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

Yuhua Zhu*

(This is a draft version)

**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 is not always a reliable approximation to the true value function. We then introduce PhiBE, a PDE-based Bellman equation that 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. 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) [20] has achieved significant success in applications inherently viewed as Markov decision processes. Remarkable milestones include its applications in Atari Games [13], AlphaGO [18], and ChatGPT [24, 15], 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, 12] and robotics [10], state changes continuously over time regardless of whether actions are discrete or not. In contrast to discrete-time decision-making applications, RL encounters challenges when applied to continuous-time decision-making processes. This paper directs its focus toward addressing continuous-time reinforcement learning problems that can be equivalently viewed as a stochastic optimal control problem with unknown dynamics [23, 6]. Since one can divide the RL problem into policy evaluation

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

and policy update [21, 11, 22, 8], we first focus on the continuous-time policy evaluation (PE) problem in this paper.

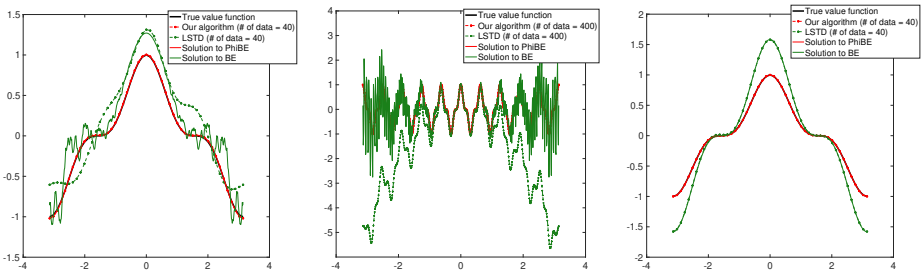
Given discrete-time trajectory data generated from the underlying dynamics, a common approach to address the continuous-time PE problem involves discretizing time and treating it as a Markov decision process. This method yields an approximated value function satisfying a Bellman equation, thereby one can use RL algorithms such as Temporal difference[20], gradient TD[17], Least square TD [4] to solve the Bellman equation. However, this paper shows that the Bellman equation is not always a good tool for solving 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, 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 with 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. 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, the proposed algorithm outperforms the RL algorithms.

**Contributions**

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




(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 $\Delta t$ unite of time, $\beta$ is the discount coefficient, and $V(s)$ is the true value function. Here, 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 model-free algorithm for solving PhiBE when only discrete-time data is available.

**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 [9, 5]. However, directly identifying the continuous dynamics is often challenging. 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. [7] 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 [7]), 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 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.

# 2 Setting

Consider the following continuous-time PE problem, where the value function $V(s) \in \mathbb{R}$ is defined as

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

and the state $s_t \in \mathbb{S} = \mathbb{R}^d$ is driven by the stochastic differential equation (SDE),

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

Here $\mu(s) \in \mathbb{R}^d, \sigma(s) \in \mathbb{R}^{d\times d}$ are unknown functions. In this paper, we assume that $\mu(s), \sigma(s)$ are Lipschitz continuous and reward function $\|r\|_{L^\infty}$ is bounded. This ensures that (2) has a unique strong solution [14] and the infinite horizon integral is bounded.

We aim to determine 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 $\Delta t$, 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 $s$ generated by the continuous dynamics (2) and collected at discrete time $j\Delta t$ is given. Here $s = \{s_0^l, s_{\Delta t}^l, \cdots, s_{m\Delta t}^l\}_{l=1}^I$ contains $I$ independent trajectories, and the initial data $s_0^i$ 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 analyses in Section 3 are conducted under Case 1. We demonstrate that the Bellman equation is not the optimal equation to solve continuous-time reinforcement learning problems under certain circumstances, 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 traditional Bellman equation. When only trajectory data is available (Case 2), one can also use the data to estimate the value function directly, referred to as model-free RL, which will be discussed in Section 4.

# 3 A PDE-based Bellman Equation (PhiBE)

In this section, we first introduce the Bellman equation in Section 3.1, 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. [20]

**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 using the given 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 Bellman equation and the true value function $V$.

**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*

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

*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, this condition 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, $\|s \cdot \nabla r(s)\|_{L^\infty} \leq C$.

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 come 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:

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

characterizes the error of the BE $\left\|V - \tilde{V}\right\|_{L^\infty}$ 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 $\Delta t$ suggests that for the same time discretization $\Delta t$, when $\beta$ 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 $\beta$ is large, the error is mainly affected by $\|r\|_{L^\infty}$, implying that the error increases when the upper bound of the reward is large.

The question that the rest of this sectionseeks to address is whether, given the same discrete-time information, i.e., the transition distributuion $\rho(s', \Delta t|s)$, time discretization $\Delta t$, and discount coefficient $\beta$, we 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 becomes 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 chain rule on the LHS of the above equation yields $\frac{d}{dt}V(s_t) = \mu(s_t) \cdot \nabla V(s_t)$, and the RHS can be written as $\beta V(s_t) - r(s_t)$, resulting in a PDE for the true value function

$$\beta V(s_t) = r(s_t) + \mu(s_t) \cdot \nabla V(s_t).$$

or equivalently,

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

Then, 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,

$$\beta \hat{V}(s) = r(s) + \frac{1}{\Delta t}(p_{\Delta t}(s) - s) \cdot \nabla \hat{V}(s), \tag{9}$$

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, and hence $\left\| \frac{d^2}{dt} s_t \right\|_{L^\infty} = \|\mu \cdot \nabla \mu\|_{L^\infty}$ is small, the error diminishes.

We refer to (9) as PhiBE, an abbreviation for the physics-informed Bellman equation, because it incorporates both the current state and the state after $\Delta t$, similar to the Bellman equation, while also resembling the form of the PDE (8) derived from the true continuous-time physical environment. However, unlike the true PDE (8) 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-

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

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, $\|\mu(s) - \hat{\mu}_2(s)\|_{L^\infty}$ 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,

$$\beta \hat{V}_i(s) = r(s) + \hat{\mu}_i(s) \cdot \nabla \hat{V}_i(s), \tag{10}$$

where

$$\hat{\mu}_i(s) = \frac{1}{\Delta t} \sum_{j=1}^{i} a_j \left( \underbrace{p_{\Delta t} \circ \cdots \circ p_{\Delta t}}_{j}(s) - s \right), \tag{11}$$

and

$$(a_0, \cdots, a_i)^\top = A^{-1} b, \quad \text{with } A_{kj} = j^k,\ b_k = \begin{cases} 0, & k \neq 1 \\ 1, & k = 1 \end{cases} \text{ for } 0 \leq j, k \leq i. \tag{12}$$

**Remark 2.** Note that $\mu_i(s)$ can be equivalently written as

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

There is an equivalent definition of $(a_0, \cdots, a_i)$, given by

$$\sum_{j=0}^{i} a_j j^k = \begin{cases} 0, & k \neq 1, \\ 1, & k = 1, \end{cases} \quad \text{for } 0 \leq j, k \leq i. \tag{13}$$

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, here it is the opposite. 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. In Lemma 3.2, we establish that $\hat{\mu}_i(s)$ is an $i$-th order approximation to $\mu(s)$. Then, in Theorem 3.3, we prove that $\hat{V}_i(s)$ is an $i$-th order approximation to $V(s)$.

**Lemma 3.2.** *Assume that $\left\|\mathcal{L}_\mu^i \mu(s)\right\|_{L^\infty}$ is bounded, then the distance between $\hat{\mu}_i(s)$ defined in* (11) *and the true dynamics can be bounded by*

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

*where*

$$C_i = \frac{\sum_{j=0}^{i} |a_j| j^{i+1}}{(i+1)!}, \quad \mathcal{L}_\mu = \mu(s) \cdot \nabla. \tag{14}$$

**Remark 3** (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 the condition. We lose some sharpness for the upper bound to make the theorem work for all general dynamics. However, we prove in Theorem 3.4 that PhiBE works when $\mu(s) = \lambda s$, and one can derive a sharper error estimate for this case.

**Theorem 3.3.** *Assume that $\|\nabla r(s)\|_{L^\infty}, \left\|\mathcal{L}_\mu^i \mu(s)\right\|_{L^\infty}$ are bounded. Additionally, assume that $\|\nabla \mu\|_{L^\infty} < \beta$, then the solution $\hat{V}_i(s)$ to the PhiBE* (10) *is an ith-order approximation to the true value function $V(s)$ defined in* (1) *with an error*

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

*where $C_i$ is a constant defined in* (14)*.*

See Section 6.2 for the proof of Lemma 3.2 and Theorem 3.3.

