of the rewards and the discount coefficient as shown in Figure 1 (See Section 5.1 for the details of Figure 1.) Hence, the ineffectiveness of RL algorithms for continuous-time RL doesn't stem from data stochasticity or insufficient sampling points; rather, it fundamentally arises from the failure of the Bellman equation as an approximation of the true value function. As shown in Figure 1, the RL algorithms are approximating the solution to the Bellman equation instead of the true value function.

The central question we aim to address in this paper is whether, with the same discrete-time information and the same computational cost, one can approximate the true solution more accurately than the Bellman equation.

We proposed a PDE-based Bellman equation, called PhiBE. which integrates discrete-time information into a continuous PDE. This approach yields a more accurate approximation of the exact solution compared to the traditional Bellman equation, particularly when the acceleration of the dynamics is small. When equipped with discrete-time transition distribution, PhiBE is a second-order PDE that contains discrete-time information. The core concept revolves around utilizing discrete-time data to approximate the dynamics rather than the value function. Furthermore, we extend this framework to higher-order PhiBE, which enhances the approximation of the true value solution with respect to the time discretization. There are two potential benefits that arise from PhiBE. Firstly, PhiBE is more robust to various reward functions compared to BE, which allows greater flexibility in designing reward functions to effiectively achieve RL objectives. Secondly, it achieves comparable error to BE with sparser data collection, enhancing the efficiency of RL algorithms. As illustrated in Figure 1, when provided with the same discrete-time information, the exact solution derived from PhiBE is closer to the true value function than BE. Additionally, we introduce a model-free algorithm for approximating the solution to PhiBE when only discrete-time data is accessible. As depicted in Figure 1, with exactly the same data and the same computational cost, the proposed algorithm outperforms the RL algorithms drastically.

Contributions

- We demonstrate that the Bellman equation is a first-order approximation in terms of time discretization and provides the error dependence on the discount coefficient, reward function, and dynamics.



(a) $\Delta t = 5, \beta = 0.1, V(s) = \cos^3(s)$.

(b) $\Delta t = 0.5, \beta = 0.1$, $V(s) = \cos^3(10s)$.

(c) $\Delta t = 0.1, \beta = 10, V(s) = \cos^3(s)$.

Figure 1: Here the data are collected every Δt unite of time, β is the discount coefficient, and $V(s)$ is the true value function. In our setting, a larger discount coefficient indicates that future rewards are discounted more. LSTD [4] is a popular RL algorithm for linear function approximation. The PhiBE is proposed in Section 3 and the algorithm is proposed in Section 4.

- We propose a PDE-based Bellman equation that combines discrete-time information with PDE formulation. Furthermore, we extend it to a higher-order approximation. Error analysis is conducted for both deterministic and stochastic cases, and the error dependence on the discount coefficient, reward function, and dynamics are explicitly derived.
- We propose a model-free algorithm for solving PhiBE when only discrete-time data is available, whose computational cost is the same as LSTD with linear function approximation. Additionally, we provide the error analysis of the proposed algorithm when the discrete-time transition dynamics are given.

Related Work There are primarily two approaches to address continuous-time RL from the stochastic optimal control perspective. One involves employing machine learning techniques to learn the dynamics from discrete-time data and subsequently transforming the problem into a classical optimal control problem with known dynamics [10, 5]. However, directly identifying the continuous dynamics is often challenging when only discrete-time data is available. Another approach involves an algorithm that converges to the true value function using continuous-time information and then discretizes it when only discrete-time data is available. For instance, [1] presents a policy gradient algorithm tailored for linear dynamics and quadratic rewards. [8] introduces a martingale loss function for continuous-time PE. These algorithms converge to the true value function when continuous-time data is available. However, when only discrete-time data is accessible, numerical summation in discrete-time is employed to approximate the continuous integral (see, for example, Algorithm 2 in [1] and Equation (19) in [8]), which is similar to the Bellman equation.

The proposed method differs fundamentally in two ways: First, unlike model-based RL approaches, which end up solving a PDE with only continuous-time information, we integrate discrete-time information into the PDE formulation. Second, unlike alternative methodologies that directly approximate the value function using discrete-time values, which could neglect the smoothness of the value function, our method results in a PDE that incorporates gradients of the value function, which ensures that the solution closely approximates the true value function under smooth dynamics.

Organization The setting of the problem is specified in Section 2. Section 3 introduces the PDE-based bellman equation, PhiBE, and establishes theoretical guarantees. In Section 4, a model-free algorithm for solving the PhiBE is proposed. Numerical experiments are conducted in Section 5.

Notation For function $f(s) \in \mathbb{R}$, $\|f\|_{L^\infty} = \sup_s f(s)$; $\|f\|_\rho = \sqrt{\int f(s)^2 \rho(s) ds}$. For vector function $f(s) \in \mathbb{R}^d$, $\nabla f(s) \in \mathbb{R}^{d\times d}$ with $(\nabla f(s))_{ij} = \partial_{s_i} f_j$; $\|f\|_2^2 = \sum_{i=1}^d f_i^2$ represents the Euclidean Norm; $\|f\|_{L^\infty} = \sqrt{\sum_{i=1}^d \|f_i(s)\|_{L^\infty}^2}$, similar with $\|f\|_\rho$. For matrix function $f(s) \in \mathbb{R}^{d\times d}$, $\nabla \cdot f = \sum_{l=1}^d \partial_{s_l} f_{i,l}$.

2 Setting

Consider the following continuous-time PE problem, where the value function $V(s) \in \mathbb{R}$, defined as follows, is the expected discounted cumulative reward starting from s,

$$V(s) = \mathbb{E}\left[\int_0^\infty e^{-\beta t} r(s_t) dt | s_0 = s\right]. \tag{1}$$

Here $\beta > 0$ is a discounted coefficient, $r(s) \in \mathbb{R}$ is a reward function, and the state $s_t \in \mathbb{S} = \mathbb{R}^d$ is driven by the stochastic differential equation (SDE),

$$ds_t = \mu(s_t)dt + \sigma(s_t)dB_t, \tag{2}$$

with unknown drift function $\mu(s) \in \mathbb{R}^d$ and unknown diffusion function $\sigma(s) \in \mathbb{R}^{d\times d}$. In this paper, we assume that $\mu(s), \sigma(s)$ are Lipschitz continuous and the reward function $\|r\|_{L^\infty}$ is bounded. This ensures that (2) has a unique strong solution [15] and the infinite horizon integral is bounded.

We aim to approximate the continuous-time value function $V(s)$ when only discrete-time information is available. To be more specific, we consider the following two cases:

case 1. The transition distribution $\rho(s', \Delta t|s)$ in discrete time Δt, driven by the continuous dynamics (2), is given. Here $\rho(s', \Delta t|s)$ represents the probability density function of $s_{\Delta t}$ given $s_0 = s$.

case 2. Trajectory data generated by the continuous dynamics (2) and collected at discrete time $j\Delta t$ is given. Here the trajectory data $s = \{s_0^l, s_{\Delta t}^l, \cdots, s_{m\Delta t}^l\}_{l=1}^I$ contains I independent trajectories, and the initial state s_0^l of each trajectory are sampled from a distribution $\rho_0(s)$.

When the discrete transition distribution is given (Case 1), one can explicitly formulate the Bellman equation. One can also estimate the discrete transition distribution from the trajectory data, which is known as model-based RL. The error analysis in Section 3 is conducted under Case 1. We demonstrate that the Bellman equation is not always the optimal equation to solve continuous-time reinforcement learning problems even when the discrete-time transition dynamics are known, and consequently, all the RL algorithms derived from it are not optimal either. To address this, we introduce a Physics-informed Bellman equation (PhiBE) and establish that its exact solution serves as a superior approximation to the true value function compared to the classical Bellman equation. When only trajectory data is available (Case 2), one can also use the data to solve the PhiBE, referred to as model-free RL, which will be discussed in Section 4.

3 A PDE-based Bellman Equation (PhiBE)

In Section 3.1, we first introduce the Bellman equation, followed by an error analysis to demonstrate why it is not always a good approximation. Then, in Section 3.2, we propose the PhiBE, a PDE-based Bellman equation, considering both deterministic case (Section 3.2.1) and stochastic case (Section 3.2.2). The error analysis provides guidance on when PhiBE is a better approximation than the BE.

3.1 Bellman equation

By approximating the definition of the value function (1) in discrete time, one obtains the approximated value function,

$$\tilde{V}(s) = \mathbb{E}\left[\sum_{j=0}^{\infty} e^{-\beta\Delta t j} r(s_{j\Delta t})\Delta t | s_0 = s\right].$$

In this way, it can be viewed as a policy evaluation problem in Markov Decision Process, where the state is $s \in \mathbb{S}$, the reward is $r(s)\Delta t$, and the discount factor is $e^{-\beta\Delta t}$ and the transition dynamics is $\rho(s', \Delta t|s)$. Therefore, the approximated value function $\tilde{V}(s)$ satisfies the following Bellman equation [21].

Definition 1 (Definition of BE)**.**

$$\tilde{V}(s) = r(s)\Delta t + e^{-\beta\Delta t}\mathbb{E}_{s_{\Delta t}\sim\rho(s',\Delta t|s)}[\tilde{V}(s_{\Delta t})|s_0 = s]. \tag{3}$$

When the discrete-time transition distribution is not given, one can utilize various RL algorithms to solve the Bellman equation (3) using the trajectory data. However, if the exact solution to the Bellman equation is not a good approximation to the true value function, then all the RL algorithms derived from it will not effectively approximate the true value function. In the theorem below, we provide an upper bound for the distance between the solution $\tilde{V}$ to the above BE and the true value function V defined in (1).

Theorem 3.1. *Assume that $\|r\|_{L^\infty}, \|\mathcal{L}_{\mu,\Sigma} r\|_{L^\infty}$ are bounded, then the solution $\tilde{V}(s)$ to the BE (3) approximates the true value function $V(s)$ defined in (1) with an error*

$$\left\|V(s) - \tilde{V}(s)\right\|_{L^\infty} \le \frac{\frac{1}{2}(\|\mathcal{L}_{\mu,\Sigma} r\|_{L^\infty} + \beta \|r\|_{L^\infty})}{\beta} \Delta t + o(\Delta t),$$

where

$$\mathcal{L}_{\mu,\Sigma} = \mu(s) \cdot \nabla + \Sigma : \nabla^2, \tag{4}$$

with $\Sigma = \sigma\sigma^\top$, and $\Sigma : \nabla^2 = \sum_{i,j} \Sigma_{ij} \partial_{s_i} \partial_{s_j}$.

Remark 1 (Assumptions on $\|\mathcal{L}_{\mu,\Sigma} r\|_{L^\infty}$)**.** *One sufficient condition for the assumption to hold is that $\|\mu(s)\|_{L^\infty}$, $\|\Sigma(s)\|_{L^\infty}$, $\left\|\nabla^k r(s)\right\|_{L^\infty}$ for $k = 0, 1, 2$ are all bounded. However, $\|\mathcal{L}_{\mu,\Sigma} r\|_{L^\infty}$ is less restrictive than the above and allows, for example, linear dynamics $\mu(s) = \lambda s, \Sigma = 0$, with the derivative of the reward decreasing faster than a linear function at infinity, $\|s \cdot \nabla r(s)\|_{L^\infty} \le C$.*

The proof of the theorem is given in Section 6.1. In fact, by expressing the true value function $V(s)$ as the sum of two integrals, one can more clearly tell where the error in the BE comes from. Note that $V(s)$, as defined in (1), can be equivalently written as,

$$\begin{aligned} V(s) =&\mathbb{E}\left[\int_0^{\Delta t} e^{-\beta t} r(s) dt + \int_{\Delta t}^{\infty} e^{-\beta t} r(s_t) dt | s_0 = s\right] \\ =&\mathbb{E}\left[\int_0^{\Delta t} e^{-\beta t} r(s) dt | s_0 = s\right] + e^{-\beta \Delta t}\mathbb{E}\left[V(s_{t+\Delta t}) | s_0 = s\right] \end{aligned} \tag{5}$$

One can interpret the Bellman equation defined in (3) as an equation resulting from approximating $\mathbb{E}\left[\frac{1}{\Delta t}\int_0^{\Delta t} e^{-\beta t} r(s_t) dt | s_0 = s\right]$ in (5) by $r(s)$. The error between these two terms can be bounded by:

$$\left|\mathbb{E}\left[\left(\frac{1}{\Delta t}\int_0^{\Delta t} e^{-\beta t} r(s_t) dt\right) - r(s_0)\middle| s_0 = s\right]\right| \le \frac{1}{2}\left(\beta \|r\|_{L^\infty} + \|\mathcal{L}_{\mu,\Sigma} r\|_{L^\infty}\right) \Delta t + o(\Delta t), \tag{6}$$

characterizes the error of $\left\|V - \tilde{V}\right\|_{L^\infty}$ in Theorem 3.1.

Theorem 3.1 indicates that the solution $\tilde{V}$ to the Bellman equation (3) approximates the true value function with a first-order error of $O(\Delta t)$. Moreover, the coefficient before Δt suggests that for the same time discretization Δt, when β is small, the error is dominated by the term $\|\mathcal{L}_{\mu,\Sigma} r(s)\|_{L^\infty}$, indicating that the error increases when the reward changes rapidly. Conversely, when β is large, the error is mainly affected by $\|r\|_{L^\infty}$, implying that the error increases when the magnitude of the reward is large.

The question that the rest of this section seeks to address is whether, given the same discrete-time information, i.e., the transition distribution $\rho(s', \Delta t|s)$, time discretization Δt, and discount coefficient β, one can achieve a more accurate estimation of the value function V compared to the Bellman equation $\tilde{V}$.

3.2 A PDE-based Bellman equation

In this section, we introduce a PDE-based Bellman equation, referred to as PhiBE. We begin by discussing the case of deterministic dynamics in Section 3.2.1 to illustrate the idea clearly. Subsequently, we extend our discussion to the stochastic case in Section 3.2.2.

3.2.1 Deterministic Dynamics

When $\sigma(s) \equiv 0$ in (2), the dynamics become deterministic, which can be described by the following ODE,

$$\frac{ds_t}{dt} = \mu(s_t). \tag{7}$$

If the discrete-time transition dynamics $p(s', \Delta t|s) = p_{\Delta t}(s)$ is given, where $p_{\Delta t}(s)$ provides the state at time $t + \Delta t$ when the state at time t is s, then the BE in deterministic dynamics reads as follows,

$$\frac{1}{\Delta t}\tilde{V}(s) = r(s) + \frac{e^{-\beta \Delta t}}{\Delta t}\tilde{V}(p_{\Delta t}(s)).$$

The key idea of the new equation is that, instead of approximating the value function directly, one approximates the dynamics. First note that the value function defined in (1) can be equivalently written as,

$$V(s_t) = \int_t^\infty e^{-\beta(\tilde{t}-t)} r(s_{\tilde{t}}) d\tilde{t},$$

which implies that,

$$\frac{d}{dt}V(s_t) = \beta \int_t^\infty e^{-\beta(\tilde{t}-t)} r(s_{\tilde{t}}) d\tilde{t} - r(s_t).$$

Using the chain rule on the LHS of the above equation yields $\frac{d}{dt}V(s_t) = \frac{d}{dt}s_t \cdot \nabla V(s_t)$, and the RHS can be written as $\beta V(s_t) - r(s_t)$, resulting in a PDE for the true value function

$$\beta V(s_t) = r(s_t) + \frac{d}{dt}s_t \cdot \nabla V(s_t). \tag{8}$$

or equivalently,

$$\beta V(s) = r(s) + \mu(s) \cdot \nabla V(s). \tag{9}$$

Applying a finite difference scheme, one can approximate $\mu(s_t)$ by

$$\mu(s_t) = \frac{d}{dt} s_t \approx \frac{1}{\Delta t}(s_{t+\Delta t} - s_t),$$

and substituting it back into (8) yields

$$\beta \hat{V}(s_t) = r(s_t) + \frac{1}{\Delta t}(s_{t+\Delta t} - s_t) \cdot \nabla \hat{V}(s_t).$$

Alternatively, this equation can be expressed in the form of a PDE as follows,

$$\beta \hat{V}(s) = r(s) + \frac{1}{\Delta t}(p_{\Delta t}(s) - s) \cdot \nabla \hat{V}(s), \tag{10}$$

Note that the error now arises from

$$\left| \mu(s_t) - \frac{s_{t+\Delta t} - s_t}{\Delta t} \right| \leq \frac{1}{2} \left\| \mu \cdot \nabla \mu \right\|_{L^\infty} \Delta t,$$

which only depends on the dynamics. As long as the dynamics change slowly, i.e., the acceleration of dynamics $\left\| \frac{d^2}{dt^2} s_t \right\|_{L^\infty} = \left\| \mu \cdot \nabla \mu \right\|_{L^\infty}$ is small, the error diminishes.

We refer to (10) as PhiBE for deterministic dynamics, an abbreviation for the physics-informed Bellman equation, because it incorporates both the current state and the state after Δt, similar to the Bellman equation, while also resembling the form of the PDE (9) derived from the true continuous-time physical environment. However, unlike the true PDE (9) and the Bellman equation, where one only possesses continuous information and the other only discrete information, PhiBE combines both continuous PDE form and discrete transition information $p_{\Delta t}(s)$.

One can derive a higher-order PhiBE by employing a higher-order finite difference scheme to approximate $\mu(s_t)$. For instance, the second-order finite difference scheme

$$\mu(s_t) \approx \hat{\mu}_2(s_t) := \frac{1}{\Delta t} \left[-\frac{1}{2}(s_{t+2\Delta t} - s_t) + 2(s_{t+\Delta t} - s_t) \right],$$

resulting in the second-order PhiBE,

$$\beta \hat{V}_2(s) = r(s) + \frac{1}{\Delta t} \left[-\frac{1}{2}(p_{\Delta t}(p_{\Delta t}(s)) - s) + 2(p_{\Delta t}(s) - s) \right] \cdot \nabla \hat{V}_2(s).$$

In this approximation, $\left\| \mu(s) - \hat{\mu}_2(s) \right\|_{L^\infty}$ has a second order error $O(\Delta t^2)$. We summarize i-th order PhiBE in deterministic dynamics in the following Definition.

