Journal of Machine Learning Research 27 (2026) 1-68 Submitted 5/24; Revised 9/26; Published 9/26

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

Yuhua Zhu YUHUAZHU@UCLA.EDU
Department of Statistics and Data Science
University of California, Los Angeles
8125 Math Sciences Building, Box 951554
Los Angeles, CA 90095-1554, USA

Editor: Maxim Raginsky

Abstract

In this paper, we study policy evaluation in continuous-time reinforcement learning (RL), where the state follows an unknown stochastic differential equation (SDE) but only discrete-time data are available. We first highlight that the discrete-time Bellman equation (BE) is not always a reliable approximation to the true value function because it ignores the underlying continuous-time structure. We then introduce a new Bellman equation, PhiBE, which integrates the discrete-time information into a continuous-time PDE formulation. By leveraging the smooth structure of the underlying dynamics, PhiBE provides a provably more accurate approximation to the true value function, especially in scenarios where the underlying dynamics change slowly or the reward oscillates. Moreover, we extend PhiBE to higher orders, providing increasingly accurate approximations.

We further develop a model-free algorithm for PhiBE under linear function approximation and establish its convergence under model misspecification, together with finite-sample guarantees. Directly applying classical discrete-time LSTD bounds to the continuous-time RL setting yields a model-misspecification bound with an $O(\Delta t^{-1/2})$ stability factor and a sample-complexity bound of order $O(\Delta t^{-2})$. By exploiting the smooth structure of the underlying dynamics, our analysis instead gives a model-misspecification bound with a stability factor independent of Δt and a sample-complexity bound of order $O(\Delta t^{-1})$. Moreover, we identify a fundamental trade-off between discretization error and sample error that is intrinsic to continuous-time policy evaluation: finer time discretization reduces bias but amplifies variance, so excessively frequent sampling does not necessarily improve performance. This is an insight that does not arise in classical discrete-time RL analyses. Numerical experiments are provided to validate the theoretical guarantees we propose.

Keywords: continuous-time reinforcement learning, continuous-time policy evaluation, optimal control with unknown dynamics, model-free reinforcement learning, higher-order approximation

1. Introduction

Reinforcement learning (RL) has achieved remarkable success in *digital environments*, with applications such as AlphaGo (Silver et al., 2016), strategic gameplay (Mnih et al., 2015), and fine-tuning large language models (Ziegler et al., 2019). In these applications, the system state changes only after an action is taken. In contrast, in many systems in the *physical world*, the state evolves continuously in time, regardless of whether actions are

©2026 Yuhua Zhu.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/24-0734.html.

taken in continuous or discrete time. Although these systems are fundamentally continuous in nature, the available data are typically collected at discrete time points. In addition, the time difference between data points can be irregular, sparse, and outside our control. Examples arise in the following applications:

- Health care: for instance, in diabetes treatment, a patient's glucose level changes continuously, but measurements are recorded only when tests are performed (Guo et al., 2023; Amos et al., 2018; Romero et al., 2024; Emerson et al., 2023; Cobelli et al., 2009; Battelino et al., 2019);
- Robotic control (Karimi et al., 2023; Kober et al., 2013; Siciliano and Villani, 1999) and autonomous driving (Sciarretta et al., 2004): the environment evolves continuously, but sensor data are collected at discrete intervals;
- Financial markets (Merton, 1975; Moody and Saffell, 2001; Wang and Zhou, 2020; Whittle, 1981): asset prices move continuously, but trades and quotes occur at discrete times;
- Plasma fusion reactors (Degrave et al., 2022): the plasma evolves in continuous time, while all experimental data are recorded as discrete-time samples via sensors.

This paper directs its focus toward addressing continuous-time RL problems (Wang et al., 2020; Doya, 2000; Basei et al., 2022; Jia and Zhou, 2022b, 2023) where the underlying dynamics follow unknown stochastic differential equations (SDEs) but only discrete-time observations are available. Since one can divide the RL problem into policy evaluation (PE) and policy update (Sutton et al., 1999; Konda and Tsitsiklis, 1999; Tallec et al., 2019; Jia and Zhou, 2022b), we first concentrate on the continuous-time PE problem in this paper.