**Remark 4** (1st-order PhiBE v.s. BE)**.** By Theorem (3.3), the distance between the first order PhiBE solution and the true value function can be bounded by

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

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

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

one observes that when the change of the reward $\|\nabla r\|_{L^\infty}$ is rapid but the change in velocity is slow, i.e., $\left\|\frac{d^2}{dt^2}s_t\right\|_{L^\infty} = \|\mu \cdot \nabla \mu\|_{L^\infty}$ 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 5** (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 3.3, 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.

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

**Theorem 3.4.** *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*

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

*where $C_i$ is defined in* (14)*.*

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 3.4, the error of the $i$-th order PhiBE for linear dynamics on $\lambda^{i+1}$. Consequently, when $\beta$ is small or the upper bound of the reward function is large, a small $\lambda$ will make the first-order PhiBE a superior approximation to the Bellman equation.

### 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 [19], the value function $V(s)$ satisfies the following PDE,

$$\beta V(s) = r(s) + \mathcal{L}_{\mu,\Sigma} V(s), \tag{15}$$

where $\mathcal{L}_{\mu,\Sigma}$ is an operator defined in (4). However, one cannot directly solve the PDE (15) 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 approximation 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,

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

where

$$\begin{aligned} \hat{\mu}_i(s) &= \frac{1}{\Delta t} \mathbb{E}_{s_{j\Delta t} \sim \rho(\cdot, j\Delta t|s)} \left[ \sum_{j=1}^{i} a_j (s_{j\Delta t} - s_0) | s_0 = s \right] \\ \hat{\Sigma}_i(s) &= \frac{1}{\Delta t} \mathbb{E}_{s_{j\Delta t} \sim \rho(\cdot, j\Delta t|s)} \left[ \sum_{j=1}^{i} a_j (s_{j\Delta t} - s_0)(s_{j\Delta t} - s_0)^\top | s_0 = s \right] \end{aligned} \tag{17}$$

where $\mathcal{L}_{\hat{\mu}_i, \hat{\Sigma}_i}$ is defined in (4), and $a = (a_0, \cdots, a_i)^\top$ is defined in (12).

Here we present the first and second order approximations. The first-order approximation is as follows:

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

and the second-order approximation reads,

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

Next, we show the solution $\hat{V}_i(s)$ to the higher order Bellman equation provides a higher-order approximation to the true value function $V(s)$. To establish $i$-th order approximation, 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)* $\max_{k,l} \sum_i \|\partial_{s_i} \Sigma_{kl}\|_{L^\infty} \leq 2\lambda_{min}$

*(c)* $\left\|\nabla^k \mu(s)\right\|_{L^\infty}, \left\|\nabla^k \Sigma(s)\right\|_{L^\infty}$ *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 $V(s)$. imposes a restriction on the change in diffusion, ensuring it remains smaller than the coercivity, which is used in proving the regularity of $\nabla V(s)$. The last assumption is employed to demonstrate that $\hat{\mu}_i$ and $\hat{\Sigma}_i$ are $i$-th approximations to $\mu, \Sigma$, respectively.

Assume that there exists a unique stationary distribution $\rho(s)$ to the stochastic dynamics that satisfies,

$$\int \mathcal{L}_{\mu,\Sigma} \phi(s) \rho(s) ds = 0. \quad \text{for } \forall \phi(s) \in C_c^\infty, \tag{18}$$

then, we define a weighted $L^2$ norm

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

**Theorem 3.5.** *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 is given in Section 6.4.

**Theorem 3.6.** *Under Assumption 1, and* $\Delta t^i \leq \frac{\lambda_{min}}{4D_{\mu,\Sigma}}$*, the solution* $\hat{V}_i(s)$ *to the i-th order PhiBE* (16) *is an i-th order approximation to the true value function* $V(s)$ *that satisfying* (15) *with an error*

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

*where* $C_{r,\mu,\Sigma,\lambda_{min}}, C_{r,\mu,\Sigma}$ *are constants defined in* (51) *depending on* $\mu(s), \Sigma(s), r(s), \lambda_{min}$*, and* $D_{\mu,\Sigma}$ *is a constant defined in* (49) *depending on* $\mu, \Sigma$ *.*

**Remark 6** (1st-order PhiBE v.s. BE)**.** By the above Theorem, the distance between the first-order PhiBE in the stochastic dynamics can be bounded by

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

where

$$\begin{aligned}
&L_{\Sigma,\rho} \leq \sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}}\left(\|\mathcal{L}\Sigma\|_\rho + \left\|\mu\mu^\top\right\|_\rho\right) + \|\nabla\mathcal{L}\Sigma\|_\rho + \left\|\nabla(\mu\mu^\top)\right\|_\rho, \\
&L_\Sigma \lesssim \|\mathcal{L}\Sigma\|_{L^\infty} + \left\|\mu\mu^\top\right\|_{L^\infty}, \\
&L_\mu \lesssim \|\mathcal{L}\mu\|_{L^\infty}, \\
&C_{\nabla\mu,\nabla\Sigma} \lesssim \|\nabla\mu\|_{L^\infty} + \|\nabla\Sigma\|_{L^\infty},
\end{aligned}$$

where $\mathcal{L}$ represents $\mathcal{L}_{\mu,\Sigma}$. This indicates that when $\lambda_{\min}$ 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, it implies that when the noise is large or the change in the dynamics is small, then the first order PhiBE solution is a better approximation to the true value function. On the other hand, if the change in the reward is small, then BE is a better approximation than the first-order PhiBE.

The proof of Theorem 3.6 is given in Section 6.5. We provide the brief proof sketch here.

**Proof sketch.** Let $e = V - \hat{V}_i$. By energy estimate of (15) and (16), one has

$$\beta \|e\|_\rho^2 = \langle \mathcal{L}_{\mu,\Sigma} e, e\rangle_\rho + \left\langle (\mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i} - \mathcal{L}_{\mu,\Sigma})\hat{V}_i, e\right\rangle_\rho .$$

By the coercivity of the operator $\mathcal{L}_{\mu,\Sigma}$ in Lemma 6.1, one can bound

$$\langle \mathcal{L}_{\mu,\Sigma} e, e\rangle_\rho \leq -\frac{\lambda_{\min}}{2} \|e\|_\rho . \tag{19}$$

In order to bound the second term $\left\langle (\mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i} - \mathcal{L}_{\mu,\Sigma})\hat{V}_i, e\right\rangle_\rho$, one requires the distance $\|\mu - \hat{\mu}_i\|, \left\|\Sigma - \hat{\Sigma}_i\right\|, \left\|\nabla(\Sigma - \hat{\Sigma}_i)\right\| \leq O(\Delta t^i)$, which is provided by Lemma 6.4, and the boundedness of $\|V\|, \|\nabla V\|$ provided by Lemma 6.2 and Lemma 6.3, respectively. Based on the above Lemmas, one can bound

$$\left\langle (\mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i} - \mathcal{L}_{\mu,\Sigma})\hat{V}_i, e\right\rangle_\rho \lesssim \Delta t^i \|\nabla e\|_\rho . \tag{20}$$

Combining the two upper bounds (19), (20) yields the bound for $\|e\|_\rho$.

# 4 Model-free Algorithm for continuous-time Policy Eveluation

In the 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. Subsequently, we introduce a model-free Galerkin method in Section 4.2.

## 4.1 Galerkin Method

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

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

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

### 4.2 Model-free Galerkin method for PhiBE

In continuous-time policy evaluation problem, one does not have access to the underlying dynamics $\mu, \Sigma$, however, the approximated dyanmcs $\hat{\mu}_i, \hat{\Sigma}_i$ 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 $A$ by $\hat{A}_i$

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

Now, 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,

$$\begin{aligned} \bar{\mu}_i(s^l_{j\Delta t}) &= \frac{1}{\Delta t}\sum_{k=1}^{i} a_k (s^l_{(k+j)\Delta t} - s^l_{j\Delta t}), \\ \bar{\Sigma}_i(s^l_{j\Delta t}) &= \frac{1}{\Delta t}\sum_{k=1}^{i} a_k (s^l_{(k+j)\Delta t} - s^l_{j\Delta t})(s^l_{(k+j)\Delta t} - s^l_{j\Delta t})^\top. \end{aligned} \tag{21}$$

Then, using the above unbiased estimate, one can approximate the matrix $\hat{A}$ and the vector $b$ by

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

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 reward or even unknown reward. We summarize the model-free Galerkin method in Algorithm 1.