Definition 2 (i-th order PhiBE in deterministic dynamics). *When the underlying dynamics are deterministic, then the i-th order PhiBE is defined as,*

$$\beta \hat{V}_i(s) = r(s) + \hat{\mu}_i(s) \cdot \nabla \hat{V}_i(s), \tag{11}$$

where

$$\hat{\mu}_i(s) = \frac{1}{\Delta t}\sum_{j=1}^{i} a_j^{(i)} \left(\underbrace{p_{\Delta t} \circ \cdots \circ p_{\Delta t}}_{j}(s) - s \right), \tag{12}$$

and

$$(a_0^{(i)}, \cdots, a_i^{(i)})^\top = A_i^{-1} b_i, \quad \text{with } (A_i)_{kj} = j^k,\ (b_i)_k = \begin{cases} 0, & k \neq 1 \\ 1, & k = 1 \end{cases} \text{ for } 0 \leq j, k \leq i. \tag{13}$$

Remark 2. *Note that $\hat{\mu}_i(s)$ can be equivalently written as*

$$\hat{\mu}_i(s) = \frac{1}{\Delta t}\sum_{j=1}^{i} a_j^{(i)} [s_{j\Delta t} - s_0 | s_0 = s].$$

There is an equivalent definition of $a^{(i)}$, given by

$$\sum_{j=0}^{i} a_j^{(i)} j^k = \begin{cases} 0, & k \neq 1, \\ 1, & k = 1, \end{cases} \quad \text{for } 0 \leq j, k \leq i. \tag{14}$$

Note that this method differs from the finite difference scheme. In the classical finite difference scheme, the dynamics $\mu(s)$ is known, and the numerical scheme is used to approximate the trajectory $s_{j\Delta t}$. However, it is the opposite here. While the dynamics $\mu(s)$ is unknown, the trajectory $s_{j\Delta t}$ is used to approximate the dynamics. Consequently, the technique used to demonstrate the convergence and convergence rate of $\hat{V}_i(s)$ is also distinct from classical numerical analysis. The error analysis of PhiBE in the deterministic dynamics is established in the following theorem.

Theorem 3.2. *Assume that $\|\nabla r(s)\|_{L^\infty}, \left\|\mathcal{L}_\mu^i \mu(s)\right\|_{L^\infty}$ are bounded. Additionally, assume that $\|\nabla \mu(s)\|_{L^\infty} < \beta$, then the solution $\hat{V}_i(s)$ to the PhiBE* (11) *is an ith-order approximation to the true value function $V(s)$ defined in* (1) *with an error*

$$\left\|\hat{V}_i(s) - V(s)\right\|_{L^\infty} \leq 2C_i \frac{\|\nabla r\|_{L^\infty} \left\|\mathcal{L}_\mu^i \mu\right\|_{L^\infty}}{(\beta - \|\nabla \mu\|_{L^\infty})^2} \Delta t^i,$$

where

$$\mathcal{L}_\mu = \mu \cdot \nabla, \tag{15}$$

and C_i is a constant defined in (36) *that only depends on the order i.*

See Section 6.2 for the proof of Theorem 3.2. Note that the boundedness assumption of $\left\|\mathcal{L}_\mu^i \mu(s)\right\|_{L^\infty}$ is required in general to establish that $\hat{\mu}_i(s)$ is an i-th order approximation to $\mu(s)$ in Lemma 6.2. However, we will see later in Remark 5 and Theorem 3.3 that one can relax this assumption in linear dynamics.

Remark 3 (1st-order PhiBE v.s. BE)**.** *By Theorem 3.2, the distance between the first-order PhiBE solution and the true value function can be bounded by*

$$\left\|\hat{V}_1 - V\right\|_{L^\infty} \leq \frac{2\left\|\nabla r\right\|_{L^\infty}\left\|\mu\cdot\nabla\mu\right\|_{L^\infty}}{(\beta - \left\|\nabla\mu\right\|_{L^\infty})^2}\Delta t.$$

Comparing it with the difference between the BE solution and the true value function in deterministic dynamics,

$$\left\|\tilde{V} - V\right\|_{L^\infty} \leq \frac{\left\|\mu\nabla r\right\|_{L^\infty} + \beta\left\|r\right\|_{L^\infty}}{2\beta}\Delta t,$$

one observes that when the change of the reward is rapid, i.e., $\left\|\nabla r\right\|_{L^\infty}$ *is large, but the change in velocity is slow, i.e.,* $\left\|\frac{d^2}{dt^2}s_t\right\|_{L^\infty} = \left\|\mu\cdot\nabla\mu\right\|_{L^\infty}$ *is small, even though both* $\hat{V}_1$ *and* $\tilde{V}$ *are first-order approximations to the true value function,* $\hat{V}_1$ *has a smaller upper bound.*

Remark 4 (Higher-order PhiBE)**.** *The advantage of the higher-order PhiBE is two-fold. Firstly, it provides a higher-order approximation, enhancing accuracy compared to the first-order PhiBE or BE. Secondly, as demonstrated in Theorem 3.2, the approximation error of the i-th order PhiBE decreases as* $\left\|\mathcal{L}_\mu^i\mu\right\|_{L^\infty}$ *decreases. if the "acceleration", i.e.,* $\frac{d^2}{dt^2}s_t = \mathcal{L}_\mu\mu$*, of the dynamics is large but the change in acceleration, i.e.,* $\frac{d^3}{dt^3}s_t = \mathcal{L}_\mu^2\mu$*, is slow, then the error reduction with the second-order PhiBE will be even more pronounced in addition to the higher-order error effect.*

Remark 5 (Assumptions on $\left\|\mathcal{L}_\mu^i\mu(s)\right\|_{L^\infty}$)**.** *A sufficient condition for* $\left\|\mathcal{L}_\mu^i\mu(s)\right\|_{L^\infty}$ *being bounded is that* $\left\|\nabla^k\mu(s)\right\|_{L^\infty}$ *are bounded for all* $0 \leq k \leq i$*. Note that the linear dynamics* $\mu(s) = \lambda s$ *does not satisfy this condition. We lose some sharpness for the upper bound to make the theorem work for all general dynamics. However, we prove in Theorem 3.3 that PhiBE works when* $\mu(s) = \lambda s$*, and one can derive a sharper error estimate for this case.*

Additionally, when the underlying dynamics are linear, one can conduct a sharper error analysis for PhiBE.

Theorem 3.3. *When the underlying dynamics follows* $\frac{d}{dt}s_t = \lambda s_t$*, then the solution to the i-th order PhiBE in deterministic dynamics approximates the true value function with an error*

$$\left\|\hat{V}_i - V\right\|_{L^\infty} \leq C_i\frac{\lambda^{i+1}\left\|s\cdot\nabla r(s)\right\|_{L^\infty}}{\beta^2}\Delta t^i + o(\Delta t^i)$$

where C_i *is a constant defined in* (36) *that only depends on the order* i*.*

The proof of the above theorem is provided in Section 6.3. We also establish the upper bound for the BE in the same dynamics, and it turns out that the upper bound in Theorem 3.1 is already sharp. According to Theorem 3.3, the error of the i-th order PhiBE for linear dynamics depends on λ^{i+1}. This means that even if the dynamics $s_t = e^{\lambda t}s_0$ changes exponentially fast when $\lambda > 0$, as long as $\lambda \sim O(1)$, PhiBE solution is still a good approximation to the true value function. Especially when $\frac{\lambda}{\Delta t} < 1$, higher order PhiBE gives a smaller approximation error in addition to the higher-order effect.

3.2.2 Stochastic dynamics

When $\sigma(s) \not\equiv 0$ is a non-degenerate matrix, then the dynamics is stochastic and driven by the SDE in (2). By Feynman–Kac theorem [20], the value function $V(s)$ satisfies the following PDE,

$$\beta V(s) = r(s) + \mathcal{L}_{\mu,\Sigma}V(s), \tag{16}$$

where $\mathcal{L}_{\mu,\Sigma}$ is an operator defined in (4). However, one cannot directly solve the PDE (16) as $\mu(s), \sigma(s)$ are unknown. In the case where one only has access to the discrete-time transition distribution $\rho(s', \Delta t|s)$, we propose an i-th order PhiBE in the stochastic dynamics to approximate the true value function $V(s)$.

Definition 3 (i-th order PhiBE in stochastic dynamics)**.** *When the underlying dynamics are stochastic, then the i-th order PhiBE is defined as,*

$$\beta \hat{V}_i(s) = r(s) + \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i}\hat{V}_i(s), \tag{17}$$

where

$$\begin{aligned}
\hat{\mu}_i(s) &= \frac{1}{\Delta t}\sum_{j=1}^{i}\mathbb{E}_{s_{j\Delta t}\sim\rho(\cdot,j\Delta t|s)}\left[a_j^{(i)}(s_{j\Delta t}-s_0)|s_0=s\right]\\
\hat{\Sigma}_i(s) &= \frac{1}{\Delta t}\sum_{j=1}^{i}\mathbb{E}_{s_{j\Delta t}\sim\rho(\cdot,j\Delta t|s)}\left[a_j^{(i)}(s_{j\Delta t}-s_0)(s_{j\Delta t}-s_0)^\top|s_0=s\right]
\end{aligned} \tag{18}$$

where $\mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i}$ is defined in (4)*, and $a^{(i)} = (a_0^{(i)}, \cdots, a_i^{(i)})^\top$ is defined in* (13)*.*

Remark 6. *There is another i-th order approximation for $\Sigma(s)$,*

$$\tilde{\Sigma}_i(s) = \frac{1}{\Delta t}\sum_{j=0}^{i}a_j^{(i)}\left(\mathbb{E}[s_{j\Delta t}^\top s_{j\Delta t}|s_0=s] - \mathbb{E}[s_{j\Delta t}|s_0=s]^\top\mathbb{E}[s_{j\Delta t}|s_0=s]\right).$$

However, the unbiased estimate for $\tilde{\Sigma}(s)$

$$\tilde{\Sigma}_i(s_0) \approx \frac{1}{\Delta t}\sum_{j=1}^{i}a_j^{(i)}\left(s_{j\Delta t}^\top s_{j\Delta t} - s'_{j\Delta t}s_{j\Delta t}\right)$$

requires two independent samples $s_{j\Delta t}, s'_{j\Delta t}$ starting from s_0, which are usually unavailable in the RL setting. This is known as the "Double Sampling" problem. One could apply a similar idea in [26, 25] to alleviate the double sampling *problem when the underlying dynamics are smooth, that is, approximating $s'_{j\Delta t} \approx s_{(j-1)\Delta t} + (s_{(j+1)\Delta t} - s_{j\Delta t})$. However, it will introduce additional bias into the approximation. We leave the study of this approximation $\tilde{\Sigma}_i(s)$ or the application of BFF on $\tilde{\Sigma}_i(s)$ for future research.*

The first and second-order approximations are presented as follows. The first-order approximation reads,

$$\hat{\mu}_1(s) = \frac{1}{\Delta t}\mathbb{E}\left[(s_{\Delta t} - s_0)|s_0 = s\right], \quad \hat{\Sigma}_1(s) = \frac{1}{\Delta t}\mathbb{E}\left[(s_{\Delta t} - s_0)(s_{\Delta t} - s_0)^\top|s_0 = s\right];$$

and the second-order approximation reads,

$$\hat{\mu}_2(s) = \frac{1}{\Delta t}\mathbb{E}\left[2(s_{\Delta t} - s_0) - \frac{1}{2}(s_{2\Delta t} - s_0)|s_0 = s\right],$$
$$\hat{\Sigma}_2(s) = \frac{1}{\Delta t}\mathbb{E}\left[2(s_{\Delta t} - s_0)(s_{\Delta t} - s_0)^\top - \frac{1}{2}(s_{2\Delta t} - s_0)(s_{2\Delta t} - s_0)^\top|s_0 = s\right].$$

Next, we show the solution $\hat{V}_i(s)$ to the ith-order PhiBE provides a ith-order approximation to the true value function $V(s)$. To establish the error analysis, the following assumptions are required.

Assumption 1. *Assumptions on the dynamics*

(a) $\lambda_{min}(\Sigma(s)) > \lambda_{min} > 0$ *for* $\forall s \in \mathbb{S}$.

(b) $\left\|\nabla^k \mu(s)\right\|_{L^\infty}, \left\|\nabla^k \Sigma(s)\right\|_{L^\infty}$ *are bounded for* $0 \leq k \leq 2i$

The first assumption ensures the coercivity of the operator $\mathcal{L}_{\mu,\Sigma}$, which is necessary to establish the regularity of $V(s)$. The second assumption is employed to demonstrate that $\hat{\mu}_i$ and $\hat{\Sigma}_i$ are i-th approximations to μ, Σ, respectively. Additionally, under the above assumption, there exists a stationary distribution $\rho(s)$ to the stochastic dynamics that satisfies [7],

$$\int \mathcal{L}_{\mu,\Sigma}\phi(s)\rho(s)ds = 0. \quad \text{for } \forall \phi(s) \in C_c^\infty. \tag{19}$$

We define a weighted L^2 norm under the above stationary distribution,

$$\langle f, g\rangle_\rho = \int f(s)g(s)\rho(s)ds, \quad \|f\|_\rho^2 = \int f^2(s)\rho(s)ds.$$

Furthermore, one can bound

$$L_\rho = \|\nabla \log \rho\|_\rho \leq \frac{1}{\lambda_{\min}}\|\mu + \nabla\cdot\Sigma\|_\rho \leq \frac{\|\mu + \nabla\cdot\Sigma\|_{L^\infty}}{\lambda_{\min}} \tag{20}$$

by Theorem 1.1 of [2].

First, the error analysis for BE in the weighted L^2 norm is presented in the following theorem.

Theorem 3.4. *Assume that* $\|r\|_\rho, \|\mathcal{L}_{\mu,\Sigma} r\|_\rho$ *are bounded, then the solution* $\tilde{V}(s)$ *to the BE* (3) *approximates the true value function* $V(s)$ *defined in* (1) *with an error*

$$\left\|V(s) - \tilde{V}(s)\right\|_\rho \leq \frac{\frac{1}{\sqrt{3}}(\|\mathcal{L}_{\mu,\Sigma} r\|_\rho + \beta \|r\|_\rho)}{\beta} \Delta t + o(\Delta t).$$

The proof of Theorem 3.4 is given in Section 6.4. Next, the error analysis for PhiBE in stochastic dynamics is presented in the following theorem.

Theorem 3.5. *Under Assumption 1, and* $\Delta t^i \leq D_{\mu,\Sigma,\beta}$*, the solution* $\hat{V}_i(s)$ *to the i-th order PhiBE* (17) *is an i-th order approximation to the true value function* $V(s)$ *defined in* (1) *with an error*

$$\left\|V(s) - \hat{V}_i(s)\right\|_\rho \leq \left(\frac{C_{r,\mu,\Sigma}}{\beta^2} + \frac{\hat{C}_{r,\mu,\Sigma}}{\beta^{3/2}}\right) \Delta t^i,$$

where $D_{\mu,\Sigma,\beta}, C_{r,\mu,\Sigma}, \hat{C}_{\mu,\Sigma,\beta}$ *are constants defined in* (42), (43) *depending on* μ, Σ, r*.*

The proof of Theorem 3.5 is given in Section 6.5.

Remark 7 (PhiBE v.s. BE)**.** *Here we discuss two cases. The first case is when the diffusion is known, that is,* $\hat{\Sigma}_i = \Sigma$*, then the distance between the PhiBE and the true value function can be bounded by*

$$\left\|\hat{V}_i - V\right\|_\rho \lesssim \frac{1}{\beta^2} \left\|\mathcal{L}^i_{\mu,\Sigma}\mu\right\|_{L^\infty} \left[\left(\frac{\|\nabla\Sigma\|_{L^\infty}}{\lambda_{min}} + \sqrt{\frac{\|\nabla\mu\|_{L^\infty}}{\lambda_{min}}}\right)\|r\|_\rho + \|\nabla r\|_\rho\right] \Delta t^i.$$

Similar to the deterministic case, the error of the i-th order PhiBE proportional to the change rate of the dynamics $\mathbb{E}[\frac{d^{i+1}}{dt^{i+1}} s_t] = \mathcal{L}^i_{\mu,\Sigma}\mu$*. One can refer to Remarks 3 and 4 for the benefit of the 1st-order PhiBE and higher-order PhiBE with respect to different dynamics.*

The second case is when both drift and diffusion are unknown. Then the distance between the first-order PhiBE and the true value function can be bounded by

$$\begin{aligned}
&\left\|\hat{V}_1 - V\right\|_\rho \\
\lesssim &\frac{\Delta t}{\beta^2}\left[\left(L_\mu + L_{\nabla\cdot\Sigma} + \frac{L_\Sigma}{\lambda_{min}}\|\mu + \nabla\cdot\Sigma\|_{L^\infty}\right)\left(\sqrt{\frac{C_\nabla}{\lambda_{min}}}\|r\|_{\rho,L^\infty} + \|\nabla r\|_{\rho,L^\infty}\right)\right] \\
&+ \frac{\Delta t}{\beta^{3/2}}\frac{1}{\sqrt{\lambda_{min}}} L_\Sigma \left(\sqrt{\frac{C_\nabla}{\lambda_{min}}}\|r\|_\rho + \|\nabla r\|_\rho\right),
\end{aligned}$$

where

$$L_\mu \lesssim \|\mathcal{L}\mu\|, \quad L_\Sigma \lesssim \left\|\mu\mu^\top + \Sigma\nabla\mu + \mathcal{L}\Sigma\right\|_{L^\infty} + \|\mu\|_{L^\infty},$$

$$L_{\nabla\cdot\Sigma} \lesssim \sqrt{\frac{C_\nabla}{\lambda_{min}}} \left\|\mu\mu^\top + \Sigma\nabla\mu + \mathcal{L}\Sigma\right\|_{L^\infty} + \left\|\nabla\cdot\left(\mu\mu^\top + \Sigma\nabla\mu + \mathcal{L}\Sigma\right)\right\|_{L^\infty} + \sqrt{\frac{C_\nabla}{\lambda_{min}}} \|\mu\|_{L^\infty}$$

$$+ \|\nabla\mu\|_{L^\infty},$$

$$\sqrt{\frac{C_\nabla}{\lambda_{min}}} \lesssim \frac{\|\nabla\mu\|_{L^\infty}}{2\lambda} + \sqrt{\frac{\|\nabla\Sigma\|_{L^\infty}}{\lambda_{min}}}, \quad \|r\|_{\rho,L^\infty} = \|r\|_\rho + \|r\|_{L^\infty}.$$

Here the operator $\mathcal{L}$ represents $\mathcal{L}_{\mu,\Sigma}$. This indicates that when λ_{min} is large or $\nabla\mu, \nabla\Sigma$ are small, the difference between $\hat{V}_1$ and V is smaller. Comparing it with the upper bound $\|\mathcal{L}r\|_\rho + \beta\|r\|_\rho$ for the BE, which is more sensitive to β and reward function, the PhiBE approximation is less sensitive to these factors. When the change in the dynamics is slow, or the noise is large, even the first-order PhiBE solution is a better approximation to the true value function.