Currently, there are two main approaches to addressing this challenge. One approach is to learn the continuous-time dynamics from discrete-time data and formulate the problem as a policy-evaluation problem with known dynamics (Kamalapurkar et al., 2016; Della Santina et al., 2023; Lee and Sutton, 2021; Vamvoudakis and Lewis, 2010). The key advantage of this approach is that it preserves the continuous-time nature of the problem, enhancing the stability and interpretability of the resulting algorithms. However, identifying continuous-time dynamics from discrete-time data is often challenging and, in general, ill-posed: there are infinitely many continuous-time dynamics that can yield the same discrete-time transitions. Consequently, a misspecified continuous-time model may introduce significant errors, which can propagate to the learned optimal policy and lead to suboptimal decision-making.

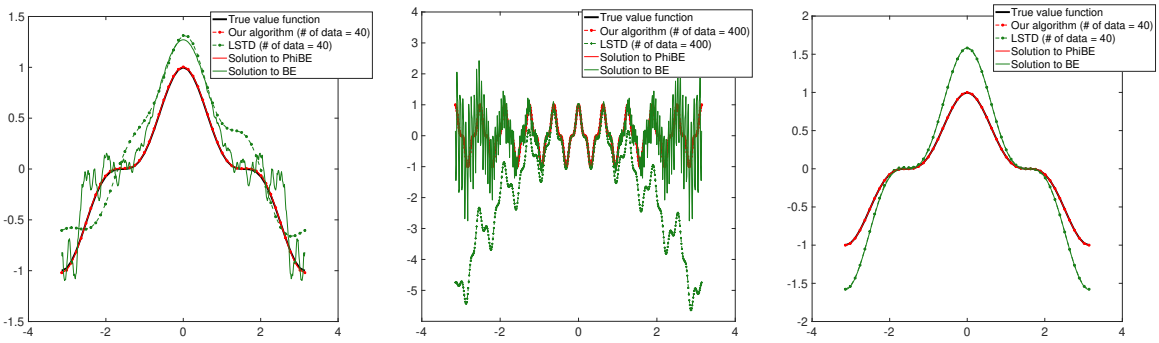
Another common approach to addressing the continuous-time PE problem involves discretizing time and treating it as a Markov reward process. This method yields an approximate value function satisfying a (discrete-time) Bellman equation (BE), thereby allowing RL algorithms such as temporal difference (TD) (Sutton and Barto, 2018), gradient TD (Precup et al., 2001), and least-squares TD (Bradtke and Barto, 1996) to be used to solve the BE. This approach is appealing for several reasons. First, it relies only on the discrete-time transition dynamics, thereby avoiding the non-identifiability issues inherent in the first approach. Second, many RL algorithms are model-free, eliminating the need to explicitly learn the transition dynamics, making these plug-and-play algorithms convenient to implement in practice. Despite these advantages, it has been observed empirically that the RL algorithms

can be sensitive to time discretization (Tallec et al., 2019; Munos, 2006; Park et al., 2021; Baird, 1994). As illustrated in Figure 1, such algorithms often yield poor approximations (see Section 5.1 for the details of Figure 1). Crucially, this failure is not due to stochasticity or insufficient data; rather, it stems from a more fundamental issue: the (discrete-time) Bellman equation is not a faithful approximation of the underlying continuous-time problem. We show in the paper that the solution to the Bellman equation is sensitive to time discretization, the rate of change of the rewards, and the discount coefficient, as shown in Figure 1. Fundamentally, the (discrete-time) Bellman equation is designed for discrete-time decision-making processes, where only the discrete-time transition dynamics are utilized. In contrast, the continuous-time dynamics considered in this paper are governed by an SDE, a structure that the (discrete-time) Bellman equation does not capture. To faithfully approximate the true continuous-time value function, it is therefore necessary to move beyond the standard MDP and (discrete-time) Bellman-equation frameworks.

The central question we aim to address in this paper is: if the dynamics are known to follow a standard SDE but only discrete-time data are available, can we design algorithms that *exploit this structure* to decisively outperform generic RL methods for policy evaluation?

First, we view the continuous-time PE problem in terms of the solution to a PDE. Then, we tackle the overall goal in two steps. The first step is to approximate the PDE operator with the discrete-time transition dynamics, and the second step is to solve the resulting equation in a model-free way. In the first step, we propose a PDE-based Bellman equation, called PhiBE, which integrates discrete-time information into a continuous PDE. PhiBE combines the advantages of both existing frameworks while mitigating their limitations. First, PhiBE is a PDE that preserves the continuous-time differential equation structure throughout the learning process. At the same time, PhiBE relies only on discrete-time transition distributions, which is similar to the BE, making it directly compatible with discrete-time data and enabling model-free algorithm design. Moreover, our framework overcomes key drawbacks of existing methods. Unlike the continuous-time PDE approach, which requires estimating the continuous-time dynamics and may suffer from identifiability issues, PhiBE depends solely on discrete-time transitions and does not require explicit dynamic modeling. Compared with the BE, PhiBE yields a more accurate approximation of the exact solution as it utilizes the underlying differential equation structure. In addition, PhiBE is more robust than BE to rapidly changing reward functions, which allows greater flexibility in designing reward functions to effectively achieve RL objectives. Moreover, higher-order PhiBE 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 the BE solution.