## 5 Numerical experiments

### 5.1 Deterministic dynamics

We first consider deterministic dynamics, where the state space $\mathbb{S}$ is defined as $\mathbb{S} = [-\pi, \pi]$, and discount coefficient $\beta = 0.1$. We consider two kinds of underlying dynamics, one is linear,

$$\frac{d}{dt}s_t = \lambda s_t, \tag{22}$$

and the other is nonlinear,

$$\frac{d}{dt}s_t = \lambda \sin^2(s_t). \tag{23}$$

**Algorithm 1** Model-free Galerkin method for $i$-th order PhiBE

**Given:** discrete time step $\Delta t$, discount coefficient $\beta$, 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 $\{\phi_i(s)\}_{i=1}^n$

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 (21).

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 $\theta$:

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

4: Output $\bar{V}(s) = \sum_{i=1}^n \theta\phi_i(s)$.

The reward is set to be $r(s) = \beta\cos(ks)^3 - \lambda s(-3k\cos(ks)^2\sin(ks))$ for the linear case and $r(s) = \beta\cos(ks)^3 - \lambda\sin^2(s)(-3k\cos(ks)^2\sin(ks))$ 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

$$\tilde{V}(s) = r(s)\Delta t + e^{-\beta\Delta t}\tilde{V}(e^{\lambda\Delta t}s), \tag{24}$$

and $i$-th order PhiBE as

$$\beta\hat{V}_i(s) = r(s) + \frac{1}{\Delta t}\left[\sum_{k=1}^{i} a_k(e^{\lambda k\Delta t}s - s)\right]\nabla\hat{V}_i(s), \tag{25}$$

respectively for $a_i$ defined in (12). 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 data we use are generated from $J$ different initial value $s_0 \sim$Unif$[-\pi, \pi]$, and each trajectory has $m = 4$ data, $\{s_0, \cdots, s_{(m-1)\Delta t}\}$. 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 $\tilde{A}$ defined as follows

$$\tilde{A} = \sum_{l=1}^{I} \sum_{j=0}^{m-1} \Phi(s_{j\Delta t}^l) \left[ \beta \Phi(s_{j\Delta t}^l) - \bar{\mu}_i(s_{j\Delta t}^l) \cdot \nabla \Phi(s_{j\Delta t}^l) - \frac{1}{2} \bar{\Sigma}_i(s_{j\Delta t}^l) : \nabla^2 \Phi(s_{j\Delta t}^l) \right]^\top \tag{26}$$

instead of $\bar{A}_i$.

In Figure 1, we compare the solution to the second-order PhiBE with the solution to BE, and the performance of LSTD with the proposed Algorithm 1 for linear dynamics (22) with $\lambda = 0.05$ and different $\Delta t, \beta, k$. 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 (25).

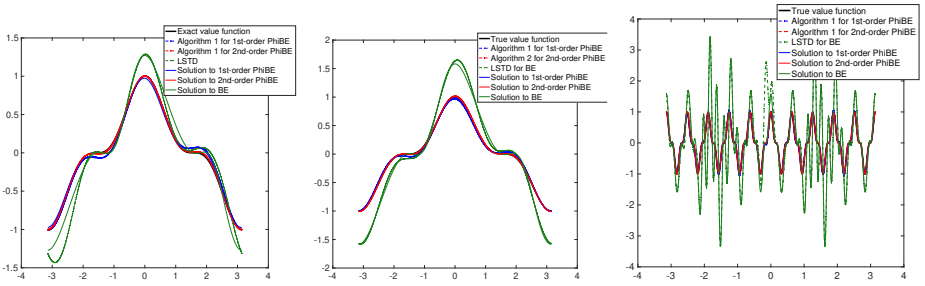
In Figure 2, we compare the solution to the first-order and second-order PhiBE with the solution to the BE, and the performance of LSTD with the proposed Algorithm 1 for nonlinear dynamics (23) for different $\Delta t, \beta, k, \lambda$.

In Figure 3, the distances of the PhiBE solution and the BE solution 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 as the amount of data increases are plotted. Here, the distance is measured using the $L^2$ norm

$$D(V, \hat{V}) = \sqrt{\int (V(s) - \hat{V}(s))^2 ds}. \tag{27}$$

In Figures 1 and 2, the solution to PhiBE is much closer to the true value function compared to the solution to BE in all the experiments. Especially, the second-order PhiBE solution is almost identical to the exact value function. Additionally, with access to only 40 or 400 data points, one can approximate the solutions to PhiBE very well. 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. Besides, unlike BE, the error increases when $\beta$ is too small or too large, twhile the error for PhiBE decays as $\beta$ 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.

In Figure 3/(a) and (b), one can observe that the solution for BE approximates the true solution in 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 performance of the algorithm does not improve, and the error in the BE solution stops decreasing when it reaches $10^{-1}$. This is because the error $\left\| \tilde{V} - V \right\| = O(\Delta t)$ 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 $O(\Delta t^i)$.

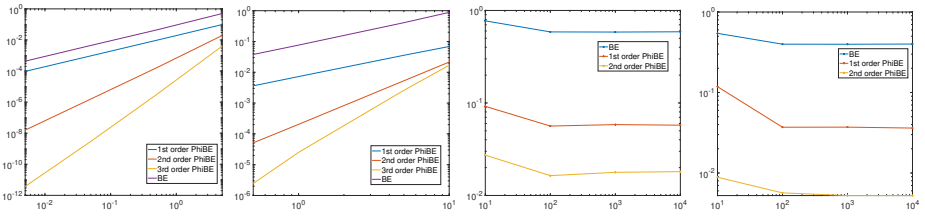



(a) Nonlinear dynamics (23) with $\Delta t = 5, \beta = 0.1, k = 1, \lambda = 0.1$

(b) Nonlinear dynamics (23) with $\Delta t = 0.1, \beta = 10, k = 10, \lambda = 5$

(c) Nonlinear dynamics (23) with $\Delta t = 0.1, \beta = 10, k = 10, \lambda = 2$

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




(a) Linear dynamics (23).

(b) Nonlinear dynamics (23)

(c) Linear dynamics (23).

(d) Nonlinear dynamics (23).

Figure 3: The $L^2$ error (27) of the solutions to PhiBE and BE with decreasing $\Delta t$ are plotted in the left two figures. The $L^2$ error (27) of the approximated solutions to PhiBE and BE with an increasing amount of data collected every $\Delta t = 5$ unit of time are plotted in the right two figures. LSTD is used to approximate the solution to BE, while Algorithm 1 is used to approximate the solution to PhiBE. 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,

$$ds(t) = \lambda s dt + \sigma dB_t,$$

with $\lambda = 0.05, \sigma = 1$. Here the reward is set to be $r(s) = \beta \cos^3(ks) - \lambda s(-3k\cos^2(ks)\sin(ks)) - \frac{1}{2}\sigma^2(6k^2\cos(s)\sin^2(ks) - 3k^2\cos^2(ks)\cos(ks))$, 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,

$$\begin{aligned}\beta\hat{V}(s) =& r(s) + \frac{1}{\Delta t}\sum_{k=1}^{i} a_k(e^{\lambda k\Delta t} - 1)s\nabla\hat{V}(s) \\ &+ \frac{1}{2\Delta t}\sum_{k=1}^{i} a_k\left[\frac{\sigma^2}{2\lambda}(e^{2\lambda k\Delta t} - 1) + (e^{\lambda k\Delta t} - 1)^2 s^2\right]\Delta\hat{V}(s);\end{aligned} \tag{28}$$

and BE as,

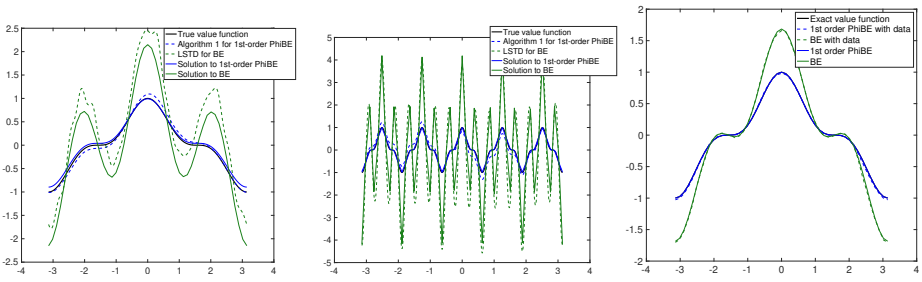
$$\begin{aligned}\tilde{V}(s) =& r(s)\Delta t + e^{-\beta\Delta t}\mathbb{E}\left[\tilde{V}(s_{t+1})|s_t = s\right] \\ =& r(s)\Delta t + e^{-\beta\Delta t}\int_{\mathbb{S}}\tilde{V}(s')\rho_{\Delta t}(s', s)ds'\end{aligned} \tag{29}$$