4 Model-free Algorithm for continuous-time Policy Eveluation

In this section, we assume that one only has access to the discrete-time trajectory data $\{s_0^l, s_{\Delta t}^l, \cdots s_{m\Delta t}^l\}_{l=1}^I$. We first revisit the Galerkin method for solving PDEs with known dynamics in Section 4.1, and we provide the error analysis of the Galerkin method for PhiBE. Subsequently, we introduce a model-free Galerkin method in Section 4.2.

4.1 Galerkin Method

Given n bases $\{\phi_i(s)\}_{i=1}^n$, the objective is to find an approximation $\bar{V}(s) = \Phi(s)^\top\theta$ to the solution $V(s)$ of the PDE,

$$\beta V(s) - \mathcal{L}_{\mu,\Sigma}V(s) = r(s),$$

where $\theta \in \mathbb{R}^n, \Phi(s) = (\phi_1(s), \cdots, \phi_n(s))^\top$, and $\mathcal{L}_{\mu,\Sigma}$ is defined in (4). The Galerkin method involves inserting the ansatz $\bar{V}$ into the PDE and then projecting it onto the finite bases,

$$\left\langle \beta\bar{V}(s) - \mathcal{L}_{\mu,\Sigma}\bar{V}(s), \Phi(s)\right\rangle_\rho = \langle r(s), \Phi(s)\rangle_\rho,$$

which results in a linear system of θ,

$$A\theta = b, \quad A = \langle\beta\Phi(s) - \mathcal{L}_{\mu,\Sigma}\Phi(s), \Phi(s)\rangle_\rho, \quad b = \langle r(s), \Phi(s)\rangle_\rho.$$

When the dynamics $\mu(s), \Sigma(s)$ are known, one can explicitly compute the matrix A and the vector b, and find the parameter $\theta = A^{-1}b$ accordingly.

In continuous-time policy evaluation problems, one does not have access to the underlying dynamics μ, Σ. However, the approximated dynamics $\hat{\mu}_i, \hat{\Sigma}_i$ is given through PhiBE. Therefore, if one has access to the discrete-time transition distribution, then the parameter $\theta = \hat{A}_i^{-1} b$ can be solved for by approximating A with $\hat{A}_i$

$$\hat{A}_i = \left\langle \beta \Phi(s) - \mathcal{L}_{\hat{\mu}_i, \hat{\Sigma}_i} \Phi(s), \Phi(s) \right\rangle_\rho,$$

where $\hat{\mu}_i, \hat{\Sigma}_i$ are defined in (18). We give the error estimate of the Galerkin method for PhiBE in the following theorem.

Theorem 4.1. *The Galerkin solution $\hat{V}_i^G(s) = \theta^\top \Phi(s)$ satisfies*

$$\left\langle (\beta - \mathcal{L}_{\hat{\mu}_i, \hat{\Sigma}_i}) \hat{V}_i^G(s), \Phi \right\rangle_\rho = \langle r(s), \Phi(s) \rangle_\rho. \tag{21}$$

When $L_\rho^\infty = \|\nabla \log \rho\|_{L^\infty}$ is bounded for the stationary distribution ρ, then as long as $\Delta t^i \leq \min\{\eta_{\mu,\Sigma,\beta}, D_{\mu,\Sigma,\beta}\}$, the Galerkin solution $\hat{V}_i^G(s)$ approximates the true value function defined in (1) *with an error*

$$\left\| \hat{V}_i^G(s) - V(s) \right\|_\rho \leq \left(\frac{C_{r,\mu,\Sigma}}{\beta^2} + \frac{\hat{C}_{r,\mu,\Sigma}}{\beta^{3/2}} \right) \Delta t^i + C_G \inf_{V^P = \theta^\top \Phi} \left\| \hat{V}_i - V^P \right\|_{H_\rho^1}.$$

When $\|\nabla \log \rho\|_{L^\infty}$ is unbounded, assume that the bases $L_\Phi^\infty = \|\Phi\|_{L^\infty}$ is bounded, then as long as $\Delta t^i \leq \min\{\hat{\eta}_{\mu,\Sigma,\beta}, D_{\mu,\Sigma,\beta}\}$, the Galerkin solution $\hat{V}_i^G(s)$ approximates the true value function defined in (1) *with an error*

$$\left\| \hat{V}_i^G(s) - V(s) \right\|_\rho \leq \left(\frac{C_{r,\mu,\Sigma}}{\beta^2} + \frac{\hat{C}_{r,\mu,\Sigma}}{\beta^{3/2}} \right) \Delta t^i + \hat{C}_G \inf_{V^P = \theta^\top \Phi} \left\| \hat{V}_i - V^P \right\|_{H_{\rho,\infty}^1}.$$

where $\eta_{\mu,\Sigma,\beta}, C_G, \hat{\eta}_{\mu,\Sigma,\beta}, \hat{C}_G, D_{\mu,\Sigma,\beta}$ are constants depending on $\mu, \Sigma, \beta, L_\rho^\infty, L_\Phi^\infty$ defined in (61), (62), (65), (66), (42) *respectively, and $\|f\|_{H_\rho^1} = \|f\|_\rho + \|\nabla f\|_\rho$, $\|f\|_{H_{\rho,\infty}^1} = \|f\|_{H_\rho^1} + \|\nabla f\|_{L^\infty}$.*

The proof of the Theorem is given in Section 6.6.

4.2 Model-free Galerkin method for PhiBE

When only discrete-time trajectory data is available, we first develop an unbiased estimate $\bar{\mu}_i, \bar{\Sigma}_i$ for $\hat{\mu}_i, \hat{\Sigma}_i$ from the trajectory data,

$$\begin{aligned} \bar{\mu}_i(s_{j\Delta t}^l) &= \frac{1}{\Delta t} \sum_{k=1}^{i} a_k^{(i)} (s_{(j+k)\Delta t}^l - s_{j\Delta t}^l), \\ \bar{\Sigma}_i(s_{j\Delta t}^l) &= \frac{1}{\Delta t} \sum_{k=1}^{i} a_k^{(i)} (s_{(j+k)\Delta t}^l - s_{j\Delta t}^l)(s_{(j+k)\Delta t}^l - s_{j\Delta t}^l)^\top, \end{aligned} \tag{22}$$

with $a^{(i)}$ defined in (13). Then, using the above unbiased estimate, one can approximate the matrix $\hat{A}$ and the vector b by

$$\bar{A}_i = \sum_{l=1}^{I}\sum_{j=0}^{m-i} \Phi(s^l_{j\Delta t}) \left[\beta\Phi(s^l_{j\Delta t}) - \mathcal{L}_{\bar{\mu}_i(s^l_{j\Delta t}),\bar{\Sigma}_i(s^l_{j\Delta t})}\Phi(s^l_{j\Delta t})\right]^\top,$$
$$\bar{b}_i = \sum_{l=1}^{I}\sum_{j=0}^{m-i} r(s^l_{j\Delta t})\Phi(s^l_{j\Delta t}).$$

By solving the linear system $\bar{A}_i\theta = \bar{b}_i$, one obtains the approximated value function $\bar{V}(s) = \Phi(s)^\top\theta$ in terms of the finite bases. Note that our algorithm can also be applied to stochastic rewards or even unknown rewards as only observation of rewards is required at discrete time. We summarize the model-free Galerkin method for deterministic and stochastic dynamics in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Model-free Galerkin method for i-th order PhiBE in deterministic dynamics

---

Given: discrete time step Δt, discount coefficient β, discrete-time trajectory data $\{(s^l_{j\Delta t}, r^l_{j\Delta t})_{j=0}^m\}_{l=1}^I$ generated from the underlying dynamics, and a finite bases $\Phi(s) = (\phi_1(s), \cdots, \phi_n(s))^\top$.

1: Calculate $\bar{A}_i$:

$$\bar{A}_i = \sum_{l=1}^{I}\sum_{j=0}^{m-i} \Phi(s^l_{j\Delta t}) \left[\beta\Phi(s^l_{j\Delta t}) - \bar{\mu}_i(s^l_{j\Delta t})\cdot\nabla\Phi(s^l_{j\Delta t})\right]^\top,$$

where $\bar{\mu}_i(s^l_{j\Delta t})$ is defined in (22).

2: Calculate $\bar{b}_i$:

$$\bar{b}_i = \sum_{l=1}^{I}\sum_{j=0}^{m-i} r^l_{j\Delta t}\Phi(s^l_{j\Delta t}).$$

3: Calculate θ:

$$\theta = \bar{A}_i^{-1}\bar{b}_i.$$

4: Output $\bar{V}(s) = \theta^\top\Phi(s)$.

---

5 Numerical experiments

5.1 Deterministic dynamics

We first consider deterministic dynamics, where the state space $\mathbb{S}$ is defined as $\mathbb{S} = [-\pi, \pi]$. We consider two kinds of underlying dynamics, one is linear,

$$\frac{d}{dt}s_t = \lambda s_t, \tag{23}$$

Algorithm 2 Model-free Galerkin method for i-th order PhiBE in stochastic dynamics

Given: discrete time step Δt, discount coefficient β, discrete-time trajectory data $\{(s^l_{j_{\Delta t}}, r^l_{j_{\Delta t}})_{j=0}^m\}_{l=1}^I$ generated from the underlying dynamics, and a finite bases $\Phi(s) = (\phi_1(s), \cdots, \phi_n(s))^\top$.

1: Calculate $\bar{A}_i$:

$$\bar{A}_i = \sum_{l=1}^I \sum_{j=0}^{m-i} \Phi(s^l_{j\Delta t}) \left[\beta\Phi(s^l_{j\Delta t}) - \bar{\mu}_i(s^l_{j\Delta t}) \cdot \nabla\Phi(s^l_{j\Delta t}) - \frac{1}{2}\bar{\Sigma}_i(s^l_{j\Delta t}) : \nabla^2\Phi(s^l_{j\Delta t})\right]^\top,$$

where $\bar{\mu}_i(s^l_{j\Delta t}), \bar{\Sigma}_i(s^l_{j\Delta t})$ are defined in (22).

2: Calculate $\bar{b}_i$:

$$\bar{b}_i = \sum_{l=1}^I \sum_{j=0}^{m-i} r^l_{j\Delta t}\Phi(s^l_{j\Delta t}).$$

3: Calculate θ:

$$\theta = \bar{A}_i^{-1}\bar{b}_i.$$

4: Output $\bar{V}(s) = \theta^\top\Phi(s)$.

and the other is nonlinear,

$$\frac{d}{dt}s_t = \lambda\sin^2(s_t). \tag{24}$$

The reward is set to be $r(s) = \beta\cos(ks)^3 - \lambda s(-3k\cos^3(ks)\sin(ks))$ for the linear case and $r(s) = \beta\cos^3(ks) - \lambda\sin^2(s)(-3k\cos^2(ks)\sin(ks))$ for the nonlinear case, where the value function can be exactly obtained, $V(s) = \cos^3(ks)$ in both cases. We use periodic bases $\{\phi_n(s_1)\}_{k=1}^{2M+1} = \frac{1}{\sqrt{\pi}}\{\frac{1}{\sqrt{2}}, \cos(ms_1), \sin(ms_1)\}_{m=1}^M$ with M large enough so that the solution can be accurately represented by these finite bases.

For the linear dynamics, the discrete-time transition dynamics are

$$p_{\Delta t}(s) = e^{\lambda\Delta t}s.$$

Hence, one can express the BE as

$$\tilde{V}(s) = r(s)\Delta t + e^{-\beta\Delta t}\tilde{V}(e^{\lambda\Delta t}s), \tag{25}$$

and i-th order PhiBE as

$$\beta\hat{V}_i(s) = r(s) + \frac{1}{\Delta t}\left[\sum_{k=1}^i a_k^{(i)}(e^{\lambda k\Delta t}s - s)\right] \cdot \nabla\hat{V}_i(s), \tag{26}$$

respectively for $a^{(i)}$ defined in (13). For the nonlinear dynamics, we approximate $p_{\Delta t}(s)$ and generate the trajectory data numerically,

$$s_{t+\delta} = s_t + \delta\lambda\sin^2(s_t)$$

with $\delta = 10^{-4}$ sufficiently small.

The trajectory data are generated from J different initial values $s_0 \sim$Unif$[-\pi, \pi]$, and each trajectory has $m = 4$ data points, $\{s_0, \cdots, s_{(m-1)\Delta t}\}$. Algorithm 1 is used to solve for the PhiBE, and LSTD is used to solve for BE. LSTD is similar to Algorithm 1 except that one uses $\tilde{A}$ derived from the BE (3),

$$\tilde{A} = \sum_{l=1}^{I} \sum_{j=0}^{m-2} \Phi(s^l_{j\Delta t}) \left[\Phi(s^l_{j\Delta t}) - e^{-\beta \Delta t}\Phi(s^l_{(j+1)\Delta t})\right]^\top \tag{27}$$

instead of $\bar{A}_i$.

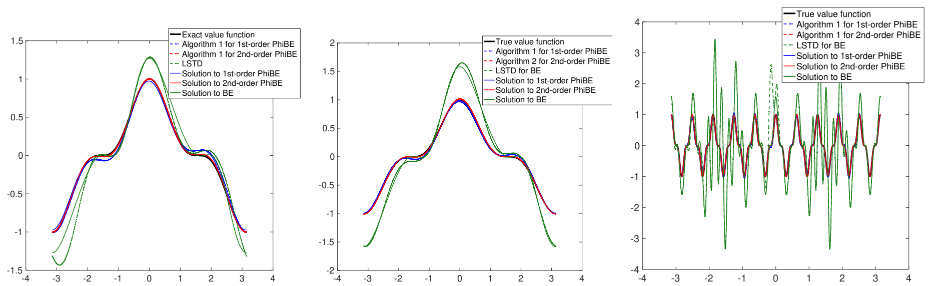
In Figure 1, the data are generated from the linear dynamics (23) with $\lambda = 0.05$ and collected at different Δt. We compare the solution to the second-order PhiBE with the solution to BE (when the discrete-time transition dynamics are known), and the performance of LSTD with the proposed Algorithm 1 (when only trajectory data are available) with different data collection interval Δt, discount coefficient β and oscillation of reward k. Note that the exact solution to BE is computed as $\tilde{V}(s) = \sum_{i=0}^{I} r(e^{\lambda \Delta t i} s)$ with I large enough, and the exact solution to PhiBE is calculated by applying the Galerkin method to (26).

In Figure 2, the data are generated from the nonlinear dynamics (24) and collected at different Δt. We compare the solutions to the first-order and second-order PhiBE with the solution to the BE (when the discrete-time transition dynamics are known), and the performance of LSTD with the proposed Algorithm 1 (when only trajectory data are available) with different $\Delta t, \beta, k, \lambda$.

In Figure 3, the distances of the solution from PhiBE, BE to the true value function are plotted as $\Delta t \to 0$; the distances of the approximated solution by Algorithm 1 and LSTD to the true value function are plotted as the amount of data increases. Here, the distance is measured using the L^2 norm

$$D(V, \hat{V}) = \sqrt{\int_{-\pi}^{\pi} (V(s) - \hat{V}(s))^2 ds}. \tag{28}$$

In Figures 1 and 2, when the discrete-time transition dynamics are known, PhiBE solution is much closer to the true value function compared to the BE solution in all the experiments. Especially, the second-order PhiBE solution is almost identical to the exact value function. Additionally, when only trajectory data is available, one can approximate the solutions to PhiBE very well with only 40 or 400 data points. Particularly, when $\Delta t = 5$ is large, the solution to PhiBE still approximates the true solution very well, which indicates that one can collect data sparsely based on PhiBE. Moreover, the solution to PhiBE is not sensitive to the oscillation of the reward function, which implies that one has more flexibility in designing the reward function in the RL problem. Besides, unlike BE, the error increases when β is too small or too large, while the error for PhiBE decays as β increases. Furthermore, it's noteworthy that in Figure 2/(b) and (c), for relatively large changes in the dynamics indicated by $\|\nabla \mu\| \leq \lambda = 5$ and 2, respectively, PhiBE still provides a good approximation.