In the second step, we introduce a model-free algorithm for approximating the solution to PhiBE when only discrete-time data are accessible. This algorithm can be viewed as a continuous-time counterpart of least-squares temporal difference (LSTD) (Bradtke and Barto, 1996) by replacing the (discrete-time) Bellman equation with the PhiBE equation. Since PhiBE enjoys a smaller discretization error than the Bellman equation, the resulting algorithm achieves significantly improved approximation accuracy for free with a slight modification of the standard RL algorithm. As depicted in Figure 1, with exactly the



(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 units 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 (Bradtke and Barto, 1996) is a popular RL algorithm for linear function approximation. The PhiBE is proposed in Section 3 and the algorithm is proposed in Section 4.

same data and the same computational cost, the proposed algorithm outperforms the RL algorithms drastically. PhiBE can also be solved in a model-based way by first estimating the discrete-time transition dynamics. This differs from model-based continuous-time RL, as discussed in Section 1.1.2 and Remark 10.

The error analysis for continuous-time reinforcement learning differs fundamentally from that of classical discrete-time RL. For continuous-time PE under function approximation, the total error can be decomposed into three components, as illustrated in Figure 2. The first component is the *discretization error*, which arises from the mismatch between the underlying continuous-time dynamics and the discrete-time observations used for learning. The second component is the *algorithmic error*, which may be caused by model misspecification or by finite iterations when iterative solvers are employed. In this work, we use an offline solver, and therefore the algorithmic error refers to a pure model misspecification error. The third component is the *finite-sample error*, which results from having only a finite number of observations. We denote by $V_{\Delta t}^{n,\Phi}$ the value function approximation obtained from n samples collected at intervals of Δt using linear basis functions $\Phi(s)$. As $n \to \infty$, the estimator converges to $V_{\Delta t}^{\Phi}$, which still differs from the true value function V due to discretization error and the limited expressiveness of the basis functions. As the approximation space is enriched, i.e., as the number of basis functions $p \to \infty$, $V_{\Delta t}^{\Phi}$ further converges to $V_{\Delta t}$, the solution of the discretized Bellman equation in the MDP framework or of PhiBE, the new equation proposed in this paper. Finally, as the data collection frequency $\Delta t \to 0$, $V_{\Delta t}$ converges to the true continuous-time value function V. Most of the existing reinforcement learning literature focuses on analyzing the finite-sample and algorithmic errors, namely the gap between $V_{\Delta t}^{n}$ and $V_{\Delta t}$ (Munos and Szepesvári, 2008; Yang and Wang, 2019; Jin et al., 2020; Cai et al., 2020; Zhang et al., 2021). In contrast, comparatively little work has addressed the discretization error arising from the time discretization itself.

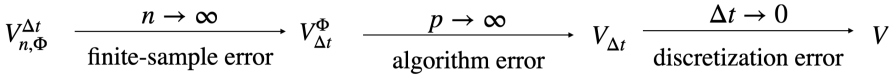


Figure 2: Error decomposition for continuous-time RL

In this paper, we analyze all three error components within the proposed PhiBE framework and its associated algorithm. In addition, we explicitly characterize the discretization error incurred by the MDP formulation, namely the gap between the (discrete-time) Bellman equation and the true continuous-time value function. In other words, the discretization error associated with the (discrete-time) Bellman equation represents the best approximation achievable by the standard policy-evaluation algorithms derived from the (discrete-time) Bellman equation, including temporal difference learning (Sutton, 1988), least-squares temporal difference methods (Bradtke and Barto, 1996), and gradient-based temporal difference methods (Sutton et al., 2009).