where

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

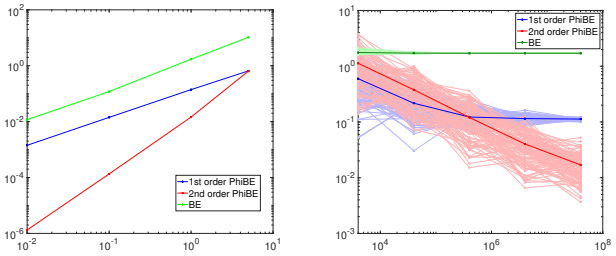
In Figure 4 , we compare the exact solution and approximated solution to PhiBE and BE, respectively, for different $\Delta t, \beta, k$. In Figure 5/(a), the decay of the error as $\Delta t \to 0$ for the exact solutions to PhiBE and BE is plotted. In Figure 5 /(b), the decay of the approximated solution to PhiBE and BE based on Algorithm 1 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 in 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 $10^{-1}$.




(a) $\Delta t = 1, \beta = 0.1, k = 1$ (b) $\Delta t = 0.1, \beta = 0.1, k = 5$ (c) $\Delta t = 0.1, \beta = 10, k = 1$

Figure 4: The solution to PhiBE and BE, when the discrete-time transition dynamics are given, are plotted in solid lines. The approximated solution to PhiBE based on Algorithm 1 and to BE based on LSTD, when discrete-time data is given, are plotted in dash lines. Both algorithms utilize the same data points.




(a) Model error w.r.t. the discretization time $\Delta t$.

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

Figure 5: The $L^2$ error (27) of the solutions to PhiBE and BE with decreasing $\Delta t$ are plotted in (a). The $L^2$ error (27) of the approximated solutions to PhiBE and BE with an increasing amount of data collected every $\Delta t = 1$ unit of time are plotted in (b). LSTD is used to approximate the solution to BE, while Algorithm 1 is used to approximate the solution to PhiBE. We set $\beta = 0.1, k = 1$ in both figures.

# 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

$$\partial_t \rho(s', t|s) = \nabla \cdot [\mu(s')\rho(s', t|s)] + \frac{1}{2} \sum_{i,j} \partial_{s_i} \partial_{s_j} [\Sigma_{ij}(s')\rho(s', t|s)]. \tag{30}$$

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

$$\begin{aligned} V(s) - \tilde{V}(s) &= \mathbb{E}\left[\sum_{i=0}^{\infty} \int_{\Delta t i}^{\Delta t(i+1)} f(t, s_t) - f(\Delta t i, s_{\Delta t i}) dt | s_0 = s\right] \\ &= \sum_{i=0}^{\infty} \int_{\Delta t i}^{\Delta t(i+1)} \left( \int_{\mathbb{S}} f(t, s')\rho(s', t|s) - f(\Delta t i, s')\rho(\Delta t i, s') ds' \right) dt \end{aligned} \tag{31}$$

Since

$$\begin{aligned} &\int_{\mathbb{S}} f(t, s')\rho(s', t|s) - f(\Delta t i, s')\rho(\Delta t i, s') ds' \\ &= \int_{\mathbb{S}} f(t, s')(\rho(s', t|s) - \rho(s', \Delta t i|s)) + (f(t, s') - f(\Delta t i, s'))\rho(s', \Delta t i|s) ds' \\ &= \int_{\mathbb{S}} f(t, s')\partial_t \rho(s', \xi_1|s)(t - \Delta t i) + \partial_t f(\xi_2, s')(t - \Delta t i)\rho(s', \Delta t i|s) ds', \quad \text{where } \xi_1, \xi_2 \in (\Delta t i, \Delta t(i+1)) \\ &= \int_{\mathbb{S}} \mathcal{L}_{\mu,\Sigma} f(t, s')\rho(s', \xi_1|s)(t - \Delta t i) ds' - \int_{\mathbb{S}} \beta e^{-\beta \xi_2} r(s')\rho(s', \Delta t i|s)(t - \Delta t i) ds' \\ &= \left( e^{-\beta t} \int_{\mathbb{S}} \mathcal{L}_{\mu,\Sigma} r(s')\rho(s', \xi_1|s) ds' - \beta e^{-\beta \xi_2} \int_{\mathbb{S}} r(s')\rho(s', \Delta t i|s) ds' \right) (t - \Delta t i) \end{aligned} \tag{32}$$

where the second equality is due to mean value theorem, and the third equality is obtained by inserting the equation (30) for $\rho(s', t|s)$ and integrating by parts.

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

Therefore, one has

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

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

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

one has,

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

□

## 6.2 Proof of Theorem 3.3

**Proof of Lemma 3.2**

*Proof.* 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,

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

where the last equality is due to the definition of $a$ in (13). Since

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

one has

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

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 $\left|\nabla^k \mu(s)\right| \le C_\mu$ for $\forall 0 \le k \le i$ are bounded, then $\left\|\mathcal{L}_\mu^i \mu(s_t)\right\|_{L^\infty}$ is bounded. This implies that

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

□

**Proof of Theorem 3.3**

*Proof.* Since $\hat{V}_i$ satisfies the PDE (10), by Feynman–Kac theorem, it is equivalently to write it as,

$$\hat{V}_i = \int_0^\infty e^{-\beta t} r(\hat{s}_t) dt$$

with

$$\frac{d}{dt}\hat{s}_t = \hat{\mu}(s_t).$$

Hence,

$$\begin{aligned}\left|V(s) - \hat{V}(s)\right| &= \left|\int_0^\infty e^{-\beta t}(r(s_t) - r(\hat{s}_t))dt\right| = \left|\int_0^\infty e^{-\beta t}\left(\int_{s_t}^{\hat{s}_t}\nabla r(s)ds\right)dt\right| \\ &\le \|\nabla r\|_{L^\infty}\int_0^\infty e^{-\beta t}\left|\hat{s}_t - s_t\right| dt\end{aligned} \tag{33}$$

where

$$\frac{d}{dt}s_t = \mu(s_t), \quad \frac{d}{dt}\hat{s}_t = \hat{\mu}_i(\hat{s}_t), \quad s_0 = \hat{s}_0 = s \tag{34}$$

Subtracting the two equations in (34) and multiplying it with $\hat{s}_t - s_t$ gives

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}|\hat{s}_t - s_t|^2 =& (\hat{\mu}_i(\hat{s}_t)) - \mu(s_t))(\hat{s}_t - s_t) \\ =& \left((\hat{\mu}_i(\hat{s}_t) - \mu(\hat{s}_t) + (\mu(\hat{s}_t) - \mu(s_t))\right)(\hat{s}_t - s_t) \\ \le& C_\mu \Delta t^i |\hat{s}_t - s_t| + \|\nabla\mu(s)\|_{L^\infty}|\hat{s}_t - s_t|^2,\end{aligned}$$

where Lemma 3.2 and mean value theorem is used in the last inequality with $C_\mu$ from Lemma 3.2.This implies

$$\frac{d}{dt}|\hat{s}_t - s_t| \le C_\mu \Delta t^i + \|\nabla\mu(s)\|_{L^\infty} |\hat{s}_t - s_t|$$

$$|\hat{s}_t - s_t| \le C_\mu \Delta t^i t + \int_0^t \|\nabla\mu(s)\|_{L^\infty} |\hat{s}_t - s_t| dt$$

$$|\hat{s}(t) - s(t)| \le (C_\mu \Delta t^i) t e^{\|\nabla\mu(s)\|_{L^\infty} t}$$

Inserting the above inequality back to (33) gives

$$\left\| V(s) - \hat{V}(s) \right\|_{L^\infty} \le \|\nabla r\|_{L^\infty} (C_\mu \Delta t^i) \int_0^\infty e^{-\beta t} t e^{\|\nabla\mu(s)\|_{L^\infty} t} dt = \frac{C_\mu \|\nabla r\|_{L^\infty}}{(\beta - \|\nabla\mu(s)\|_{L^\infty})^2} \Delta t^i$$

□

## 6.3 Proof of Theorem 3.4

*Proof.*

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

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

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

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

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

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,

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

On the other hand, 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,

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

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

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

Therefore,

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

□

## 6.4 Proof of Theorem 3.5

We first present the property of the operator $\mathcal{L}_{\mu,\Sigma}$ and $\partial_{s_i} \mathcal{L}_{\mu,\Sigma}$ in the following Lemma.