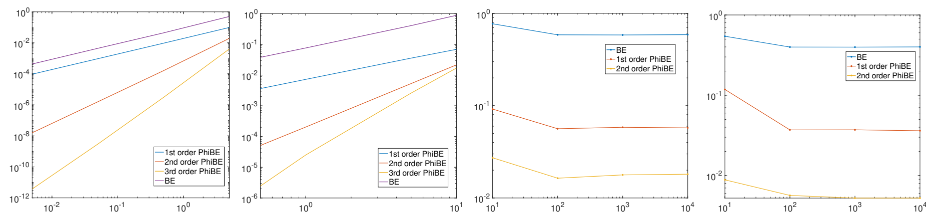
(a) Nonlinear dynamics (24) with $\Delta t = 5, \beta = 0.1, k = 1, \lambda = 0.1$

(b) Nonlinear dynamics (24) with $\Delta t = 0.1, \beta = 10, k = 10, \lambda = 5$

(c) Nonlinear dynamics (24) with $\Delta t = 0.1, \beta = 10, k = 10, \lambda = 2$

Figure 2: The PhiBE solution and the BE solution, when the discrete-time transition dynamics are given, are plotted in solid lines. The approximated PhiBE solution based on Algorithm 1 and the approximated BE solution based on LSTD, when discrete-time data is given, are plotted in dash lines. Both algorithms utilize the same data points.

In Figure 3/(a) and (b), one can observe that the solution for BE approximates the true solution in the first order, while the solution for i-th order PhiBE approximates the true solution in i-th order. In Figure 3/(c) and (d), one can see that as the amount of data increases, the error from the LSTD algorithm stops decreasing when it reaches 10^{-1}. This is because the error between BE and the true value function $\left\|\tilde{V} - V\right\| = O(\Delta t)$ dominates the data error. On the other hand, for higher-order PhiBE, as the amount of data increases, the performance of the algorithm improves, and the error can achieve $O(\Delta t^i)$.



(a) Linear dynamics (24). (b) Nonlinear dynamics (24). (c) Linear dynamics (24). (d) Nonlinear dynamics (24).

Figure 3: The L^2 error (28) of the PhiBE solutions and the BE solutions with decreasing Δt are plotted in the left two figures. The L^2 error (28) of the approximated PhiBE solutions and the approximated BE solutions with increasing amount of data collected every $\Delta t = 5$ unit of time are plotted in the right two figures. We set $\lambda = 0.05, \beta = 0.1, k = 1$ in both linear and nonlinear cases.

5.2 Stochastic dynamics

We consider the Ornstein–Uhlenbeck process,

$$ds(t) = \lambda s dt + \sigma dB_t, \tag{29}$$

with $\lambda = 0.05, \sigma = 1$. Here the reward is set to be $r(s) = \beta\cos^3(ks) - \lambda s(-3k\cos^2(ks)\sin(ks)) - \frac{1}{2}\sigma^2(6k^2\cos(s)\sin^2(ks) - 3k^2\cos^3(ks))$, where the value function can be exactly obtained, $V(s) = \cos^3(ks)$. For OU process, since the conditional density function for s_t given $s_0 = s$ follows the normal distribution with expectation $se^{\lambda t}$, variance $\frac{\sigma^2}{2\lambda}(e^{2\lambda t} - 1)$. Both PhiBE and BE have explicit forms. One can express PhiBE as,

$$\begin{aligned}\beta\hat{V}_i(s) =& r(s) + \frac{1}{\Delta t}\sum_{k=1}^{i} a_k^{(i)}(e^{\lambda k\Delta t} - 1)s\nabla\hat{V}(s)\\ &+ \frac{1}{2\Delta t}\sum_{k=1}^{i} a_k^{(i)}\left[\frac{\sigma^2}{2\lambda}(e^{2\lambda k\Delta t} - 1) + (e^{\lambda k\Delta t} - 1)^2 s^2\right]\Delta\hat{V}(s);\end{aligned} \tag{30}$$

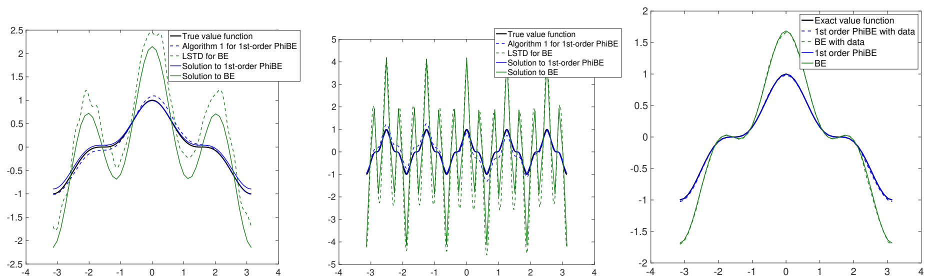
and BE as,

$$\begin{aligned}\tilde{V}(s) =& r(s)\Delta t + e^{-\beta\Delta t}\mathbb{E}\left[\tilde{V}(s_{t+1})|s_t = s\right]\\ =& r(s)\Delta t + e^{-\beta\Delta t}\int_{\mathbb{S}}\tilde{V}(s')\rho_{\Delta t}(s', s)ds',\end{aligned} \tag{31}$$

where

$$\rho_{\Delta t}(s', s) = \frac{1}{\sqrt{2\pi}\hat{\sigma}}\exp\left(-\frac{1}{2\hat{\sigma}^2}(s' - se^{\lambda\Delta t})^2\right), \quad \text{with } \hat{\sigma} = \frac{\sigma^2}{2\lambda}(e^{2\lambda\Delta t} - 1).$$

In Figure 4, we compare the exact solution and approximated solution to PhiBE and BE, respectively, for different $\Delta t, \beta, k$. In Figure 5/(a), the decay of



(a) Stochastic dynamics (29) with $\Delta t = 1, \beta = 0.1, k = 1$.

(b) Stochastic dynamics (29) with $\Delta t = 0.1, \beta = 0.1, k = 5$.

(c) Stochastic dynamics (29) with $\Delta t = 0.1, \beta = 10, k = 1$.

Figure 4: The PhiBE solution and the BE solution, when the discrete-time transition dynamics are given, are plotted in solid lines. The approximated PhiBE solution based on Algorithm 2 and the approximated BE solution based on LSTD, when discrete-time data is given, are plotted in dash lines. Both algorithms utilize the same data points.

the error as $\Delta t \to 0$ for the exact solutions to PhiBE and BE are plotted. In Figure 5 /(b), the decay of the approximated solution to PhiBE and BE based on Algorithm 2 and LSTD are plotted with an increasing amount of data.

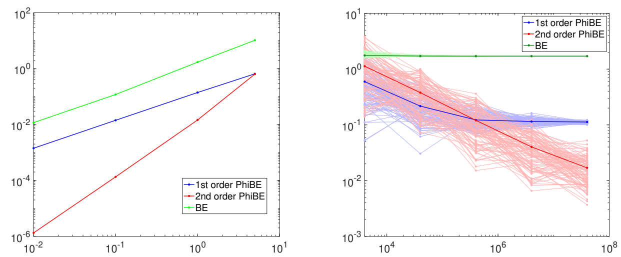
We observe similar performance in the stochastic dynamics as in the deterministic dynamics, as shown in Figures 4 and 5. In Figure 5, the variance of the higher order PhiBE is larger than that of the first-order PhiBE because it involves more future steps. However, note that the error is plotted on a logarithmic scale. Therefore, when the error is smaller, although the variance appears to have the same width on the plot, it is actually much smaller. Particularly, when the amount of the data exceeds 10^6, the variance is smaller than 10^{-1}.

6 Proofs

6.1 Proof of Theorem 3.1.

Proof. Let $\rho(s', t|s)$ be the probability density function of s_t that starts from $s_0 = s$, then it satisfies the following PDE

$$\partial_t \rho(s', t|s) = \nabla \cdot [\mu(s')\rho(s', t|s)] + \frac{1}{2}\sum_{i,j} \partial_{s_i}\partial_{s_j}[\Sigma_{ij}(s')\rho(s', t|s)]. \tag{32}$$



(a) Model error w.r.t. the discretization time Δt.

(b) Approximation error w.r.t the amount of data.

Figure 5: The L^2 error (28) of the PhiBE solutions and the BE solutions with decreasing Δt are plotted in (a). The L^2 error (28) of the approximated PhiBE solutions and the approximated BE solutions with increasing amount of data collected every $\Delta t = 5$ unit of time are plotted in (b). We set $\beta = 0.1, k = 1$ in both figures.

with initial data $\rho(s', 0|s) = \delta_s(s')$. Let $f(t,s) = e^{-\beta t} r(s)$, then

$$\begin{aligned} V(s) - \tilde{V}(s) &= \mathbb{E}\left[\sum_{i=0}^{\infty} \int_{\Delta t i}^{\Delta t(i+1)} f(t, s_t) - f(\Delta t i, s_{\Delta t i}) dt | s_0 = s\right] \\ &= \sum_{i=0}^{\infty} \int_{\Delta t i}^{\Delta t(i+1)} \left(\int_{\mathbb{S}} f(t,s')\rho(s',t|s) - f(\Delta t i, s')\rho(\Delta t i, s') ds'\right) dt. \end{aligned} \tag{33}$$

Since

$$\begin{aligned} &\int_{\mathbb{S}} f(t,s')\rho(s',t|s) - f(\Delta t i, s')\rho(\Delta t i, s')ds' \\ =& \int_{\mathbb{S}} f(t,s')(\rho(s',t|s) - \rho(s',\Delta t i|s)) + (f(t,s') - f(\Delta t i, s'))\rho(s',\Delta t i|s)ds' \\ =& \int_{\mathbb{S}} f(t,s')\partial_t \rho(s',\xi_1|s)(t - \Delta t i) + \partial_t f(\xi_2, s')(t-\Delta t i)\rho(s',\Delta t i|s)ds' \\ &\text{where } \xi_1, \xi_2 \in (\Delta t i, \Delta t(i+1)) \\ =& \int_{\mathbb{S}} \mathcal{L}_{\mu,\Sigma} f(t,s'), \rho(s',\xi_1|s)(t-\Delta t i)ds' - \int_{\mathbb{S}} \beta e^{-\beta \xi_2} r(s')\rho(s',\Delta t i|s)(t-\Delta t i)ds' \\ =& \left(e^{-\beta t}\int_{\mathbb{S}} \mathcal{L}_{\mu,\Sigma} r(s')\rho(s',\xi_1|s)ds' - \beta e^{-\beta\xi_2}\int_{\mathbb{S}} r(s')\rho(s',\Delta t i|s)ds'\right)(t-\Delta t i), \end{aligned} \tag{34}$$

where the second equality is due to the mean value theorem, and the third equality is obtained by inserting the equation (32) for $\rho(s', t|s)$ and integrating by parts. Therefore, for $t \in [\Delta t i, \Delta t(i+1)]$,

$$\begin{aligned} &\left|\int_{\mathbb{S}} f(t,s')\rho(s',t|s) - f(\Delta t i, s')\rho(\Delta t i, s')ds'\right| \\ \le& \left\|\mathcal{L}_{\mu,\Sigma} r\right\|_{L^\infty} e^{-\beta\Delta t i}(t-\Delta t i) + \beta e^{-\beta \Delta t i}\left\|r\right\|_{L^\infty}(t-\Delta t i). \end{aligned}$$

Therefore, one has

$$
\begin{aligned}
&\left\|V(s)-\tilde{V}(s)\right\|_{L^\infty}\\
\leq&\sum_{i=0}^{\infty}\int_{\Delta ti}^{\Delta t(i+1)}\left\|\mathcal{L}_{\mu,\Sigma}r\right\|_{L^\infty}e^{-\beta\Delta ti}(t-\Delta ti)+\beta e^{-\beta\Delta ti}\left\|r\right\|_{L^\infty}(t-\Delta ti)dt\\
\leq&\left(\left\|\mathcal{L}_{\mu,\Sigma}r\right\|_{L^\infty}+\beta\left\|r\right\|_{L^\infty}\right)\sum_{i=0}^{\infty}e^{-\beta\Delta ti}\int_{\Delta ti}^{\Delta t(i+1)}(t-\Delta ti)dt\\
\leq&\frac{1}{2}\left(\left\|\mathcal{L}_{\mu,\Sigma}r\right\|_{L^\infty}+\beta\left\|r\right\|_{L^\infty}\right)\sum_{i=0}^{\infty}e^{-\beta\Delta ti}\Delta t^2=\frac{C}{1-e^{-\beta\Delta t}}\Delta t^2=\frac{C}{\beta}\Delta t+C\left(\frac{1}{1-e^{-\beta\Delta t}}\Delta t^2-\frac{\Delta t}{\beta}\right),
\end{aligned}
$$

where $C=\frac{1}{2}\left(\left\|\mathcal{L}_{\mu,\Sigma}r\right\|_{L^\infty}+\beta\left\|r\right\|_{L^\infty}\right)$. Since

$$
\lim_{\Delta t\to 0}C\left(\frac{1}{1-e^{-\beta\Delta t}}\Delta t^2-\frac{\Delta t}{\beta}\right)\frac{1}{\Delta t}=0,
$$

one has,

$$
\left\|V(s)-\tilde{V}(s)\right\|_{L^\infty}=\frac{L\Delta t}{\beta}+o(\Delta t).
$$

□

6.2 Proof of Theorem 3.2

Note that the true value function V and the i-th order PhiBE solution $\hat{V}$ satisfies

$$
\beta V(s)=r(s)+\mu(s)\cdot\nabla V(s),\quad \beta\hat{V}_i(s)=r(s)+\hat{\mu}_i(s)\cdot\nabla\hat{V}_i(s).
$$

First, by the following lemma, one can bound $\left\|V-\hat{V}_i\right\|_{L^\infty}$ with $\left\|\mu-\hat{\mu}_i\right\|_{L^\infty}$.

Lemma 6.1. *For function V and $\hat{V}$ satisfies,*

$$
\beta V(s)=r(s)+\mu(s)\cdot\nabla V(s),\quad \beta\hat{V}(s)=r(s)+\hat{\mu}(s)\cdot\nabla\hat{V}(s),
$$

the distance between V and $\hat{V}$ can be bounded by

$$
\left\|V-\hat{V}\right\|_{L^\infty}\leq\frac{2\left\|\mu-\hat{\mu}\right\|_{L^\infty}\left\|\nabla r\right\|_{L^\infty}}{(\beta-\left\|\nabla\mu\right\|_{L^\infty})^2}.
$$

(See Section 6.2.1 for the proof of the above lemma.) Therefore, one has

$$
\left\|V-\hat{V}_i\right\|_{L^\infty}\leq\frac{\left\|\mu-\hat{\mu}_i\right\|_{L^\infty}\left\|\nabla r\right\|_{L^\infty}}{(\beta-\left\|\nabla\mu\right\|_{L^\infty})^2}. \tag{35}
$$

Then by the following lemma, one can further bound $\left\|\mu-\hat{\mu}_i\right\|_{L^\infty}$.

Lemma 6.2. *The distance between $\hat{\mu}_i(s)$ defined in* (12) *and the true dynamics can be bounded by*

$$\|\hat{\mu}_i(s) - \mu(s)\|_{L^\infty} \le C_i \left\|\mathcal{L}_\mu^i \mu(s)\right\|_{L^\infty} \Delta t^i,$$

where $\mathcal{L}_\mu$ is defined in (15)*, and*

$$C_i = \frac{\sum_{j=0}^{i} |a_j^{(i)}| j^{i+1}}{(i+1)!}. \tag{36}$$

(See Section 6.2.2 for the proof of the above lemma.) Hence, one completes the proof by applying the above lemma to (35)

$$\left\|V - \hat{V}_i\right\|_{L^\infty} \le \frac{C_i \left\|\mathcal{L}_\mu^i \mu\right\|_{L^\infty} \|\nabla r\|_{L^\infty}}{(\beta - \|\nabla \mu\|_{L^\infty})^2} \Delta t^i.$$

6.2.1 Proof of Lemma 6.1

By Feynman–Kac theorem, it is equivalently to write $\hat{V}$ as,

$$\hat{V} = \int_0^\infty e^{-\beta t} r(\hat{s}_t) dt \quad \text{with} \quad \frac{d}{dt}\hat{s}_t = \hat{\mu}(s_t).$$

Hence,

$$\begin{aligned} \left|V(s) - \hat{V}(s)\right| &= \left|\int_0^\infty e^{-\beta t}(r(s_t) - r(\hat{s}_t))dt\right| = \left|\int_0^\infty e^{-\beta t}\left(\int_{s_t}^{\hat{s}_t} \nabla r(s) ds\right) dt\right| \\ &\le \|\nabla r\|_{L^\infty} \int_0^\infty e^{-\beta t} \left|\hat{s}_t - s_t\right| dt, \end{aligned} \tag{37}$$

where

$$\frac{d}{dt}s_t = \mu(s_t), \quad \frac{d}{dt}\hat{s}_t = \hat{\mu}(\hat{s}_t), \quad s_0 = \hat{s}_0 = s. \tag{38}$$

Subtracting the two equations in (38) and multiplying it with $(\hat{s}_t - s_t)^\top$ gives

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\hat{s}_t - s_t\|^2 =& (\hat{\mu}(\hat{s}_t)) - \mu(s_t))(\hat{s}_t - s_t) \\ =& \left((\hat{\mu}(\hat{s}_t) - \mu(\hat{s}_t) + (\mu(\hat{s}_t) - \mu(s_t))\right)(\hat{s}_t - s_t) \\ \le& \|\mu - \hat{\mu}\|_{L^\infty} \|\hat{s}_t - s_t\| + \|\nabla \mu(s)\|_{L^\infty} \|\hat{s}_t - s_t\|^2 \\ \le& \frac{1}{2\epsilon}\|\mu - \hat{\mu}\|_{L^\infty}^2 + \left(\frac{\epsilon}{2} + \|\nabla \mu(s)\|_{L^\infty}\right)\|\hat{s}_t - s_t\|^2, \quad \text{for } \forall \epsilon > 0, \end{aligned}$$

where the mean value theorem is used in the first inequality. This implies

$$\|\hat{s}_t - s_t\|_2 \le \frac{1}{\sqrt{\epsilon}} \|\mu - \hat{\mu}\|_{L^\infty} t^{1/2} e^{(\epsilon/2 + \|\nabla \mu\|_{L^\infty})t}.$$

Inserting the above inequality back to (37) gives

$$\left\|V(s)-\hat{V}(s)\right\|_{L^\infty} \leq \frac{1}{\sqrt{\epsilon}}\left\|\nabla r\right\|_{L^\infty}\left\|\mu-\hat{\mu}\right\|_{L^\infty}\int_0^\infty e^{-(\beta-\epsilon/2-\|\nabla\mu\|_{L^\infty})t}t^{1/2}dt$$
$$=\frac{\sqrt{\pi}\left\|\mu-\hat{\mu}\right\|_{L^\infty}\left\|\nabla r\right\|_{L^\infty}}{2\sqrt{\epsilon}(\beta-\epsilon/2-\left\|\nabla\mu(s)\right\|_{L^\infty})^{3/2}}.$$

Assigning $\epsilon=\frac{1}{2}(\beta-\left\|\nabla\mu(s)\right\|_{L^\infty})$ to the above inequality completes the proof.