Our analysis leads to two notable observations. First, in terms of discretization error, PhiBE can provide a more accurate approximation than the MDP framework when the reward function exhibits high oscillation and the system dynamics evolve smoothly. This regime is common in practice because reward functions are often designed to vary sharply in order to distinguish rewards from penalties, while physical systems typically evolve smoothly over time. These properties suggest that the MDP framework may be suboptimal for certain continuous-time control problems, a limitation that has also been observed empirically (De Asis and Sutton, 2024). Second, for continuous-time PE, we identify an inherent trade-off with respect to the time discretization Δt. More frequent data collection reduces discretization bias but increases statistical variance, requiring more samples to achieve the same accuracy. To the best of our knowledge, the only prior work reporting a similar trade-off is (Zhang et al., 2023), which provides analytical results only for linear-quadratic policy evaluation without function approximation. In contrast, our results apply to general controlled diffusion processes and incorporate function approximation.

Contributions

- We show that the classical Bellman equation, which does not exploit the underlying continuous-time structure, is a first-order approximation to the continuous-time PE problem, whose error deteriorates when the reward oscillates. To address this limitation, we develop the PhiBE framework that explicitly utilizes the structure of the underlying dynamics, yielding provably smaller approximation error than the Bellman equation when rewards oscillate or dynamics evolve slowly. Furthermore, we derive higher-order extensions to further enhance the approximation accuracy.

- We propose a Galerkin method to approximate the solution to PhiBE under linear function approximation when the discrete-time transition dynamics are given, and establish its convergence under model misspecification. Directly applying the classical LSTD bound to the continuous-time setting yields a model-misspecification bound with an $O(\Delta t^{-1/2})$ stability factor. By exploiting the smooth structure of the underlying dynamics, our analysis instead yields an approximation bound with a stability factor independent of Δt.

- We propose a model-free algorithm to approximate the Galerkin solution to PhiBE when only data are available, and provide a sample-complexity bound guaranteeing an ϵ-accurate approximation. Under consistent continuous-time scaling, directly applying the classical LSTD bound gives sample complexity of order $O(\epsilon^{-2}\beta^{-4}\Delta t^{-2})$. By contrast, our bound is $O(\epsilon^{-2}\beta^{-4}\Delta t^{-1})$, improving the dependence on the sampling interval from Δt^{-2} to Δt^{-1}.
- Combining all the errors, we identify an inherent trade-off with respect to the time discretization Δt in continuous-time PE. More frequent data collection reduces discretization bias but increases statistical variance, requiring more samples to achieve the same accuracy. To the best of our knowledge, such an explicit trade-off, together with a nonasymptotic characterization of its dependence on Δt, n, and β, has not been systematically analyzed in prior work.
- The PhiBE formulation integrates discrete-time information with a continuous-time PDE. By leveraging the continuous-time structure, it achieves smaller discretization error than classical RL. Moreover, due to its structural similarity to the discrete-time Bellman equation, existing RL algorithms can be seamlessly adapted to the PhiBE framework, achieving improved approximation accuracy essentially for free.

1.1 Related Work

1.1.1 Continuous-Time RL

There are primarily two approaches to addressing continuous-time RL from the continuous-time 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 model-based continuous-time RL problem (Kamalapurkar et al., 2016; Della Santina et al., 2023; Lee and Sutton, 2021; Vamvoudakis and Lewis, 2010). However, the underlying continuous dynamics are often unidentifiable when only discrete-time data are available. Another approach involves developing an algorithm for the continuous-time optimal control problem and then adapting it to discrete-time data (Basei et al., 2022; Jia and Zhou, 2022a,b). Typically, numerical summation in discrete time is used to approximate continuous integrals (e.g., Algorithm 2 of Basei et al. (2022) or Equation (19) of Jia and Zhou (2022a)). While these methods offer valuable theoretical insights, two issues remain underexplored. First, most analyses guarantee convergence to the true value function when continuous-time data are available, but do not explicitly account for discretization error arising when only discrete samples are available. Second, although these works assume dynamics governed by a standard SDE, the resulting algorithms are not tailored to this structure and remain broadly applicable to other dynamics. By contrast, the central question of this paper is: when the dynamics follow a standard SDE but only discrete data are available, can we exploit this structure to build algorithms that outperform generic RL methods? To the best of our knowledge, this is the first work to address this question systematically and establish corresponding theoretical guarantees.

The proposed method fundamentally differs from the existing literature in three key aspects. First, unlike model-based approaches, which end up solving a PDE with continuous-time information, we integrate discrete-time information into the PDE formulation. Second,

unlike alternative methodologies that view the problem as a Markov reward process and neglect the smoothness of the underlying dynamics, 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. Third, in contrast to most continuous-time RL algorithms that require long trajectory data, our formulation depends only on the current and next states. Consequently, algorithms derived from it require access only to transition pairs rather than full trajectories, making the approach more flexible and broadly applicable to diverse data sources.