**Lemma 6.1.** *Under Assumption 1/(a), for the operator $\mathcal{L}_{\mu,\Sigma}$ defined in* (4)*, one has*

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

*Additionally, Assumption 1/(b) holds, one has,*

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

*where $C_{\nabla\mu,\nabla\Sigma}$ is defined in* (35) *depending on the first derivatives of $\mu, \Sigma$.*

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

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

where the last inequality is because of the definition of the stationary solution (18) and the positivity of the matrix $\Sigma(s)$.

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

$$
\begin{aligned}
&\sum_i \left\langle \partial_{s_i} \mathcal{L}_{\mu,\Sigma} V(s), V(s) \right\rangle_\rho \\
\leq& \sum_i \left( \left\langle \partial_{s_i} \mu \cdot \nabla V, \partial_{s_i} V \right\rangle_\rho + \frac{1}{2} \left\langle \partial_{s_i} \Sigma : \nabla^2 V, \partial_{s_i} V \right\rangle_\rho \right) - \frac{\lambda_{\min}}{2} \sum_i \left\| \nabla \partial_{s_i} V \right\|_\rho^2 \\
\leq& \sum_{i,k} \left\| \partial_{s_i} \mu_k \right\|_{L^\infty} \left\| \partial_{s_k} V \right\|_\rho \left\| \partial_{s_i} V \right\|_\rho + \frac{1}{2} \sum_{i,k,l} \left\| \partial_{s_i} \Sigma_{kl} \right\|_{L^\infty} \left\| \partial_{s_k} \partial_{s_l} V \right\|_\rho \left\| \partial_{s_i} V \right\|_\rho - \frac{\lambda_{\min}}{2} \sum_{k,i} \left\| \partial_{s_k} \partial_{s_i} V \right\|_\rho^2 \\
\leq& \frac{1}{2} \left[ \left( \max_k \sum_i \left\| \partial_{s_i} \mu_k \right\|_{L^\infty} \right) \sum_k \left\| \partial_{s_k} V \right\|_\rho^2 + \left( \max_i \sum_k \left\| \partial_{s_i} \mu_k \right\|_{L^\infty} \right) \sum_i \left\| \partial_{s_i} V \right\|_\rho^2 \right. \\
&\left. + \frac{1}{2} \left( \max_{k,l} \sum_i \left\| \partial_{s_i} \Sigma_{kl} \right\|_{L^\infty} \right) \sum_{k,l} \left\| \partial_{s_k} \partial_{s_l} V \right\|_\rho^2 + \frac{1}{2} \left( \max_i \sum_{k,l} \left\| \partial_{s_i} \Sigma_{kl} \right\|_{L^\infty} \right) \sum_i \left\| \partial_{s_i} V \right\|_\rho^2 \right] \\
&- \frac{\lambda_{\min}}{2} \sum_{k,i} \left\| \partial_{s_k} \partial_{s_i} V \right\|_\rho^2 \\
\leq& \frac{C_{\nabla\mu,\nabla\Sigma}}{2} \left\| \nabla V \right\|_\rho^2 ,
\end{aligned}
$$

where Assumption 1/(b) is applied in the last inequality, and

$$
C_{\nabla\mu,\nabla\Sigma} = \max_k \sum_i \left\| \partial_{s_i} \mu_k \right\|_{L^\infty} + \max_i \sum_k \left\| \partial_{s_i} \mu_k \right\|_{L^\infty} + \frac{1}{2} \max_i \sum_{k,l} \left\| \partial_{s_i} \Sigma_{kl} \right\|_{L^\infty} . \tag{35}
$$

□

**Proof of Theorem 3.5** Now we are ready to prove Theorem 3.5.

*Proof.* By (31), one has

$$\begin{aligned}&\left\|V(s)-\tilde{V}(s)\right\|_\rho\\ \leq&\sum_{i=0}^{\infty}\sqrt{\Delta t\int_{\Delta ti}^{\Delta t(i+1)}\left\|\int_{\mathbb{S}}f(t,s')\rho(s',t|s)-f(\Delta ti,s')\rho(\Delta ti,s')ds'\right\|_\rho^2 dt},\end{aligned}\tag{36}$$

where the Jensen's inequality is used. By (32), one has,

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

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

$$\partial_t p_i(s,t)=\mathcal{L}_{\mu,\Sigma}p_i(s,t)$$

with initial data

$$p_1(0,s)=\mathcal{L}_{\mu,\Sigma}r(s),\quad p_2(0,s)=r(s).$$

By Lemma (6.1), one has

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

which implies,

$$\left\|p_i(t)\right\|_\rho\leq\left\|p_i(0)\right\|_\rho$$

Therefore, one has

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

Inserting it back to (36) yields,

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

which completes the proof. □

## 6.5 Proof of Theorem 3.6

Lemma 6.2 and Lemma 6.3 are both related to the true value function $V$ satisfying (15).

**Lemma 6.2.** *For $V(s)$ satisfying (15), one has*

$$\|V\|_{L^\infty} \leq \frac{1}{\beta} \|r\|_{L^\infty}$$

*Proof.* By the definition of $V$ in (1), one has,

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

□

**Lemma 6.3.** *Under Assumption 1/(a), (b), for $V(s)$ satisfying (15), one has*

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

*where $C_{r,\nabla\mu,\nabla\Sigma}$ is a constant defined in (38) depending on $\lambda_{min}, r(s)$ and the first derivatives of $r(s), \mu(s), \Sigma(s)$.*

*Proof.* Based on Lemma 6.1, one has

$$\beta \|V\|_\rho^2 - \langle r(s), V(s) \rangle_\rho \leq -\frac{\lambda_{\min}}{2} \sum_i \|\partial_{s_i} V(s)\|^2$$

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

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

$$\frac{C_{\nabla\mu,\nabla\Sigma}\beta}{\lambda_{\min}} \|V\|^2 + \beta \|\nabla V(s)\|_\rho^2 \leq C_{r,\nabla\mu,\nabla\Sigma} \sqrt{\|V\|_\rho + \|\nabla V\|_\rho} \tag{37}$$

where

$$C_{r,\nabla\mu,\nabla\Sigma} = \sqrt{2} \max\left\{ \frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}} \|r\|_\rho, \|\nabla r\|_\rho \right\} \tag{38}$$

Therefore, one has

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

As $C_{\nabla\mu,\nabla\Sigma}$ is an upper bound and $\lambda_{\min}$ is a lower bound, so one could always assume that $C_{\nabla\mu,\nabla\Sigma} \geq 1$ and $\lambda_{\min} \leq 1$, which implies,

$$\sqrt{\|V\|^2 + \|\nabla V(s)\|^2} \leq \frac{C_{r,\nabla\mu,\nabla\Sigma}}{\beta}.$$

□

Lemma 6.4 and Lemma 6.5 are related to the distance between the true dynamics and the approximated dynamics $\mu - \hat{\mu}_i$, $\Sigma - \hat{\Sigma}$ and the true operator and the approximated operator $\mathcal{L}_{\mu,\Sigma} - \mathcal{L}_{\hat{\mu},\hat{\Sigma}}$.