6.2.2 Proof of Lemma 6.2

By Taylor expansion, one has

$$s_{j\Delta t}=\sum_{k=0}^{i}\frac{(j\Delta t)^k}{k!}\left(\frac{d^k}{dt^k}s_t\Big|_{t=0}\right)+\frac{(j\Delta t)^{i+1}}{(i+1)!}\left(\frac{d^{i+1}}{dt^{i+1}}s_t\Big|_{t=\xi_j}\right)$$

with $\xi_j\in(0,j\Delta t)$. Inserting it into $\hat{\mu}_i(s)$ gives,

$$\begin{aligned}
\hat{\mu}_i(s)=&\frac{1}{\Delta t}\sum_{j=0}^{i}a_j^{(i)}[s_{j\Delta t}|s_0=s]\\
=&\frac{1}{\Delta t}\sum_{j=0}^{i}a_j^{(i)}\left[\sum_{k=0}^{i}\left(\frac{d^k}{dt^k}s_t\Big|_{t=0}\right)\frac{(\Delta tj)^k}{k!}+\left(\frac{d^{i+1}}{dt^{i+1}}s_t\Big|_{t=\xi_j}\right)\frac{(\Delta tj)^{i+1}}{(i+1)!}\right]\\
=&\frac{1}{\Delta t}\sum_{k=0}^{i}\left(\frac{d^k}{dt^k}s_t\Big|_{t=0}\right)\frac{(\Delta t)^k}{k!}\sum_{j=0}^{i}a_j^{(i)}j^k+\frac{1}{\Delta t}\sum_{j=0}^{i}a_j^{(i)}\left(\frac{d^{i+1}}{dt^{i+1}}s_t\Big|_{t=\xi_j}\right)\frac{(\Delta tj)^{i+1}}{(i+1)!}\\
=&\left(\frac{d}{dt}s_t\Big|_{t=0}\right)+\frac{\Delta t^i}{(i+1)!}\sum_{j=0}^{i}a_j^{(i)}j^{i+1}\left(\frac{d^{i+1}}{dt^{i+1}}s_t\Big|_{t=\xi_j}\right),
\end{aligned}$$

where the last equality is due to the definition of $a^{(i)}$ in (14). Since

$$\frac{d}{dt}s_t\Big|_{t=0}=\mu(s_0)=\mu(s),$$

one has

$$|\hat{\mu}_i(s)-\mu(s)|=\frac{\Delta t^i}{(i+1)!}\left|\sum_{j=0}^{i}a_j^{(i)}j^{i+1}\left(\frac{d^{i+1}}{dt^{i+1}}s_t\Big|_{t=\xi_j}\right)\right|.$$

Since

$$\frac{d^{i+1}}{dt^{i+1}}s_t=\frac{d^i}{dt^i}(\mu(s_t))=\mathcal{L}_\mu^i\mu(s_t),$$

then as long as $\left\|\mathcal{L}_\mu^i\mu(s_t)\right\|_{L^\infty}$ is bounded, one has

$$\left\|\hat{\mu}_i(s)-\mu(s)\right\|_{L^\infty}\leq\frac{\left\|\mathcal{L}_\mu^i\mu(s)\right\|_{L^\infty}}{(i+1)!}\sum_{j=0}^{i}|a_j^{(i)}|j^{i+1}\Delta t^i=C_i\left\|\mathcal{L}_\mu^i\mu(s)\right\|_{L^\infty}\Delta t^i.$$

6.3 Proof of Theorem 3.3

Proof. The first-order PhiBE solution satisfies

$$\beta\hat{V}_1(s) = r(s) + \frac{1}{\Delta t}(e^{\lambda\Delta t} - 1)s \cdot \nabla\hat{V}_1(s).$$

By setting $\hat{\lambda} = \frac{1}{\Delta t}(e^{\lambda\Delta t} - 1)$, one can write $\hat{V}_1(s)$ equivalently as

$$\hat{V}_1(s) = \int_0^\infty e^{-\beta t} r(e^{\hat{\lambda}t}s)dt,$$

which yields,

$$\begin{aligned}
\left|V(s) - \hat{V}_1(s)\right| &= \left|\int_0^\infty e^{-\beta t}\left(r(e^{\lambda t}s) - r(e^{\hat{\lambda}t}s)\right)dt\right| \\
&= \left|\int_0^\infty e^{-\beta t}\int_{\hat{\lambda}}^{\lambda} s \cdot \nabla r(e^{\tilde{\lambda}t}s)te^{\tilde{\lambda}t}d\tilde{\lambda}dt\right| \\
&= \left|\int_{\hat{\lambda}}^{\lambda}\left(\int_0^\infty e^{-\beta t}u \cdot \nabla r(u)tdt\right)d\tilde{\lambda}\right| \\
&\leq \|u \cdot \nabla r(u)\|_{L^\infty}\left|\int_{\hat{\lambda}}^{\lambda}\left(\int_0^\infty te^{-\beta t}dt\right)d\tilde{\lambda}\right| \\
&= \frac{1}{\beta^2}\|u \cdot \nabla r(u)\|_{L^\infty}|\lambda - \hat{\lambda}|,
\end{aligned}$$

where the second equality is obtained by applying the integral residual of Taylor expansion, and the third equality is obtained by setting $u = e^{\tilde{\lambda}t}s$. Since

$$|\lambda - \hat{\lambda}| = C_i\lambda^{i+1}\Delta t + o(\Delta t),$$

one has

$$\left|V(s) - \hat{V}_1(s)\right| \leq \frac{C_i\lambda^{i+1}\Delta t}{\beta^2}\|u \cdot \nabla r(u)\|_{L^\infty}\Delta t + o(\Delta t).$$

Next, we prove the upper bound for BE solution in the linear dynamics.

$$V(s)-\tilde{V}(s)=\sum_{i=0}^{\infty}\int_{\Delta ti}^{\Delta t(i+1)} e^{-\beta t}r(e^{\lambda t}s)-e^{-\beta\Delta ti}r(e^{\lambda\Delta ti}s)dt$$

$$=\sum_{i=0}^{\infty}\int_{\Delta ti}^{\Delta t(i+1)} \left(e^{-\beta t}-e^{-\beta\Delta ti}\right) r(e^{\lambda t}s)+\sum_{i=0}^{\infty}e^{-\beta\Delta ti}\int_{\Delta ti}^{\Delta t(i+1)}\left(r(e^{\lambda t}s)-r(e^{\lambda\Delta ti}s)\right)dt$$

$$\le \|r(s)\|_{L^\infty}\left(\int_0^\infty e^{-\beta t}dt-\sum_{i=0}^{\infty}e^{-\beta\Delta ti}\Delta t\right)+\sum_{i=0}^{\infty}e^{-\beta\Delta ti}\int_{\Delta ti}^{\Delta t(i+1)}\int_{\Delta ti}^{t} s\cdot\nabla r(e^{\lambda\tilde{t}}s)\lambda e^{\lambda\tilde{t}}d\tilde{t}dt,$$

where we use the integral residual of the Taylor expansion for the second term,

$$\le \|r(s)\|_{L^\infty}\left(\frac{1}{\beta}-\frac{\Delta t}{1-e^{-\beta\Delta t}}\right)+\|u\cdot\nabla r(u)\|_{L^\infty}\sum_{i=0}^{\infty}e^{-\beta\Delta ti}\left|\int_{\Delta ti}^{\Delta t(i+1)}\int_{\Delta ti}^{t}\lambda e^{\lambda\tilde{t}}e^{-\lambda\tilde{t}}d\tilde{t}dt\right|,$$

where we set $u=e^{\lambda\tilde{t}}s,\quad s=e^{-\lambda\tilde{t}}u$,

$$\le \|r(s)\|_{L^\infty}\left(\frac{1}{\beta}-\frac{\Delta t}{1-e^{-\beta\Delta t}}\right)+\frac{1}{2}|\lambda|\,\|u\cdot\nabla r(u)\|_{L^\infty}\frac{\Delta t^2}{1-e^{-\beta\Delta t}}.$$

Since

$$\lim_{\Delta t\to 0}\frac{\frac{1}{\beta}-\frac{\Delta t}{1-e^{-\beta\Delta t}}}{\frac{\Delta t}{2}}=1,\quad \frac{\frac{\Delta t}{1-e^{-\beta\Delta t}}}{\frac{1}{\beta}}=1,$$

which implies,

$$\left\|V(s)-\tilde{V}(s)\right\|_{L^\infty}\le\frac{1}{\beta}\left(\frac{\beta}{2}\|r(s)\|_{L^\infty}+\frac{|\lambda|}{2}\|u\cdot\nabla r(u)\|_{L^\infty}\right)\Delta t.$$

□

6.4 Proof of Theorem 3.4

We first present the property of the operator $\mathcal{L}_{\mu,\Sigma}$ that will be frequently used later in the following Proposition.

Proposition 6.3. *For the operator $\mathcal{L}_{\mu,\Sigma}$ defined in (4), under Assumption 1/(a), one has*

$$\langle\mathcal{L}_{\mu,\Sigma}V(s),V(s)\rangle_\rho\le-\frac{\lambda_{min}}{2}\|\nabla V\|_\rho^2;$$

$$\sum_i\langle\partial_{s_i}\mathcal{L}_{\mu,\Sigma}V(s),\partial_{s_i}V(s)\rangle_\rho\le C_{\nabla\mu,\nabla\Sigma}\|\nabla V\|_\rho^2;$$

$$\langle \mathcal{L}_{\mu,\Sigma} f(s), g(s)\rangle_\rho \le \left[\left(\|\mu\|_{L^\infty} + \frac{1}{2}\|\nabla\cdot\Sigma\|_{L^\infty}\right)\|g\|_\rho + \frac{1}{2}\|\Sigma\|_{L^\infty}\|\nabla g\|_\rho\right]\|\nabla f\|_\rho$$
$$+ \frac{1}{2}\|\Sigma\|_{L^\infty}\|\nabla\log\rho\|_\rho\|g\|_\rho\|\nabla f\|_{L^\infty};$$
$$\langle \mathcal{L}_{\mu,\Sigma} f(s), g(s)\rangle_\rho \le \left[\left(\|\mu\|_{L^\infty} + \frac{1}{2}\|\nabla\cdot\Sigma\|_{L^\infty} + \frac{1}{2}\|\Sigma\|_{L^\infty}\|\nabla\log\rho\|_{L^\infty}\right)\|g\|_\rho\right.$$
$$\left.+\frac{1}{2}\|\Sigma\|_{L^\infty}\|\nabla g\|_\rho\right]\|\nabla f\|_\rho,$$

where $C_{\nabla\mu,\nabla\Sigma}$ is defined in (39) *depending on the first derivatives of* μ, Σ.

Proof. Inserting the operator $\mathcal{L}_{\mu,\Sigma}$, and applying integral by parts gives,

$$\langle \mathcal{L}_{\mu,\Sigma} V(s), V(s)\rangle_\rho = \langle \mu\cdot\nabla V, V\rangle_\rho - \frac{1}{2}\sum_{i,j}\left\langle \partial_{s_j}(\Sigma_{ij}V\rho), \partial_{s_i}V\right\rangle$$
$$=\sum_i\left\langle \mu_i\rho, \partial_{s_i}\left(\frac{1}{2}V^2\right)\right\rangle - \frac{1}{2}\sum_{i,j}\left\langle \partial_{s_j}(\Sigma_{ij}\rho), \partial_{s_i}\left(\frac{1}{2}V^2\right)\right\rangle - \frac{1}{2}\sum_{i,j}\left\langle(\partial_{s_j}V)\Sigma_{ij}, \partial_{s_i}V\right\rangle_\rho$$
$$=-\sum_i\left\langle \partial_{s_i}(\mu_i\rho), \frac{1}{2}V^2\right\rangle + \frac{1}{2}\sum_{i,j}\left\langle \partial_{s_i}\partial_{s_j}(\Sigma_{ij}\rho), \frac{1}{2}V^2\right\rangle - \frac{1}{2}\int(\nabla V)^\top\Sigma(\nabla V)\rho\, ds$$
$$=\left\langle \nabla\cdot\left(-\mu\rho + \frac{1}{2}\nabla\cdot(\Sigma\rho)\right), \frac{1}{2}V^2\right\rangle - \frac{1}{2}\int(\nabla V)^\top\Sigma(\nabla V)\rho\, ds$$
$$\le -\frac{\lambda_{\min}}{2}\|\nabla V\|_\rho^2,$$

where the last inequality is because of the definition of the stationary solution (19) and the positivity of the matrix $\Sigma(s)$ in Assumption 1.

For the second part of the Lemma, first note that

$$\partial_{s_i}\mathcal{L}_{\mu,\Sigma}V = \partial_{s_i}\mu\cdot\nabla V + \frac{1}{2}\partial_{s_i}\Sigma : \nabla^2 V + \mathcal{L}_{\mu,\Sigma}\partial_{s_i}V.$$

Therefore, applying the first part of the Lemma gives

$$\sum_i\langle \partial_{s_i}\mathcal{L}_{\mu,\Sigma}V(s), V(s)\rangle_\rho$$
$$\le\sum_i\left(\langle \partial_{s_i}\mu\cdot\nabla V, \partial_{s_i}V\rangle_\rho + \frac{1}{2}\left\langle \partial_{s_i}\Sigma : \nabla^2 V, \partial_{s_i}V\right\rangle_\rho\right) - \frac{\lambda_{\min}}{2}\sum_i\|\nabla\partial_{s_i}V\|_\rho^2$$
$$\le\|\nabla\mu\|_{L^\infty}\|\nabla V\|_\rho^2 + \frac{1}{2}\|\nabla\Sigma\|_{L^\infty}\left\|\nabla^2 V\right\|_\rho\|\nabla V\|_\rho - \frac{\lambda_{\min}}{2}\left\|\nabla^2 V\right\|_\rho^2$$
$$=-\frac{\lambda_{\min}}{2}\left(\left\|\nabla^2 V\right\|_\rho - \frac{\|\nabla\Sigma\|_{L^\infty}}{2\lambda_{\min}}\|\nabla V\|_\rho\right)^2 + \left(\frac{\|\nabla\Sigma\|_{L^\infty}^2}{8\lambda_{\min}} + \|\nabla\mu\|_{L^\infty}\right)\|\nabla V\|_\rho^2$$
$$\le\frac{C_{\nabla\mu,\nabla\Sigma}}{2}\|\nabla V\|_\rho^2,$$

where

$$C_{\nabla\mu,\nabla\Sigma} = \frac{\|\nabla\Sigma\|_{L^\infty}^2}{4\lambda_{\min}} + 2\|\nabla\mu\|_{L^\infty}. \tag{39}$$

For the last two inequalities, one notes

$$\begin{aligned}
&\langle \mathcal{L}_{\mu,\Sigma} f, g\rangle_\rho \\
=&\langle \mu\cdot\nabla f, g\rangle_\rho - \frac{1}{2}\left[\langle \nabla f\cdot\nabla\cdot\Sigma, g\rangle_\rho + \langle \nabla f\Sigma, \nabla g\rangle_\rho + \left\langle \nabla f\Sigma, \frac{\nabla\rho}{\rho} g\right\rangle_\rho\right] \\
\le& \left[\left(\|\mu\|_{L^\infty} + \frac{1}{2}\|\nabla\cdot\Sigma\|_{L^\infty}\right)\|g\|_\rho + \frac{\|\Sigma\|_{L^\infty}}{2}\|\nabla g\|_\rho\right]\|\nabla f\|_\rho + \left\langle \nabla f\frac{\Sigma}{2}, g\nabla\log\rho\right\rangle_\rho.
\end{aligned}$$

By bounding the last term differently,

$$\frac{1}{2}\|\nabla\log\rho\|_{L^\infty}\|\Sigma\|_{L^\infty}\|g\|_\rho\|\nabla f\|_\rho \quad \text{or,} \quad \frac{1}{2}\|\nabla\log\rho\|_\rho\|\Sigma\|_{L^\infty}\|g\|_\rho\|\nabla f\|_{L^\infty},$$

one ends up with the last two inequalities of the Lemma.

□

Proof of Theorem 3.4 Now we are ready to prove Theorem 3.4.