**Lemma 6.4.** *Under Assumption 1, for $\hat{\mu}(s), \hat{\Sigma}(s)$ defined in* (17)*, 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*

$$\max_k \sqrt{\sum_l \left\|\partial_{s_l}(\hat{\Sigma}_i - \Sigma)_{kl}\right\|_\rho^2} \le L_{\Sigma,\rho} \Delta t^i, \tag{39}$$

*where $L_\mu, L_\Sigma, L_{\Sigma,\rho}$ are constants depending on $\mu, \Sigma, i$ defined in* (41)*,* (44)*,* (48)*, respectively.*

*Proof.* Therefore, $\hat{\mu}_i$ can be written as

$$\hat{\mu}_i(s) = \frac{1}{\Delta t} \sum_{j=1}^{i} a_j \left( \int s' \rho(s', j\Delta t|s) ds' - s \right), \tag{40}$$

where $\rho(s', t|s)$ is defined in (30). 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 (40) yields,

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

The first part can be written as

$$\begin{aligned} I =& \frac{1}{\Delta t} \left( \sum_{j=1}^{i} a_j \right) \int_{\mathbb{S}} s\rho(s', 0|s) ds' + \int_{\mathbb{S}} s' \partial_t \rho(s', 0|s) ds' - \frac{1}{\Delta t} \left( \sum_{j=1}^{i} a_j \right) s \\ =& \int_{\mathbb{S}} s' \left( \nabla \cdot [\mu(s')\rho(s', 0)|s] + \frac{1}{2} \partial_{s_i} \partial_{s_j} [\Sigma_{ij}(s')\rho(s', 0|s)] \right) ds' \\ =& \int_{\mathbb{S}} \mathcal{L}_{\mu,\Sigma}(s')\rho(s', 0|s) ds' \end{aligned}$$

Therefore

$$I = \int_{\mathbb{S}} \mu(s)\rho(s', 0|s)ds' = \mu(s).$$

which implies that

$$\begin{aligned}
&\|\hat{\mu}_i(s) - \mu(s)\|_{L^\infty} = II \\
=&\frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j \int_0^{j\Delta t} \left(\int_{\mathbb{S}} \mathcal{L}^{i+1}_{\mu,\Sigma}(s')\rho(s', t|s)ds'\right) t^i dt \\
=&\frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j \int_0^{j\Delta t} \left(\int_{\mathbb{S}} \mathcal{L}^{i}_{\mu,\Sigma}(\mu(s))\rho(s', t|s)ds'\right) t^i dt \\
\leq&\frac{\left\|\mathcal{L}^i_{\mu,\Sigma}\mu(s)\right\|_{L^\infty}}{\Delta t i!}\sum_{j=1}^{i}|a_j| \int_0^{j\Delta t} t^i dt \\
\leq&L_\mu \Delta t^i.
\end{aligned}$$

where

$$L_\mu = \frac{\sum_{j=1}^{i}|a_j|j^{i+1}}{(i+1)!}\left\|\mathcal{L}^i_{\mu,\Sigma}\mu(s)\right\|_{L^\infty}. \tag{41}$$

To prove the second inequality in the lemma, first note that

$$\begin{aligned}
\left(\hat{\Sigma}_i(s)\right)_{kl} =&\frac{1}{\Delta t}\mathbb{E}\left[\sum_{j=1}^{i} a_j (s_{j\Delta t} - s_0)_k (s_{j\Delta t} - s_0)_l | s_0 = s\right] \\
=&\frac{1}{\Delta t}\int_{\mathbb{S}}\sum_{j=1}^{i}(s'-s)_k(s'-s)_l\rho(s', j\Delta t|s)ds' \\
=&\frac{1}{\Delta t}\sum_{j=1}^{i} a_j \int_{\mathbb{S}}(s'-s)_k(s'-s)_l\rho(s', 0|s)ds' + \int_{\mathbb{S}}(s'-s)_k(s'-s)_l\partial_t\rho(s', 0|s)ds' \\
&+\frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j \int_0^{j\Delta t}\left(\int_{\mathbb{S}}(s'-s)_k(s'-s)_l\partial_t^{i+1}\rho(s', t|s)ds'\right)t^i dt \\
=&\Sigma_{kl}(s) + \frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j \int_0^{j\Delta t}\left(\int_{\mathbb{S}}\mathcal{L}^i_{\mu,\Sigma}\left(\mu(s')_k(s'-s)_l + \mu(s')_l(s'-s)_k + \frac{1}{2}\Sigma(s')_{kl}\right)\rho(s', t|s)ds'\right)t^i d
\end{aligned} \tag{42}$$

Note that

$$\begin{aligned}
&\mathcal{L}^i_{\mu,\Sigma}\left(\mu(s')_k(s'-s)_l + \mu(s')_l(s'-s)_k + \frac{1}{2}\Sigma(s')_{kl}\right) \\
=&h(s')_{kl} + f(s')_k(s'-s)_l + f(s')_l(s'-s)_k
\end{aligned}$$

where

$$\begin{aligned} h(s')_{kl} =& \frac{1}{2}\mathcal{L}^i_{\mu,\Sigma}\Sigma(s')_{kl} + \mathcal{L}^i_{\mu,\Sigma}\left[\mu(s')_k(s'-s)_l + \mu(s')_l(s'-s)_k\right] \\ & - \mathcal{L}^i_{\mu,\Sigma}\mu(s')_k(s'-s)_l - \mathcal{L}^i_{\mu,\Sigma}\mu(s')_l(s'-s)_k, \\ f(s')_k =& \mathcal{L}^i_{\mu,\Sigma}\mu(s')_k. \end{aligned} \tag{43}$$

Note that $h(s)$ is a function that only depends on $\Sigma, \mu$ and is independent of $(s'-s)$. Thus

$$\begin{aligned} & \left| \int_{\mathbb{S}} \mathcal{L}^i_{\mu,\Sigma} \left( \mu(s')_k(s'-s)_l + \mu(s')_l(s'-s)_k + \frac{1}{2}\Sigma(s')_{kl} \right) \rho(s',t|s)ds' \right| \\ \leq & \|h(s)_{kl}\|_{L^\infty} + \|f(s)_k\|_{L^\infty} \int_{\mathbb{S}} |(s'-s)_l|\rho(s',t|s)ds' + \|f(s)_l\|_{L^\infty} \int_{\mathbb{S}} |(s'-s)_k|\rho(s',t|s)ds'. \end{aligned}$$

Since

$$\int_S (|(s'-s)_l|\rho(s',t|s)ds' \leq \sqrt{\int_S |(s'-s)_l^2\rho(s',t|s)ds'}$$

and

$$\begin{aligned} & \partial_t \int_{\mathbb{S}} (s'-s)_l^2 \rho(s',t|s)ds' \\ = & \int (s'-s)_l^2 \nabla(\mu(s')\rho(t,s'|s))ds' + \frac{1}{2}\sum_{i,j}\int (s'-s)_l^2 \partial_{s_i}\partial_{s_j}(\Sigma(s')_{ij}\rho(t,s'|s))ds' \\ = & -\int 2(s'-s)_l\mu(s')_l\rho(t,s'|s)ds + \frac{1}{2}\int \Sigma(s')_{ll}\rho(t,s'|s)ds' \\ = & \int (s'-s)_l^2\rho(t,s'|s)ds' + \|\mu(s)_l\|^2_{L^\infty} + \frac{1}{2}\|\Sigma(s)_{ll}\|_{L^\infty} \\ & \int_{\mathbb{S}} (s'-s)_l^2\rho(s',t|s)ds' \leq \left( \|\mu(s)_l\|^2_{L^\infty} + \frac{1}{2}\|\Sigma(s)_{ll}\|_{L^\infty} \right) te^t, \end{aligned}$$

where the last inequality is due to $\int_{\mathbb{S}}(s'-s)_l^2\rho(s',0|s)ds' = 0$ and the Grownwall inequality. Hence, one has,

$$\begin{aligned} & \int_0^{j\Delta t}\int_{\mathbb{S}} |(s'-s)_l|\rho(s',t)ds't^i dt \\ \leq & \left( \|\mu(s)_l\|^2_{L^\infty} + \frac{1}{2}\|\Sigma(s)_{ll}\|_{L^\infty} \right) \int_0^{j\Delta t} e^{t/2}t^{1/2+i}dt \\ \leq & \left( \|\mu(s)_l\|^2_{L^\infty} + \frac{1}{2}\|\Sigma(s)_{ll}\|_{L^\infty} \right) (j\Delta t)^{i+3/2} + o((j\Delta t)^{i+3/2}). \end{aligned}$$

Therefore,

$$\begin{aligned} & \int_0^{j\Delta t}\int_{\mathbb{S}} \mathcal{L}^i_{\mu,\Sigma} \left( \mu(s')_k(s'-s)_l + \mu(s')_l(s'-s)_k + \frac{1}{2}\Sigma(s')_{kl} \right) \rho(s',t|s)ds't^i dt \\ \leq & \frac{1}{i+1}\|h(s)_{kl}\|_{L^\infty}(j\Delta t)^{i+1} + O(\Delta t^{i+3/2}), \end{aligned}$$

which implies

$$\left|\left(\hat{\Sigma}_i(s)\right)_{kl} - \Sigma_{kl}(s)\right| \leq L_\Sigma \Delta t^i + o(\Delta t^i),$$

where

$$L_\Sigma = \frac{\sum_{j=1}^{i} |a_j| j^{i+1}}{(i+1)!} \left\|h(s)_{kl}\right\|_{L^\infty}, \quad \text{with } h(s)_{kl} \text{ defined in (43).} \tag{44}$$