Proof. By (33), one has

$$\begin{aligned}
&\left\|V(s) - \tilde{V}(s)\right\|_\rho \\
\le& \sum_{i=0}^{\infty}\sqrt{\Delta t\int_{\Delta ti}^{\Delta t(i+1)}\left\|\int_{\mathbb{S}} f(t,s')\rho(s',t|s) - f(\Delta ti, s')\rho(\Delta ti, s')ds'\right\|_\rho^2 dt},
\end{aligned} \tag{40}$$

where the Jensen's inequality is used. By (34), one has for $t \in [\Delta ti, \Delta t(i+1)]$

$$\begin{aligned}
&\left\|\int_{\mathbb{S}} f(t,s')\rho(s',t|s) - f(\Delta ti, s')\rho(\Delta ti, s')ds'\right\|_\rho \\
\le& e^{-\beta\Delta ti}(t-\Delta ti)\left(\|p_1(\xi_1, s)\|_\rho + \beta\|p_2(\Delta ti, s)\|_\rho\right),
\end{aligned}$$

where

$$p_1(s,t) = \int_{\mathbb{S}} \mathcal{L}_{\mu,\Sigma} r(s')\rho(s',t|s)ds', \quad p_2(s,t) = \int_{\mathbb{S}} r(s')\rho(s',t|s)ds'.$$

Note that both $p_1(s,t)$ and $p_2(s,t)$ satisfies

$$\partial_t p_i(s,t) = \mathcal{L}_{\mu,\Sigma} p_i(s,t), \quad \text{with initial data } p_1(0,s) = \mathcal{L}_{\mu,\Sigma} r(s), \quad p_2(0,s) = r(s).$$

By Proposition 6.3, one has

$$\frac{1}{2}\|p_i(t)\|_\rho^2 \le -\frac{\lambda_{\min}}{2}\|\nabla p_i(t)\|_\rho^2 \le 0,$$

which implies,

$$\|p_i(t)\|_\rho \le \|p_i(0)\|_\rho .$$

Therefore, one has

$$\begin{aligned}&\left\|\int_{\mathbb{S}} f(t,s')\rho(s',t|s) - f(\Delta ti, s')\rho(\Delta ti, s')ds'\right\|_\rho \\ \le & e^{-\beta\Delta ti}(t-\Delta ti)\left(\|\mathcal{L}_{\mu,\Sigma} r(s)\|_\rho + \beta\|r(s)\|_\rho\right).\end{aligned}$$

Inserting it back to (40) yields,

$$\begin{aligned}&\left\|V(s)-\tilde{V}(s)\right\|_\rho \\ \le & \left(\|\mathcal{L}_{\mu,\Sigma} r(s)\|_\rho + \beta\|r(s)\|_\rho\right)\sum_{i=0}^{\infty}\sqrt{\Delta t e^{-2\beta\Delta ti}\int_{\Delta ti}^{\Delta t(i+1)}(t-\Delta ti)^2 dt} \\ = & \left(\|\mathcal{L}_{\mu,\Sigma} r(s)\|_\rho + \beta\|r(s)\|_\rho\right)\frac{1}{\sqrt{3}}\Delta t^2\sum_{i=0}^{\infty} e^{-\beta\Delta ti} \\ = & \frac{1}{\sqrt{3}\beta}\left(\|\mathcal{L}_{\mu,\Sigma} r(s)\|_\rho + \beta\|r(s)\|_\rho\right) + o(\Delta t),\end{aligned}$$

which completes the proof. □

6.5 Proof of Theorem 3.5

First note that $V, \hat{V}_i$ satisfies,

$$\mathcal{L}_{\mu,\Sigma} V = \beta V - r, \quad \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i}\hat{V}_i = \beta\hat{V}_i - r.$$

By the following Lemma, one can bound $\left\|V-\hat{V}_i\right\|_\rho$ by the distance between μ, Σ and $\hat{\mu}_i, \hat{\Sigma}_i$.

Lemma 6.4. *For $V, \hat{V}$ satisfying*

$$\mathcal{L}_{\mu,\Sigma} V = \beta V - r, \quad \mathcal{L}_{\hat{\mu},\hat{\Sigma}}\hat{V} = \beta\hat{V} - r,$$

under Assumption 1/(a), if $\|\hat{\mu}-\mu\|_{L^\infty} \le C_\mu$, $\left\|\hat{\Sigma}-\Sigma\right\|_{L^\infty} \le C_\Sigma$, $\left\|\nabla\cdot(\hat{\Sigma}-\Sigma)\right\|_{L^\infty} \le C_{\nabla\cdot\Sigma}$, $\|\nabla\log\rho\|_\rho \le L_\rho$, and $C_\mu + \frac{1}{2}C_\Sigma \le \sqrt{\frac{\beta\lambda_{min}}{2}}, C_\Sigma \le \frac{\lambda_{min}}{2}$, one has

$$\left\|V-\hat{V}\right\|_\rho \le \left[\frac{2C_\mu + C_{\nabla\cdot\Sigma}}{\beta}\left(1+\frac{C_\Sigma}{\lambda_{min}}\right) + \frac{C_\Sigma}{\sqrt{\beta\lambda_{min}}}\right]\|\nabla V\|_\rho + \frac{2C_\Sigma L_\rho}{\beta}\left\|\nabla\hat{V}\right\|_{L^\infty}.$$

(See Section 6.5.1 for the proof of the above lemma) Then we further apply the following lemma regarding the distance between μ, Σ and $\hat{\mu}_i, \hat{\Sigma}_i$.

Lemma 6.5. *Under Assumption 1, for $\hat{\mu}(s), \hat{\Sigma}(s)$ defined in* (18)*, one has*

$$\|\hat{\mu}_i(s) - \mu(s)\|_{L^\infty} \le L_\mu \Delta t^i, \quad \left\|\hat{\Sigma}_i(s)_{kl} - \Sigma(s)_{kl}\right\|_{L^\infty} \le L_\Sigma \Delta t^i + o(\Delta t^i),$$

and

$$\left\|\nabla \cdot (\hat{\Sigma} - \Sigma)\right\|_{L^\infty}^2 \le L_{\nabla\cdot\Sigma} \Delta t^i + o(\Delta t^i), \tag{41}$$

where $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}$ are constants depending on μ, Σ, i defined in (46), (49), (51), *respectively.*

(See Section 6.5.2 for the proof of the above lemma) Combine the above two lemmas, one can bound

$$\left\|V - \hat{V}_i\right\| \le \left[\left(\frac{2L_\mu + L_{\nabla\cdot\Sigma}}{\beta}\left(1 + \frac{L_\Sigma \Delta t^i}{\lambda_{\min}}\right) + \frac{L_\Sigma}{\sqrt{\beta\lambda_{\min}}}\right)\|\nabla V\|_\rho + \frac{2L_\Sigma L_\rho}{\beta}\left\|\nabla \hat{V}\right\|_{L^\infty}\right]\Delta t^i + o(\Delta t^i)$$

for

$$\Delta t^i \le D_{\mu,\Sigma,\beta}, \quad D_{\mu,\Sigma,\beta} = \min\left\{\frac{\lambda_{\min}}{2L_\Sigma}, \frac{\sqrt{2\beta\lambda_{\min}}}{2L_\mu + L_{\nabla\cdot\Sigma}}\right\}, \tag{42}$$

with $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}$ defined in (46), (49), (57). Furthermore, by the following lemma on the upper bound for $\|\nabla V\|_\rho, \left\|\nabla \hat{V}\right\|_{L^\infty}$,

Lemma 6.6. *Under Assumption 1/(a), for $V(s)$ satisfying* (16)*, one has*

$$\|\nabla V(s)\|_\rho \le \frac{1}{\beta}\left(\sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{min}}}\|r\|_\rho + \|\nabla r\|_\rho\right)$$

$$\|\nabla V(s)\|_{L^\infty} \le \frac{1}{\beta}\left(\sqrt{\frac{2C_{\nabla\mu,\nabla\Sigma}}{\lambda_{min}}}\|r\|_{L^\infty} + \|\nabla r\|_{L^\infty}\right) + o(\Delta t^{i/2})$$

where $C_{\nabla\mu,\nabla\Sigma}$ is a constant defined in (39) *that depends on $\nabla\mu(s), \nabla\Sigma(s)$.*

(See Section 6.5.3 for the proof of the above lemma) one has,

$$\begin{aligned}
\left\|V - \hat{V}_i\right\|_\rho \le& \left[\left(\frac{2L_\mu + L_{\nabla\cdot\Sigma}}{\beta^2} + \frac{L_\Sigma}{\beta\sqrt{\beta\lambda_{\min}}}\right)\left(\sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}}\|r\|_\rho + \|\nabla r\|_\rho\right)\right.\\
&\left.+\frac{2L_\Sigma L_\rho}{\beta^2}\left(\sqrt{\frac{2C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}}\|r\|_{L^\infty} + \|\nabla r\|_{L^\infty}\right)\right]\Delta t^i + o(\Delta t^i)\\
\le& \frac{C_{r,\mu,\Sigma}}{\beta^2} + \frac{\hat{C}_{r,\mu,\Sigma}}{\beta^{3/2}},
\end{aligned}$$

where

$$
\begin{aligned}
C_{r,\mu,\Sigma} =& (2L_\mu + L_{\nabla\cdot\Sigma}) \left(\sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}} \|r\|_\rho + \|\nabla r\|_\rho \right) \\
&+ 2L_\Sigma L_\rho \left(\sqrt{\frac{2C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}} \|r\|_{L^\infty} + \|\nabla r\|_{L^\infty} \right), \qquad (43) \\
\hat{C}_{r,\mu,\Sigma} =& \frac{L_\Sigma}{\sqrt{\lambda_{\min}}} \left(\sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}} \|r\|_\rho + \|\nabla r\|_\rho \right),
\end{aligned}
$$

with $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}, L_\rho, C_{\nabla\mu,\nabla\Sigma}$ defined in (46), (49), (57), (20), (39).

6.5.1 Proof of Lemma 6.4

Substracting the second equation from the first one and let $e(s) = V(s) - \hat{V}(s)$ gives,

$$
\beta e = \mathcal{L}_{\mu,\Sigma} e + (\mathcal{L}_{\mu,\Sigma} - \mathcal{L}_{\hat{\mu},\hat{\Sigma}})\hat{V}.
$$

Multiply the above equation with $e(s)\rho(s)$ and integrate it over $s \in \mathbb{S}$, one has,

$$
\begin{aligned}
&\beta \|e\|_\rho^2 = \langle \mathcal{L}_{\mu,\Sigma} e, e \rangle_\rho + \left\langle \mathcal{L}_{\hat{\mu}-\mu,\hat{\Sigma}-\Sigma} \hat{V}, e \right\rangle_\rho \\
\le& -\frac{\lambda_{\min}}{2} \|\nabla e\|_\rho^2 + \left(C_\mu + \frac{1}{2} C_{\nabla\cdot\Sigma} \right) \|e\|_\rho \left\| \nabla \hat{V} \right\|_\rho + \frac{1}{2} C_\Sigma \|\nabla e\|_\rho \left\| \nabla \hat{V} \right\|_\rho \\
&+ \frac{1}{2} C_\Sigma L_\rho \|e\|_\rho \left\| \nabla \hat{V} \right\|_{L^\infty} \\
\le& -\left(\frac{\lambda_{\min}}{2} - c_2 \right) \|\nabla e\|_\rho^2 + \Big(c_1 \|e\|_\rho + c_2 \|\nabla V\|_\rho \Big) \|\nabla e\|_\rho + \left(c_1 \|\nabla V\|_\rho + c_3 \left\| \nabla \hat{V} \right\|_{L^\infty} \right) \|e\|_\rho,
\end{aligned} \tag{44}
$$

where the first and third equations in Proposition 6.3 are used for the first inequality, $\left\| \nabla \hat{V} \right\|_\rho \le \|\nabla V\|_\rho + \|\nabla e\|_\rho$ are used for the second inequality, and $c_1 = C_\mu + \frac{1}{2} C_{\nabla\cdot\Sigma}$, $c_2 = \frac{1}{2} C_\Sigma$, $c_3 = \frac{1}{2} C_\Sigma L_\rho$. Under the assumption that $c_2 \le \frac{\lambda_{\min}}{4}$, one has

$$
\begin{aligned}
\beta \|e\|_\rho^2 \le& -\frac{\lambda_{\min}}{4} \left(\|\nabla e\|_\rho^2 - \frac{2}{\lambda_{\min}} (c_1 \|e\|_\rho + c_2 \|\nabla V\|_\rho) \right)^2 \\
&+ \frac{1}{\lambda_{\min}} (c_1 \|e\|_\rho + c_2 \|\nabla V\|_\rho)^2 + \left(c_1 \|\nabla V\|_\rho + c_3 \left\| \nabla \hat{V} \right\|_{L^\infty} \right) \|e\|_\rho \\
\le& \frac{c_1^2}{\lambda_{\min}} \|e\|_\rho^2 + \left[\left(\frac{2c_1c_2}{\lambda_{\min}} + c_1 \right) \|\nabla V\|_\rho + c_3 \left\| \nabla \hat{V} \right\|_{L^\infty} \right] \|e\|_\rho + \frac{c_2^2}{\lambda_{\min}} \|\nabla V\|_\rho^2.
\end{aligned}
$$

Under the assumption that $c_1^2 \le \frac{1}{2} \beta \lambda_{\min}$, one has

$$
\frac{\beta}{2} \|e\|_\rho^2 \le \left[\frac{2}{\beta} \left(\frac{2c_1c_2}{\lambda_{\min}} + c_1 \right)^2 + \frac{c_2^2}{\lambda_{\min}} \right] \|\nabla V\|_\rho^2 + \frac{2c_3^2}{\beta} \|\nabla V\|_{L^\infty}^2 + \frac{\beta}{4} \|e\|_\rho^2,
$$

which yields,

$$\|e\|_\rho \leq \left[\frac{2c_1}{\beta}\left(\frac{2c_2}{\lambda_{\min}}+1\right)+\frac{2c_2}{\sqrt{\beta\lambda_{\min}}}\right]\|\nabla V\|_\rho + \frac{3c_3}{\beta}\left\|\nabla \hat{V}\right\|_{L^\infty}.$$

6.5.2 Proof of Lemma 6.5

The proof of Lemma 6.5 replies the following two lemmas, which we will prove later.

Lemma 6.7. *Define operator* $\Pi_{i,\Delta t}f(s) = \frac{1}{\Delta t}\mathbb{E}[\sum_{j=1}^{i} a_j^{(i)} f(s_{j\Delta t} - s_0)|s_0 = s]$ *with* $a_j^{(i)}$ *defined in* (14) *and* $f(0) = 0$*, then*

$$\Pi_{i,\Delta t}f(s) = \mathcal{L}_{\mu,\Sigma}f(0) + \frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j^{(i)} \int_0^{j\Delta t} \mathbb{E}[\mathcal{L}_{\mu,\Sigma}^{i+1} f(s_t - s_0)|s_0 = s]t^i dt.$$

Lemma 6.8. *For* $p(s,t) = \mathbb{E}[f(s_t)|s_0 = s]$ *with* s_t *driven by the SDE* (2)*, then under Assumption 1/(a), one has*

$$\|\nabla p(s,t)\|_{L^\infty} \leq \sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{min}}}\|f\|_{L^\infty} + \|\nabla f\|_{L^\infty}, \quad \text{with } C_{\nabla\mu,\nabla\Sigma} \text{ defined in (39).}$$

For $p(s,t) = \mathbb{E}[f(s_t)(s_t - s_0)|s_0 = s]$ *with* s_t *driven by the SDE* (2)*, then under Assumption 1/(a), one has*

$$\|p(s,t)\|_{L^\infty} \leq \|\mu\|_{L^\infty}\sqrt{e^t - 1}.$$

$$\|\nabla p(s,t)\|_{L^\infty} \leq \left(\sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{min}}}\|\mu(s)\|_{L^\infty} + \|\nabla\mu(s)\|_{L^\infty}\right)\sqrt{e^t - 1}.$$

Now we are ready to prove Lemma 6.5. By Lemma 6.7, one has

$$\hat{\mu}_i(s) = \mu(s) + \frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j^{(i)}\left(\int_0^{j\Delta t}\mathbb{E}[\mathcal{L}_{\mu,\Sigma}^i \mu(s_t)|s_0 = s]t^i dt\right), \tag{45}$$

which implies that

$$\|\hat{\mu}_i(s) - \mu(s)\|_{L^\infty} \leq \frac{\left\|\mathcal{L}_{\mu,\Sigma}^i\mu(s)\right\|_{L^\infty}}{\Delta t i!}\sum_{j=1}^{i}|a_j^{(i)}|\int_0^{j\Delta t} t^i dt \leq L_\mu \Delta t^i,$$

where

$$L_\mu = \hat{C}_i\left\|\mathcal{L}_{\mu,\Sigma}^i\mu(s)\right\|_{L^\infty} \quad \text{with } \hat{C}_i = \sum_{j=1}^{i}\frac{|a_j^{(i)}|j^{i+1}}{(i+1)!} \text{ defined in (36).} \tag{46}$$

To prove the second inequality in the lemma, first apply Lemma 6.7, one has

$$\begin{aligned}\hat{\Sigma}_i(s) =&\Sigma(s)\\ &+\frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j^{(i)}\int_0^{j\Delta t}\mathbb{E}[\mathcal{L}^i_{\mu,\Sigma}\left(\mu(s_t)(s_t-s_0)^\top+(s_t-s_0)\mu^\top(s_t)+\Sigma(s_t)\right)|s_0=s]t^i dt.\end{aligned} \tag{47}$$

Note that

$$\begin{aligned}h(s_t):=&\mathcal{L}^i_{\mu,\Sigma}\left(\mu(s_t)(s_t-s_0)^\top+(s_t-s_0)\mu^\top(s_t)+\Sigma(s_t)\right)\\ &-\left[\mathcal{L}^i_{\mu,\Sigma}\mu(s_t)(s_t-s_0)^\top+(s_t-s_0)(\mathcal{L}^i_{\mu,\Sigma}\mu(s_t))^\top\right]\end{aligned} \tag{48}$$

is a function that only depends on the derivative $\nabla^j\Sigma, \nabla^j\mu$ up to $2i$-th order, which can be bounded under Assumption 1/(b). Thus applying the second inequality of Lemma 6.8 yields

$$\begin{aligned}&\left|\mathbb{E}[\mathcal{L}^i_{\mu,\Sigma}\left(\mu(s_t)(s_t-s_0)^\top+(s_t-s_0)\mu^\top(s_t)+\Sigma(s_t)\right)|s_0=s]\right|\\ \le&\left\|h(s)\right\|_{L^\infty}+2\left\|\mu(s)\right\|_{L^\infty}e^{t/2}.\end{aligned}$$