For the estimate of $\left\|\partial_{s_l}\left(\Sigma(s)\right)_{kl} - \partial_{s_l}\left(\hat{\Sigma}_i(s)\right)_{kl}\right\|_\rho$, first one can obtain the following equation similar to (42),

$$\partial_{s_l}\left(\hat{\Sigma}_i(s)\right)_{kl} = \partial_{s_l}\Sigma_{kl}(s) + \frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j \int_0^{j\Delta t} \partial_{s_l} p(s,t) t^i dt$$

where

$$p(s,t) = \int_{\mathbb{S}} h(s')_{kl} + f(s')_k (s'-s)_l + f(s')_l (s'-s)_k ds',$$

with $h, f$ defined in (43). Therefore,

$$\begin{aligned}
&\sum_l \left\|\partial_{s_l}\left(\Sigma(s)\right)_{kl} - \partial_{s_l}\left(\hat{\Sigma}_i(s)\right)_{kl}\right\|_\rho^2 \\
=&\sum_l \left(\frac{1}{\Delta t i!}\right)^2 \int_{\mathbb{S}} \left(\sum_{j=1}^{i} a_j \int_0^{j\Delta t} \partial_{s_l} p(s,t) t^i dt\right)^2 \rho(s) ds \\
\leq&\sum_l \left(\frac{1}{\Delta t i!}\right)^2 i \sum_{j=1}^{i} a_j^2 \int_{\mathbb{S}} \left(\int_0^{j\Delta t} \partial_{s_l} p(s,t) t^i dt\right)^2 \rho(s) ds \\
\leq&\left(\frac{1}{\Delta t i!}\right)^2 i \sum_{j=1}^{i} a_j^2 j\Delta t \int_0^{j\Delta t} \left(\sum_l \int_{\mathbb{S}} \partial_{s_l} p(s,t)^2 \rho(s) ds\right) t^{2i} dt \\
\leq&\left(\frac{1}{\Delta t i!}\right)^2 i \sum_{j=1}^{i} a_j^2 j\Delta t \int_0^{j\Delta t} \left\|\nabla p(t)\right\|_\rho^2 t^{2i} dt
\end{aligned} \tag{45}$$

Next, we will estimate $\left\|\nabla p(t)\right\|_\rho^2$. Note that $p(s,t)$ satisfies the following forward Kolmogorov equation [16],

$$\partial_t p(s,t) = \mathcal{L}_{\mu,\Sigma} p(s,t), \quad \text{with } p(s,0) = h(s)_{kl}. \tag{46}$$

By Lemma 6.1, one has

$$\begin{aligned}
&\frac{1}{2}\partial_t \left\|p\right\|_\rho^2 = \left\langle \mathcal{L}_{\mu,\Sigma} p, p\right\rangle_\rho \leq -\frac{\lambda_{\min}}{2}\left\|\nabla p\right\|_\rho^2 \\
&\frac{1}{2}\partial_t \left\|\nabla p\right\|_\rho^2 = \left\langle \nabla \mathcal{L}_{\mu,\Sigma} p, \nabla p\right\rangle_\rho \leq \frac{C_{\nabla\mu,\nabla\Sigma}}{2}\left\|\nabla p\right\|_\rho^2
\end{aligned} \tag{47}$$

Multiply $\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}$ to the first equation gives

$$\partial_t \left( \frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}} \|p\|_\rho^2 + \|\nabla p\|_\rho^2 \right) \leq 0,$$

which implies that

$$\|\nabla p(t)\|_\rho^2 \leq \frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}} \|p(0)\|_\rho^2 + \|\nabla p(0)\|_\rho^2 = \frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}} \|h_{kl}\|_\rho^2 + \|\nabla h_{kl}\|_\rho^2 .$$

Inserting the above inequality back to (45), one has,

$$\begin{aligned}
&\sum_l \left\| \partial_{s_l} \left(\Sigma(s)\right)_{kl} - \partial_{s_l} \left(\hat{\Sigma}_i(s)\right)_{kl} \right\|_\rho^2 \\
\leq& \left(\frac{1}{\Delta t i!}\right)^2 \left( \frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}} \|h_{kl}\|_\rho^2 + \|\nabla h_{kl}\|_\rho^2 \right) i \sum_{j=1}^{i} a_j^2 j\Delta t \int_0^{j\Delta t} t^{2i} dt \\
=& \left(\frac{1}{\Delta t i!}\right)^2 \left( \frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}} \|h_{kl}\|_\rho^2 + \|\nabla h_{kl}\|_\rho^2 \right) i \sum_{j=1}^{i} a_j^2 \frac{j^{2i+2}}{2i+1} \Delta t^{2i+2},
\end{aligned}$$

which implies

$$\sqrt{\sum_l \left\| \partial_{s_l} \left(\Sigma(s)\right)_{kl} - \partial_{s_l} \left(\hat{\Sigma}_i(s)\right)_{kl} \right\|_\rho^2} \leq L_{\Sigma,\rho} \Delta t^i$$

where

$$L_{\Sigma,\rho} = \frac{\sum_{j=1}^{i} |a_j| j^{i+1}}{(i+1)!} (i+1) \left( \sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}} \|h_{kl}\|_\rho + \|\nabla h_{kl}\|_\rho \right), \tag{48}$$

with $h(s)_{kl}$ defined in (43).

□

**Lemma 6.5.** *Under Assumption 1, for $\hat{\mu}_i(s), \hat{\Sigma}_i(s)$ defined in* (17)*, one has*

$$\left\langle (L_{\mu,\Sigma} - L_{\hat{\mu}_i,\hat{\Sigma}_i}) f, g \right\rangle_\rho \leq D_{\mu,\Sigma,\lambda_{min}} \Delta t^i \|f\|_{L^\infty} \|\nabla g\|_\rho + \frac{dL_\Sigma}{2} \Delta t^i \|\nabla f\|_\rho \|\nabla g\|_\rho + dL_\mu \Delta t^i \|\nabla f\|_\rho \|g\|_\rho$$

*where $L_\mu, L_\Sigma$ are defined in* (41) *and* (44)*, respectively, and $D_{\mu,\Sigma,\lambda_{min}}$ is a constant depending on $\mu, \Sigma, \lambda_{min}$ and the dimension $d$ of $\mathbb{S}$, which are defined in* (49)*.*

*Proof.* By extending the the inner product, one has

$$\begin{aligned}
&\left\langle (\mathcal{L}_{\mu,\Sigma} - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i}) f, g \right\rangle_\rho \\
=& \sum_k \int (\mu - \hat{\mu}_i)_k \partial_{s_k} f g \rho ds - \frac{1}{2} \sum_{k,l} \int \partial_{s_l} \left[ (\Sigma - \hat{\Sigma}_i)_{kl} f \rho \right] \partial_{s_k} g ds \\
\leq& dL_\mu \Delta t^i \|\nabla f\|_\rho \|g\|_\rho - \frac{1}{2} \sum_{k,l} \int \partial_{s_l} \left[ (\Sigma - \hat{\Sigma}_i)_{kl} f \rho \right] \partial_{s_k} g ds.
\end{aligned}$$