Hence, one has,

$$\begin{aligned}&\int_0^{j\Delta t}\mathbb{E}[\mathcal{L}^i_{\mu,\Sigma}\left(\mu(s_t)(s_t-s_0)^\top+(s_t-s_0)\mu^\top(s_t)+\Sigma(s_t)\right)|s_0=s]t^i dt\\ \le&\frac{1}{i+1}\left\|h\right\|_{L^\infty}(j\Delta t)^{i+1}+2\left\|\mu(s)\right\|_{L^\infty}\int_0^{j\Delta t}(2+t)t^i dt,\quad \text{for } j\Delta t\le 3\\ =&\frac{1}{i+1}\left(\left\|h\right\|_{L^\infty}+4\left\|\mu\right\|_{L^\infty}\right)(j\Delta t)^{i+1}+2\left\|\mu(s)\right\|_{L^\infty}\frac{1}{i+2}(j\Delta t)^{i+2},\end{aligned}$$

where $e^{t/2}\le 2+t$ for $t\le 3$ are used in the first inequality. Plugging the above inequality back to (47) implies

$$\left\|\hat{\Sigma}_i(s)-\Sigma(s)\right\|_{L^\infty}\le L_\Sigma\Delta t^i+o(\Delta t^i),$$

where

$$L_\Sigma=\hat{C}_i(\left\|h(s)\right\|_{L^\infty}+4\left\|\mu\right\|_{L^\infty})\quad\text{with } \hat{C}_i, h(s) \text{ defined in (46), (48).} \tag{49}$$

To prove the third inequality, one first takes $\nabla\cdot$ to (47),

$$\nabla\cdot\hat{\Sigma}_i(s)=\nabla\cdot\Sigma(s)+\frac{1}{\Delta t i!}\sum_{j=1}^{i}a_j^{(i)}\int_0^{j\Delta t}\nabla\cdot p(s,t)t^i dt,$$

where

$$p(s,t)=\mathbb{E}[h(s_t)|s_0=s]+\mathbb{E}[\mathcal{L}^i_{\mu,\Sigma}\mu(s_t)(s_t-s_0)^\top+(s_t-s_0)(\mathcal{L}^i_{\mu,\Sigma}\mu(s_t))^\top|s_0=s],$$

with h defined in (48). Therefore, by the first and third inequalities in Lemma 6.8, and denoting $c_1 = \sqrt{C_{\nabla\mu,\nabla\Sigma}/\lambda_{\min}}\, \|h\|_{L^\infty} + \|\nabla\cdot h\|_{L^\infty}$, $c_2 = \sqrt{C_{\nabla\mu,\nabla\Sigma}/\lambda_{\min}}\, \|\mu\|_{L^\infty} + \|\nabla\cdot\mu\|_{L^\infty}$, one has

$$
\begin{aligned}
&\sum_l \left\|\nabla\cdot\hat{\Sigma}_i(s) - \nabla\cdot\Sigma(s)\right\|_{L^\infty} \le \frac{1}{\Delta t i!}\sum_{j=1}^{i}|a_j^{(i)}|\left(\int_0^{j\Delta t} c_1 t^i + 2c_2 e^{t/2} t^i dt\right) \\
\le&\frac{1}{\Delta t i!}\sum_{j=1}^{i}|a_j^{(i)}|\left(\int_0^{j\Delta t}(c_1+4c_2)t^i + 2c_2 t^{i+1} dt\right) = L_{\nabla\cdot\Sigma}\Delta t^i + o(\Delta t^i),
\end{aligned}
\tag{50}
$$

where

$$
\begin{aligned}
L_{\nabla\cdot\Sigma} =&\hat{C}_i\left[\sqrt{C_{\nabla\mu,\nabla\Sigma}/\lambda_{\min}}\,\|h\|_{L^\infty} + \|\nabla\cdot h\|_{L^\infty}\right. \\
&\left.+4\left(\sqrt{C_{\nabla\mu,\nabla\Sigma}/\lambda_{\min}}\,\|\mu\|_{L^\infty} + \|\nabla\cdot\mu\|_{L^\infty}\right)\right],
\end{aligned}
\tag{51}
$$

with $\hat{C}_i, h(s), C_{\nabla\mu,\nabla\Sigma}$ defined in (46), (48), (39).

Proof of Lemma 6.7

Proof. First note that

$$
\Pi_{i,\Delta t} f(s) = \frac{1}{\Delta t}\sum_{j=1}^{i} a_j^{(i)}\int_{\mathbb{S}} f(s'-s)\rho(s', j\Delta t|s)ds', \tag{52}
$$

where $\rho(s',t|s)$ is defined in (32). By Taylor's expansion, one has

$$
\rho(s', j\Delta t|s) = \sum_{k=0}^{i}\partial_t^k\rho(s',0|s)\frac{(j\Delta t)^k}{k!} + \frac{1}{i!}\int_0^{j\Delta t}\partial_t^{i+1}\rho(s',t|s)t^i dt.
$$

Inserting the above equation into (52) yields,

$$
\begin{aligned}
\Pi_{i,\Delta t} f(s) =&\underbrace{\frac{1}{\Delta t}\sum_{k=0}^{i}\left(\sum_{j=1}^{i} a_j^{(i)} j^k\right)\frac{(\Delta t)^k}{k!}\int_{\mathbb{S}} f(s'-s)\partial_t^k\rho(s',0|s)ds'}_{I} \\
&+\underbrace{\frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j^{(i)}\left(\int_{\mathbb{S}}\int_0^{j\Delta t} f(s'-s)\partial_t^{i+1}\rho(s',t|s)t^i dt ds'\right)}_{II}.
\end{aligned}
$$

By the definition of $a_j^{(i)}$, the first part can simplified to

$$
\begin{aligned}
I =&\frac{1}{\Delta t}\left(\sum_{j=1}^{i} a_j^{(i)}\right)\int_{\mathbb{S}} f(s'-s)\rho(s',0|s)ds' + \int_{\mathbb{S}} f(s'-s)\partial_t\rho(s',0|s)ds' \\
=&\frac{\sum_{j=1}^{i} a_j^{(i)}}{\Delta t} f(0) + \int_{\mathbb{S}}\mathcal{L}_{\mu,\Sigma} f(s'-s)\rho(s',0|s)ds' = \mathcal{L}_{\mu,\Sigma} f(0).
\end{aligned}
$$

Apply integration by parts, the second part can be written as

$$II = \frac{1}{\Delta t i!} \sum_{j=1}^{i} a_j^{(i)} \int_0^{j\Delta t} \mathbb{E}[\mathcal{L}_{\mu,\Sigma}^{i+1} f(s_t - s_0)|s_0 = s] t^i dt,$$

which completes the proof. □

Proof of Lemma 6.8

Proof. Note that $p(s,t)$ satisfies the following forward Kolmogorov equation [17],

$$\partial_t p(s,t) = \mathcal{L}_{\mu,\Sigma} p(s,t), \quad \text{with } p(s,0) = f(s).$$

Multiplying p to the above equation and let $q_l = \partial_{s_l} p$, with Assumption 1/(a), one has

$$\partial_t(\frac{1}{2}p^2) = \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}p^2) - \frac{1}{2} q^\top \Sigma q, \quad \partial_t(\frac{1}{2}p^2) \le \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}p^2) - \frac{\lambda_{\min}}{2} \|q\|_2^2. \tag{53}$$

On the other hand, q_l satisfies

$$\partial_t q_l = \mathcal{L}_{\mu,\Sigma} q_l + \mathcal{L}_{\partial_{s_l}\mu, \partial_{s_l}\Sigma} p, \quad \text{with } q(0,s) = \partial_{s_l} f(s).$$

Multiplying q_l to the above equation and then summing it over l gives,

$$\begin{aligned}
\partial_t(\frac{1}{2}\|q\|_2^2) =& \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|q\|_2^2) - \frac{1}{2}\sum_l (\nabla q_l)^\top \Sigma (\nabla q_l) + q^\top \nabla\mu \cdot q + \sum_l \frac{1}{2}(\partial_{s_l}\Sigma : \nabla q) q_l, \\
\partial_t(\frac{1}{2}\|q\|_2^2) \le& \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|q\|_2^2) - \frac{\lambda_{\min}}{2}\|\nabla q\|_2^2 + \|\nabla\mu\|_2 \|q\|_2^2 + \frac{1}{2}\|\nabla\Sigma\|_2 \|\nabla q\|_2 \|q\|_2 \\
\le& \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|q\|_2^2) + \left(\|\nabla\mu\|_{L^\infty} + \frac{\|\nabla\Sigma\|_{L^\infty}^2}{8\lambda_{\min}} \right) \|q\|_2^2.
\end{aligned}$$

Adding the above inequality to $c\times$(53) with $c = \frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}$ and $C_{\nabla\mu,\nabla\Sigma}$ defined in (39), one has

$$\partial_t \left(\frac{1}{2}p^2 + \frac{c}{2}\|q\|_2^2 \right) \le \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}p^2 + \frac{c}{2}\|q\|_2^2).$$

Let $g(s,t) = \frac{1}{2}p^2 + \frac{c}{2}\|q\|_2^2$, then $\partial_t g \le \mathcal{L}_{\mu,\Sigma} g$. Let $g_1(s,t) = \mathbb{E}[cf(s_t)^2/2 + \|\nabla f(s_t)\|_2^2/2 | s_0 = s]$, then $g_1(s,t)$ satisfying $\partial_t g_1 = \mathcal{L}_{\mu,\Sigma} g_1$, with $g_1(0,s) = g(0,s)$. Since $\|g_1(t,s)\|_{L^\infty} \le \frac{c}{2}\|f\|_{L^\infty}^2 + \frac{1}{2}\|\nabla f\|_{L^\infty}$, by comparison theorem, one has for $\forall s \in \mathbb{S}$

$$g(t,s) \le g_1(t,s) \le \frac{1}{2}(c\|f\|_{L^\infty}^2 + \|\nabla f\|_{L^\infty}^2),$$

which completes the proof for the first inequality.

For the second $p(t,s) = \mathbb{E}[f(s_t)(s_t - s_0)|s_0 = s]$, first note that it satisfies the following PDE,

$$\partial_t p = \mathcal{L}_{\mu,\Sigma} p + \mu(s), \quad \text{with } p(0,s) = 0.$$

Multiplying it with $p^\top$ and letting $q = \nabla p$ gives,

$$\partial_t(\frac{1}{2}\|p\|_2^2) \leq \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|p\|_2^2) - \frac{\lambda_{\min}}{2}\|q\|_2^2 + \frac{1}{2a}\|\mu\|_{L^\infty}^2 + \frac{a}{2}\|p\|_2^2, \quad \text{for } \forall a > 0. \tag{54}$$

Let $g(t,s) = \frac{1}{2}\|p\|_2^2 e^{-at} + \frac{1}{2a^2}\|\mu\|_{L^\infty}^2 e^{-at}$, then,

$$\partial_t g \leq \mathcal{L}_{\mu,\Sigma} g, \quad \text{with } g(0,s) = \frac{\|\mu\|_{L^\infty}^2}{2a^2}. \tag{55}$$

Similarly, by comparison theorem, $\|g(t,s)\|_{L^\infty} \leq \|g(0,s)\|_{L^\infty} = \frac{\|\mu\|_{L^\infty}^2}{2a^2}$, which implies,

$$\|p(t,\cdot)\|_{L^\infty} \leq \frac{\|\mu\|_{L^\infty}}{a}\sqrt{e^{at}-1} = \|\mu\|_{L^\infty}\sqrt{e^t - 1}.$$

where the last equality is obtained by selecting $a = 1$. On the other hand, taking ∇ to (55) and multiply $q^\top$ to it gives,

$$\begin{aligned}
\partial_t(\frac{1}{2}\|q\|_2^2) =& \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|q\|_2^2) - \frac{1}{2}\sum_l (\nabla q_l)^\top \Sigma (\nabla q_l) + q^\top \nabla\mu \cdot q + \sum_l \frac{1}{2}(\partial_{s_l}\Sigma : \nabla q) q_l + q \cdot \nabla \mu, \\
\partial_t(\frac{1}{2}\|q\|_2^2) \leq& \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|q\|_2^2) - \frac{\lambda_{\min}}{2}\|\nabla q\|_2^2 + \|\nabla\mu\|_2\|q\|_2^2 + \frac{1}{2}\|\nabla\Sigma\|_2\|\nabla q\|_2\|q\|_2 + \frac{1}{2a}\|\nabla\mu\|_{L^\infty}^2 \\
&+ \frac{a}{2}\|q\|_2^2 \\
\leq& \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|q\|_2^2) + \left(\|\nabla\mu\|_{L^\infty} + \frac{\|\nabla\Sigma\|_{L^\infty}^2}{8\lambda_{\min}}\right)\|q\|_2^2 + \frac{1}{2a}\|\nabla\mu\|_{L^\infty}^2 + \frac{a}{2}\|q\|_2^2.
\end{aligned}$$

Adding the above inequality to $c\times$(54) with $c = \frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}$ and $C_{\nabla\mu,\nabla\Sigma}$ defined in (39), one has

$$\partial_t\left(\frac{1}{2}p^2 + \frac{c}{2}\|q\|_2^2\right) \leq \mathcal{L}_{\mu,\Sigma}\left(\frac{1}{2}p^2 + \frac{c}{2}\|q\|_2^2\right) + \frac{1}{2a}\left(\|\nabla\mu\|_{L^\infty}^2 + c\|\mu\|_{L^\infty}^2\right) + a\left(\frac{1}{2}\|p\|_2^2 + \frac{c}{2}\|q\|_2^2\right).$$

Let $g(s,t) = \left(\frac{c}{2}p^2 + \frac{1}{2}\|q\|_2^2\right)e^{-at} + \frac{1}{2a^2}(c\|\mu\|_{L^\infty}^2 + \|\nabla\mu\|_{L^\infty}^2)e^{-at}$, then

$$\partial_t g \leq \mathcal{L}_{\mu,\Sigma} g, \quad \text{with } g(s,0) = \frac{1}{2a^2}(c\|\mu\|_{L^\infty}^2 + \|\nabla\mu\|_{L^\infty}^2).$$

By the comparison theorem, one has $g(s,t) \leq 0$, which implies

$$\frac{c}{2}\|p\|_{L^\infty}^2 + \frac{1}{2}\|q\|_{L^\infty}^2 \leq \frac{1}{2a^2}(c\|\mu\|_{L^\infty}^2 + \|\nabla\mu\|_{L^\infty}^2)(e^{at} - 1)$$

$$\|q\|_{L^\infty} \leq (\sqrt{c}\|\mu\|_{L^\infty} + \|\nabla\mu\|_{L^\infty})\sqrt{e^t - 1}$$

□

6.5.3 Proof of Lemma 6.6

Based on Proposition 6.3, one has

$$\beta \|V\|_\rho^2 - \langle r(s), V(s)\rangle_\rho \leq -\frac{\lambda_{\min}}{2} \|\nabla V(s)\|_\rho^2;$$
$$\beta \|\nabla V\|_\rho^2 - \langle \nabla r(s), \nabla V(s)\rangle_\rho \leq \frac{C_{\nabla\mu,\nabla\Sigma}}{2} \|\nabla V\|_\rho^2.$$

Multiplying $\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}$ to the first inequality and adding it to the second one gives

$$\begin{aligned}&\beta\left(\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}\|V\|^2 + \|\nabla V(s)\|_\rho^2\right)\\ \leq&\frac{1}{2\beta}\left(\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}\|r\|_\rho^2 + \|\nabla r\|_\rho^2\right) + \frac{\beta}{2}\left(\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}\|V\|_\rho^2 + \|\nabla V\|_\rho^2\right),\end{aligned} \tag{56}$$

which implies

$$\|\nabla V(s)\|_\rho^2 \leq \frac{1}{\beta}\left(\sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}}\|r\|_\rho + \|\nabla r\|_\rho\right).$$

For the second inequality, Let $p(s,t) = \mathbb{E}[r(s_t)|s_0 = s]$, then by Lemma 6.8, one has

$$\begin{aligned}\left\|\nabla \hat{V}(s)\right\|_{L^\infty} &= \left\|\int_0^\infty e^{-\beta t}\nabla p(s,t)dt\right\|_{L^\infty} \leq \int_0^\infty e^{-\beta t}\|\nabla p(s,t)\|_{L^\infty} dt\\ \leq&\int_0^\infty e^{-\beta t}\left(\sqrt{\frac{C_{\nabla\hat{\mu},\nabla\hat{\Sigma}}}{\lambda_{\min}}}\|r\|_{L^\infty} + \|\nabla r\|_{L^\infty}\right)dt\\ \leq&\frac{1}{\beta}\left(\frac{1}{\sqrt{\lambda_{\min}}}\sqrt{2C_{\nabla\mu,\nabla\Sigma} + \frac{1}{2\lambda_{\min}}\left\|\nabla(\Sigma-\hat{\Sigma})\right\|_{L^\infty}^2 + 2\|\nabla(\mu-\hat{\mu})\|_{L^\infty}}\|r\|_{L^\infty} + \|\nabla r\|_{L^\infty}\right).\end{aligned}$$

Similar to the estimate of the $\left\|\nabla\cdot(\Sigma-\hat{\Sigma})\right\|_{L^\infty}$ in Lemma 6.2, one has,

$$\left\|\nabla(\Sigma-\hat{\Sigma}_i)\right\|_{L^\infty} \leq L_{\nabla\Sigma}\Delta t^i + o(\Delta t^i),$$

where

$$L_{\nabla\Sigma} = \hat{C}_i\left[\sqrt{C_{\nabla\mu,\nabla\Sigma}/\lambda_{\min}}\|h\|_{L^\infty} + \|\nabla h\|_{L^\infty} + 4\left(\sqrt{C_{\nabla\mu,\nabla\Sigma}/\lambda_{\min}}\|\mu\|_{L^\infty} + \|\nabla\mu\|_{L^\infty}\right)\right], \tag{57}$$

with $\hat{C}_i, h(s), C_{\nabla\mu,\nabla\Sigma}$ defined in (46), (48), (39). Taking ∇ to (45) gives,

$$\|\nabla(\mu-\hat{\mu}_i)\|_{L^\infty} \leq \frac{1}{\Delta t i!}\sum_{j=1}^{i}|a_j^{(i)}|\int_0^{j\Delta t}\|\nabla p(s,t)\|\, t^i dt$$

with

$$p(s,t) = \mathbb{E}[\mathcal{L}^i_{\mu,\Sigma}\mu(s_t)|s_0 = s].$$

By Lemma 6.8, one has,

$$\|\nabla(\mu - \hat{\mu}_i)\|_{L^\infty} \leq L_{\nabla\mu}\Delta t^i$$

with

$$L_{\nabla\mu} = C_i\left(\sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}}\left\|\mathcal{L}^i_{\mu,\Sigma}\mu(s)\right\|_{L^\infty} + \left\|\nabla\mathcal{L}^i_{\mu,\Sigma}\mu(s)\right\|_{L^\infty}\right). \tag{58}$$