In addition, one has,

$$
\begin{aligned}
&\sum_{k,l}\int \partial_{s_l}\left[(\hat{\Sigma}-\Sigma)_{kl} f\rho\right]\partial_{s_k} g ds \\
=&\sum_{k,l}\int \partial_{s_l}(\hat{\Sigma}-\Sigma)_{kl} f\rho\partial_{s_k} g ds + \int (\hat{\Sigma}-\Sigma)_{kl}\partial_{s_l} f\rho\partial_{s_k} g ds + \int (\hat{\Sigma}-\Sigma)_{kl} f\partial_{s_l}\rho\partial_{s_k} g ds \\
\leq& \|f\|_{L^\infty}\left(\max_k \sum_l \left\|\partial_{s_l}(\hat{\Sigma}-\Sigma)_{kl}\right\|_\rho\right)\left(\sum_k \|\partial_{s_k} g\|_\rho\right) + \lambda_{\max}\|\nabla f\|_\rho \|\nabla g\|_\rho \\
&+ \max_{k,l}\left\|(\hat{\Sigma}-\Sigma)_{kl}\right\|_{L^\infty}\|f\|_{L^\infty}\left(\sum_l \left\|\frac{\partial_{s_l}\rho}{\rho}\right\|_\rho\right)\left(\sum_k \|\partial_{s_k} g\|_\rho\right) \\
\leq& d\|f\|_{L^\infty}\left(\max_k \sqrt{\sum_l \left\|\partial_{s_l}(\hat{\Sigma}-\Sigma)_{kl}\right\|_\rho^2}\right)\|\nabla g\|_\rho + \lambda_{\max}\|\nabla f\|_\rho\|\nabla g\|_\rho \\
&+ d\max_{k,l}\left\|(\hat{\Sigma}-\Sigma)_{kl}\right\|_{L^\infty}\|f\|_{L^\infty}\left\|\frac{\nabla\rho}{\rho}\right\|_\rho\|\nabla g\|_\rho
\end{aligned}
$$

where

$$\lambda_{\max} = \text{the maximum absolute eigenvalue of } (\hat{\Sigma}-\Sigma).$$

By the Gershgorin circle theorem, one has

$$\lambda_{\max} \leq \max_k \sum_l |(\hat{\Sigma}-\Sigma)_{kl}| \leq dL_\Sigma \Delta t^i.$$

By applying Theorem 1.1 of [2], one has,

$$\left\|\frac{\nabla\rho}{\rho}\right\|_\rho \leq \frac{1}{\lambda_{\min}}\|\mu+\nabla\cdot\Sigma\|_\rho$$

Therefore, one has

$$
\begin{aligned}
&\sum_{k,l}\int \partial_{s_l}\left[(\hat{\Sigma}-\Sigma)_{kl} e\rho\right]\partial_{s_k} e ds \\
\leq& \left(dL_{\Sigma,\rho}\|f\|_{L^\infty}\|\nabla g\|_\rho + dL_\Sigma\|\nabla f\|_\rho\|\nabla g\|_\rho + \frac{dL_\Sigma}{\lambda_{\min}}\|\mu+\nabla\cdot\Sigma\|_\rho\|f\|_{L^\infty}\|\nabla g\|_\rho\right)\Delta t^i \\
=& \left[2D_{\mu,\Sigma,\lambda_{\min}}\|f\|_{L^\infty}\|\nabla g\|_\rho + dL_\Sigma\|\nabla f\|_\rho\|\nabla g\|_\rho\right]\Delta t^i
\end{aligned}
$$

where

$$D_{\mu,\Sigma,\lambda_{\min}} = \frac{d}{2}L_{\Sigma,\rho} + \frac{dL_\Sigma}{2\lambda_{\min}}\|\mu+\nabla\cdot\Sigma\|_\rho. \tag{49}$$

where $L_{\Sigma,\rho}, L_\Sigma$ are defined in (48) and (44). □

**Proof of Theorem 3.6**

*Proof.* First note that $V, \hat{V}$ satisfies,

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

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,

$$\begin{aligned}
&\beta \|e\|_\rho^2 = \langle \mathcal{L}_{\mu,\Sigma} e, e \rangle_\rho + \left\langle (\mathcal{L}_{\hat{\mu},\hat{\Sigma}} - \mathcal{L}_{\mu,\Sigma})\hat{V}, e \right\rangle_\rho \\
\leq & -\frac{\lambda_{\min}}{2} \|\nabla e\|_\rho^2 + d D_{\mu,\Sigma,\lambda_{\min}} \Delta t^i \left\|\hat{V}\right\|_{L^\infty} \|\nabla e\|_\rho + \frac{d}{2} L_\Sigma \Delta t^i \left\|\nabla \hat{V}\right\|_\rho \|\nabla e\|_\rho \\
&+ \frac{d}{2} L_\mu \Delta t^i \left\|\nabla \hat{V}\right\|_\rho \|e\|_\rho \\
\leq & -(\frac{\lambda_{\min}}{2} - \frac{d}{2} L_\Sigma \Delta t^i) \|\nabla e\|_\rho^2 + \left( \frac{d D_{\mu,\Sigma,\lambda_{\min}} \Delta t^i}{\beta} \|r\|_{L^\infty} + \frac{d}{2} L_\Sigma \Delta t^i \|\nabla V\|_\rho \right) \|\nabla e\|_\rho \\
&+ \frac{d}{2} L_\mu \Delta t^i \|\nabla e\|_\rho \|e\|_\rho + \frac{d}{2} L_\mu \Delta t^i \|\nabla V\|_\rho \|e\|_\rho \\
\leq & -\left( \frac{\lambda_{\min}}{2} - \frac{d}{2} L_\Sigma \Delta t^i - \frac{d^2}{4\beta} L_\mu^2 \Delta t^{2i} \right) \|\nabla e\|_\rho^2 + \frac{\beta}{2} \|e\|_\rho^2 + \frac{d^2}{4\beta} L_\mu^2 \Delta t^{2i} \|\nabla V\|_\rho^2 \\
&+ \left( \frac{d}{\beta} D_{\mu,\Sigma,\lambda_{\min}} \|r\|_{L^\infty} + \frac{d}{2\beta} L_\Sigma C_{r,\nabla\mu,\nabla\Sigma} \right) \Delta t^i \|\nabla e\|_\rho \\
\leq & -\left( \frac{\lambda_{\min}}{2} - \frac{d}{2} L_\Sigma \Delta t^i - \frac{d^2}{4\beta} L_\mu^2 \Delta t^{2i} \right) \|\nabla e\|_\rho^2 + \sqrt{\lambda_{\min}} C_1 \|\nabla e\|_\rho + \frac{\beta}{2} \|e\|_\rho^2 + C_2^2,
\end{aligned} \tag{50}$$

where

$$C_1 = \frac{d\Delta t^i}{\beta\sqrt{\lambda_{\min}}} \left( D_{\mu,\Sigma,\lambda_{\min}} \|r\|_{L^\infty} + \frac{1}{2} L_\Sigma C_{r,\nabla\mu,\nabla\Sigma} \right), \quad C_2 = \frac{d\Delta t^i}{2\beta^{3/2}} L_\mu C_{r,\nabla\mu,\nabla\Sigma}$$

with $D_{\mu,\Sigma,\lambda_{\min}}$ defined in (49), $L_\mu, L_\Sigma$ defined in (41), (44), and $C_{r,\nabla\mu,\nabla\Sigma}$ defined in (38). Here the first inequality applies Lemma 6.5, the second inequality uses $\nabla\hat{V} = \nabla V - \nabla e$ and $\left\|\hat{V}\right\|_{L^\infty} \leq \frac{1}{\beta} \|r\|_{L^\infty}$, and the last inequality applies Lemma 6.3. Under the assumption that

$$\frac{d}{2} L_\Sigma \Delta t^i + \frac{d^2}{4\beta} L_\mu^2 \Delta t^{2i} \leq \frac{\lambda_{\min}}{4}$$

the RHS of (50) can be bounded by

$$\text{RHS of (50)} \leq C_1^2 + C_2^2 + \frac{\beta}{2} \|e\|_\rho^2$$

which implies

$$\|e\|_\rho \leq \sqrt{\frac{2}{\beta}}(C_1 + C_2) = \frac{\sqrt{\beta}C_{r,\mu,\Sigma,\lambda_{\min}} + C_{r,\mu,\Sigma}}{\beta^2}\Delta t^i.$$

where

$$C_{r,\mu,\Sigma,\lambda_{\min}} = \frac{d\sqrt{2}}{\sqrt{\lambda_{\min}}}\left(D_{\mu,\Sigma,\lambda_{\min}}\|r\|_{L^\infty} + \frac{1}{2}L_\Sigma C_{r,\nabla\mu,\nabla\Sigma}\right), \quad C_{r,\mu,\Sigma} = \frac{d}{\sqrt{2}}L_\mu C_{r,\nabla\mu,\nabla\Sigma} \tag{51}$$

with $D_{\mu,\Sigma,\lambda_{\min}}$ defined in (49), $L_\mu, L_\Sigma$ defined in (41), (44), and $C_{r,\nabla\mu,\nabla\Sigma}$ defined in (38).

□

# 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] 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.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

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

[15] 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.

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

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

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.