Therefore,

$$\left\|\nabla\hat{V}(s)\right\|_{L^\infty} \leq \frac{1}{\beta}\left[\left(\sqrt{\frac{2C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}} + \frac{L_{\nabla\Sigma}}{\sqrt{2}\lambda_{\min}}\Delta t^i + \sqrt{\frac{2L_{\nabla\mu}}{\lambda_{\min}}}\Delta t^{i/2}\right)\|r\|_{L^\infty} + \|\nabla r\|_{L^\infty}\right] + o(\Delta t^i).$$

6.6 Proof of Theorem 4.1

The i-th approximation $\hat{V}_i(s)$ can be divided into two parts,

$$\hat{V}_i(s) = \hat{V}_i^P(s) + e_i^P(s) \quad \text{with} \quad \hat{V}_i^P(s) = \sum_{k=1}^n \hat{v}_k\phi_k(s), e_i^P(s) = \hat{V}_i(s) - \hat{V}_i^P(s),$$

where $\hat{V}_i^P(s)$ could be any functions in the linear space spanned by $\Phi(s)$. Note that $\hat{V}_i(s)$ satisfies

$$\left\langle(\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i})\hat{V}_i(s), \Phi\right\rangle_\rho = \langle r(s), \Phi(s)\rangle_\rho,$$

which can be divided into two parts,

$$\left\langle(\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i})\hat{V}_i^P(s), \Phi\right\rangle_\rho = \langle r(s), \Phi(s)\rangle_\rho - \left\langle(\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i})e_i^P(s), \Phi\right\rangle_\rho,$$

subtract the above equation from the Galerkin equation (21) gives

$$\left\langle(\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i})(\hat{V}_i^G(s) - \hat{V}_i^P(s)), \Phi\right\rangle_\rho = \left\langle(\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i})e_i^P(s), \Phi\right\rangle_\rho.$$

Let $e_i^G(s) = \hat{V}_i^G(s) - \hat{V}_i^P(s) = \sum_{k=1}^n e_k\phi_k(s)$, then multiplying $(e_1, \cdots, e_n)$ to the above equation yields,

$$\begin{aligned} \left\langle(\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i})e_i^G, e_i^G\right\rangle_\rho &= \left\langle(\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i})e_i^P, e_i^G\right\rangle_\rho \\ \left\langle(\beta - \mathcal{L}_{\mu,\Sigma})e_i^G, e_i^G\right\rangle_\rho &= \left\langle\mathcal{L}_{\hat{\mu}_i-\mu,\hat{\Sigma}_i-\Sigma}e_i^G, e_i^G\right\rangle_\rho + \left\langle(\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i})e_i^P, e_i^G\right\rangle_\rho. \end{aligned} \tag{59}$$

When $L_\rho^\infty = \|\nabla\log\rho\|_{L^\infty}$ is bounded, by applying the last inequality of Lemma 6.3 and Lemma 6.5, one has

$$\begin{aligned} \beta\left\|e_i^G\right\|_\rho^2 + \frac{\lambda_{\min}}{2}\left\|\nabla e_i^G\right\|_\rho^2 \leq & c_1\left\|\nabla e_i^G\right\|_\rho^2 + c_2\left\|e_i^G\right\|_\rho\left\|\nabla e_i^G\right\|_\rho + \beta\left\|e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho \\ & + c_3\left\|\nabla e_i^P\right\|_\rho\left\|\nabla e_i^G\right\|_\rho + c_4\left\|\nabla e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho, \end{aligned}$$

where

$$
\begin{aligned}
c_1 &= \frac{L_\Sigma}{2}\Delta t^i, \quad c_2 = \left(L_\mu + \frac{L_{\nabla\cdot\Sigma}}{2} + \frac{L_\Sigma L_\rho^\infty}{2}\right)\Delta t^i, \\
c_3 &= \frac{1}{2}\left(\|\Sigma\|_{L^\infty} + L_\Sigma \Delta t^i\right), \\
c_4 &= \left(\|\mu\|_{L^\infty} + \frac{\|\nabla\cdot\Sigma\|_{L^\infty}}{2} + \frac{\|\Sigma\|_{L^\infty} L_\rho^\infty}{2}\right) + \left(L_\mu + \frac{L_{\nabla\cdot\Sigma}}{2} + \frac{L_\Sigma L_\rho^\infty}{2}\right)\Delta t^i
\end{aligned} \tag{60}
$$

with $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}$ defined in (46), (49), (57). Under the assumption that $c_1 \le \frac{\lambda_{\min}}{8}$, $\frac{c_2}{2} \le \min\left\{\frac{\lambda_{\min}}{8}, \frac{\beta}{2}\right\}$, i.e.,

$$
\Delta t^i \le \eta_{\mu,\Sigma,\beta}, \quad \eta_{\mu,\Sigma,\beta} = \frac{\min\left\{\frac{\lambda_{\min}}{2}, 2\beta\right\}}{2L_\mu + L_{\nabla\cdot\Sigma} + L_\Sigma \max\{L_\rho^\infty, 4\}} \tag{61}
$$

with $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}$ defined in (46), (49), (57), one has

$$
\begin{aligned}
\frac{\beta}{2}\left\|e_i^G\right\|_\rho^2 + \frac{\lambda_{\min}}{4}\left\|\nabla e_i^G\right\|_\rho^2 \le& \beta\left\|e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho + c_3\left\|\nabla e_i^P\right\|_\rho\left\|\nabla e_i^G\right\|_\rho, \\
&+ c_4\left\|\nabla e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho \\
\left\|e_i^G\right\|_\rho \le& \sqrt{8}\left\|e_i^P\right\|_\rho + \sqrt{\frac{4}{\beta}\left(\frac{2c_3^2}{\lambda_{\min}} + \frac{2c_4^2}{\beta}\right)}\left\|\nabla e_i^P\right\|_\rho,
\end{aligned}
$$

which implies that

$$
\left\|\hat{V}_i - \hat{V}_i^G\right\|_\rho \le \left\|e_i^P\right\|_\rho + \left\|e_i^G\right\|_\rho \le (\sqrt{8}+1)\left\|e_i^P\right\|_\rho + \sqrt{\frac{4}{\beta}\left(\frac{2c_3^2}{\lambda_{\min}} + \frac{2c_4^2}{\beta}\right)}\left\|\nabla e_i^P\right\|_\rho.
$$

Since the above inequality holds for all V_i^P in the linear space spanned by $\{\Phi\}$, therefore,

$$
\left\|\hat{V}_i - \hat{V}_i^G\right\|_\rho \le C_G \inf_{V=\theta^\top\Phi}\left\|\hat{V}_i - V\right\|_{H_\rho^1}
$$

where

$$
C_G = \max\left\{\sqrt{8}+1, \sqrt{\frac{4}{\beta}\left(\frac{2c_3^2}{\lambda_{\min}} + \frac{2c_4^2}{\beta}\right)}\right\}, \quad \|f\|_{H_\rho^1} = \|f\|_\rho + \|\nabla f\|_\rho. \tag{62}
$$

with c_3, c_4 defined in (60). When $L_\rho^\infty = \|\nabla \log\rho\|_{L^\infty}$ is not bounded, by applying the last second inequality of Lemma 6.3 and Lemma 6.5 to (59), one has

$$
\begin{aligned}
\beta\left\|e_i^G\right\|_\rho^2 + \frac{\lambda_{\min}}{2}\left\|\nabla e_i^G\right\|_\rho^2 \le& c_1\left\|\nabla e_i^G\right\|_\rho^2 + c_5\left\|e_i^G\right\|_\rho\left\|\nabla e_i^G\right\|_\rho + c_6\left\|e_i^G\right\|_{L^\infty}\left\|\nabla e_i^G\right\|_\rho \\
+ \beta\left\|e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho &+ c_3\left\|\nabla e_i^P\right\|_\rho\left\|\nabla e_i^G\right\|_\rho + c_7\left\|\nabla e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho + c_8\left\|\nabla e_i^P\right\|_{L^\infty}\left\|e_i^G\right\|_\rho,
\end{aligned} \tag{63}
$$

where

$$
\begin{aligned}
c_5 &= \left(L_\mu + \frac{L_{\nabla\cdot\Sigma}}{2}\right)\Delta t^i, \quad c_8 = \frac{\|\Sigma\|_{L^\infty} L_\rho}{2} + \frac{L_\Sigma L_\rho}{2}\Delta t^i, \\
c_6 &= \frac{L_\Sigma L_\rho}{2}\Delta t^i, \quad c_7 = \left(\|\mu\|_{L^\infty} + \frac{\|\nabla\cdot\Sigma\|_{L^\infty}}{2}\right) + \left(L_\mu + \frac{L_{\nabla\cdot\Sigma}}{2}\right)\Delta t^i,
\end{aligned} \tag{64}
$$

with $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}, L_\rho$ defined in (46), (49), (57), (20). Since

$$
\left\|e_i^G\right\|_{L^\infty} = \left\|e^\top\Phi\right\|_{L^\infty} \le \|e\|_2 \|\Phi\|_{L^\infty} \le \frac{1}{\sqrt{\hat{\lambda}}}\left\|e_i^G\right\|_\rho \|\Phi\|_{L^\infty},
$$

where $\hat{\lambda}$ is the smallest eigenvalue of the matrix B, where $B_{ij} = \int \phi_i(s)\phi_i(s)\rho(s)ds$. Note that the matrix B is always positive definite when $\{\phi_i(s)\}$ are linear independent bases w.r.t. the weighted L^2 norm. let $L_\Phi^\infty = \|\Phi\|_{L^\infty}$, then (63) can be rewritten as

$$
\begin{aligned}
\beta\left\|e_i^G\right\|_\rho^2 + \frac{\lambda_{\min}}{2}\left\|\nabla e_i^G\right\|_\rho^2 \le{}& c_1\left\|\nabla e_i^G\right\|_\rho^2 + c_9\left\|e_i^G\right\|_\rho\left\|\nabla e_i^G\right\|_\rho + \beta\left\|e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho \\
&+ c_3\left\|\nabla e_i^P\right\|_\rho\left\|\nabla e_i^G\right\|_\rho + c_7\left\|\nabla e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho + c_8\left\|\nabla e_i^P\right\|_{L^\infty}\left\|e_i^G\right\|_\rho,
\end{aligned}
$$

where

$$
c_9 = \left(L_\mu + \frac{L_{\nabla\cdot\Sigma}}{2} + \frac{L_\Sigma L_\rho L_\Phi^\infty}{2\sqrt{\hat{\lambda}}}\right)\Delta t^i.
$$

When $c_1 \le \frac{\lambda_{\min}}{8}$, $\frac{c_9}{2} \le \min\{\frac{\lambda_{\min}}{8}, \frac{\beta}{2}\}$, i.e.,

$$
\Delta t^i \le \hat{\eta}_{\mu,\Sigma,\beta}, \quad \hat{\eta}_{\mu,\Sigma,\beta} = \frac{\min\left\{\frac{\lambda_{\min}}{2}, 2\beta\right\}}{2L_\mu + L_{\nabla\cdot\Sigma} + L_\Sigma \max\left\{\frac{L_\rho L_\Phi^\infty}{\sqrt{\hat{\lambda}}}, 4\right\}}, \tag{65}
$$

with $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}$ defined in (46), (49), (57), then one has

$$
\begin{aligned}
\frac{\beta}{2}\left\|e_i^G\right\|_\rho^2 + \frac{\lambda_{\min}}{4}\left\|\nabla e_i^G\right\|_\rho^2 \le{}& \beta\left\|e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho + c_3\left\|\nabla e_i^P\right\|_\rho\left\|\nabla e_i^G\right\|_\rho \\
&+ c_7\left\|\nabla e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho + c_8\left\|\nabla e_i^P\right\|_{L^\infty}\left\|e_i^G\right\|_\rho, \\
\left\|e_i^G\right\|_\rho \le \sqrt{12}\left\|e_i^P\right\|_\rho +{}& \sqrt{\frac{4c_3^2}{\beta\lambda_{\min}} + \frac{12c_7^2}{\beta^2}}\left\|\nabla e_i^P\right\|_\rho + \frac{\sqrt{12}c_8}{\beta}\left\|\nabla e_i^P\right\|_{L^\infty},
\end{aligned}
$$

which implies that

$$
\left\|\hat{V}_i - \hat{V}_i^G\right\|_\rho \le (\sqrt{12}+1)\left\|e_i^P\right\|_\rho + \sqrt{\frac{4c_3^2}{\beta\lambda_{\min}} + \frac{12c_7^2}{\beta^2}}\left\|\nabla e_i^P\right\|_\rho + \frac{\sqrt{12}c_8}{\beta}\left\|\nabla e_i^P\right\|_{L^\infty}.
$$

Since the above inequality holds for all V_i^P in the linear space spanned by $\{\Phi\}$, therefore,

$$
\left\|\hat{V}_i - \hat{V}_i^G\right\|_\rho \le \hat{C}_G \inf_{V=\theta^\top\Phi}\left\|\hat{V}_i - V\right\|_{H^1_{\rho,\infty}},
$$

where

$$\hat{C}_G = \max\left\{\sqrt{12}+1, \sqrt{\frac{4c_3^2}{\beta\lambda_{\min}} + \frac{12c_7^2}{\beta^2}}, \frac{\sqrt{12}c_8}{\beta}\right\}, \quad \|f\|_{H^1_{\rho,\infty}} = \|f\|_{H^1_\rho} + \|\nabla f\|_{L^\infty}. \tag{66}$$

with c_3, c_7, c_8 defined in (60), (64).

7 Conclusion

In this paper, we introduce PhiBE, a PDE-based Bellman equation that integrates discrete-time information into continuous-time PDEs. The new BE outperforms the classical Bellman equation in approximating the continuous-time policy evaluation problem, particularly in scenarios where underlying dynamics evolve slowly. Importantly, the approximation error of PhiBE depends on the dynamics, making it more robust against changes in reward structures. This property allows greater flexibility in designing reward functions to effectively achieve RL objectives. Furthermore, we propose higher-order PhiBE, which offers superior approximations to the true value function. One can achieve the same error with sparse data, enhancing the learning efficiency.

This paper serves as the first step for using PhiBE in continuous-time RL, laying the groundwork for future research directions. Specifically, we defer the investigation of the degenerate diffusion case and sample complexity considerations for subsequent studies. Additionally, our methodology can also be extended to encompass broader RL settings.

References

[1] Matteo Basei, Xin Guo, Anran Hu, and Yufei Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. The Journal of Machine Learning Research, 23(1):8015–8048, 2022.

[2] Vladimir I Bogachev, N Krylov, and Michael Röckner. Regularity of invariant measures: the case of non-constant diffusion part. journal of functional analysis, 138(1):223–242, 1996.

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316, 2016.

[4] Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. Machine learning, 22:33–57, 1996.

[5] Cosimo Della Santina, Christian Duriez, and Daniela Rus. Model-based control of soft robots: A survey of the state of the art and open challenges. IEEE Control Systems Magazine, 43(3):30–65, 2023.

[6] Kenji Doya. Reinforcement learning in continuous time and space. Neural computation, 12(1):219–245, 2000.

[7] Wen Huang, Min Ji, Zhenxin Liu, and Yingfei Yi. Steady states of fokker–planck equations: I. existence. Journal of Dynamics and Differential Equations, 27:721–742, 2015.

[8] Yanwei Jia and Xun Yu Zhou. Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. The Journal of Machine Learning Research, 23(1):6918–6972, 2022.

[9] Yanwei Jia and Xun Yu Zhou. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. The Journal of Machine Learning Research, 23(1):12603–12652, 2022.

[10] Rushikesh Kamalapurkar, Patrick Walters, and Warren E Dixon. Model-based reinforcement learning for approximate optimal regulation. Automatica, 64:94–104, 2016.

[11] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. The International Journal of Robotics Research, 32(11):1238–1274, 2013.

[12] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. Advances in neural information processing systems, 12, 1999.

[13] Jason Kong, Mark Pfeiffer, Georg Schildbach, and Francesco Borrelli. Kinematic and dynamic vehicle models for autonomous driving control design. In 2015 IEEE intelligent vehicles symposium (IV), pages 1094–1099. IEEE, 2015.

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. nature, 518(7540):529–533, 2015.

[15] Bernt Oksendal. Stochastic differential equations: an introduction with applications. Springer Science & Business Media, 2013.

[16] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.

[17] Grigorios A Pavliotis. Stochastic processes and applications. Springer, 2016.

[18] Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In ICML, pages 417–424, 2001.

[19] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. nature, 529(7587):484–489, 2016.

[20] Daniel W Stroock and SR Srinivasa Varadhan. Multidimensional diffusion processes, volume 233. Springer Science & Business Media, 1997.

[21] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.

[22] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems, 12, 1999.

[23] Corentin Tallec, Léonard Blier, and Yann Ollivier. Making deep q-learning methods robust to time discretization. In International Conference on Machine Learning, pages 6096–6104. PMLR, 2019.

[24] Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. The Journal of Machine Learning Research, 21(1):8145–8178, 2020.

[25] Yuhua Zhu, Zachary Izzo, and Lexing Ying. Borrowing from the future: Addressing double sampling in model-free control. In Mathematical and Scientific Machine Learning, pages 1099–1136. PMLR, 2022.

[26] Yuhua Zhu and Lexing Ying. Borrowing from the future: An attempt to address double sampling. In Mathematical and scientific machine learning, pages 246–268. PMLR, 2020.

[27] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593, 2019.