To sum up, our work complements this line of research by bridging the gap between continuous-time algorithms and discrete-time data. We treat the discrete-time data before the algorithm design. The PhiBE formulation offers a systematic way to approximate the continuous-time problem from discrete-time data. First, from a theoretical standpoint, because our learning algorithms retain the continuous-time PDE form, existing convergence analyses developed in the PDE literature (Guo et al., 2025; Kerimkulov et al., 2020; Tang et al., 2025, 2022) can be directly applied or adapted to analyze the proposed methods. Second, from a numerical perspective, PhiBE and the (discrete-time) Bellman equation are structurally similar and both depend only on the current and next states. Therefore, most existing RL algorithms developed for solving Bellman equations can be directly applied to PhiBE.

1.1.2 Statistical Estimation of Diffusion Dynamics from Discrete Observations

One model-based approach is to estimate the continuous-time drift and diffusion coefficients directly from discrete observations. Representative methods include kernel estimation (Florens-Zmirou, 1993), parametric contrast estimation (Kessler, 1997), penalized projection estimation (Comte et al., 2007; Schmisser, 2013), maximum likelihood (Pedersen, 1995; Aït-Sahalia, 2002), and operator- or neural-network-based methods (Hansen and Scheinkman, 1995; Gobet et al., 2004; Oga and Koike, 2024). Most studies either fix the sampling interval Δt when deriving rates in n or give asymptotic results as $\Delta t \to 0$. Nonasymptotic risk bounds depending explicitly on both n and Δt are available for joint drift and diffusion estimation in the one-dimensional stationary ergodic setting (Comte et al., 2007) and for drift estimation in multidimensional stationary or mixing settings (Schmisser, 2013; Oga and Koike, 2024).

A second model-based approach is to estimate the discrete-time transition dynamics directly. Representative nonparametric results include adaptive transition-density estimation for homogeneous Markov chains (Lacour, 2007; Sart, 2014), transition-operator estimation motivated by multidimensional diffusions observed at low frequency (Löffler and Picard, 2021), and diffusion transition-density estimation at a fixed lag (Comte and Marie, 2025). These results treat the Markov kernel or transition lag as fixed and therefore do not characterize the joint dependence of estimation error on n and a shrinking Δt. An estimated transition model can be inserted into either the Bellman equation or the conditional moments defining PhiBE, but propagating transition-estimation error to the final solution requires a separate analysis. Classical simulation-lemma arguments provide such propagation bounds for discrete-time Bellman equations (Kearns and Singh, 2002). Remark 10

explains the corresponding plug-in construction for PhiBE, whose end-to-end error analysis is left for future work.

The statistical analysis in this paper instead concerns the proposed model-free route. It estimates the value function from data without first fitting a model of either the continuous-time dynamics or the discrete-time transitions. The finite-sample guarantees established here apply only to this route. They explicitly characterize how the data-to-solution error depends jointly on n and Δt. Together with our discretization and function-approximation bounds, they control the error of the final policy-evaluation approximation.

1.1.3 Numerical Schemes for the HJB Equation

Although the formulation of the PhiBE closely resembles the PDE associated with the continuous-time PE problem, our method is fundamentally different from traditional numerical schemes for solving PDEs in three key aspects. First, classical schemes require full knowledge of the system dynamics, whereas our model-free algorithm does not. Second, numerical schemes approximate trajectories based on known dynamics, while our approach leverages exact trajectory data to approximate the underlying dynamics. Consequently, the corresponding error analysis framework also differs. Third, the objectives of error analysis are distinct. Numerical schemes typically focus on the convergence order of the discretization, whereas one of the aims in this paper is to compare the first-order PhiBE formulation with the discrete-time MDP formulation. Since both are first-order formulations, our analysis emphasizes the leading constant, which ultimately determines the practical performance of continuous-time PE. (See more discussion after Remark 6.)

1.1.4 Model-Free Algorithms for RL

Due to the structural similarity between PhiBE and the (discrete-time) Bellman equation, our model-free algorithm for solving PhiBE is closely related to model-free algorithms for the BE, particularly the LSTD method (Bradtke and Barto, 1996). Conceptually, the proposed solver serves as the continuous-time analogue of LSTD by replacing the (discrete-time) Bellman equation with the PhiBE equation. Both approaches are model-free, employ linear function approximation, and can be derived via a Galerkin projection followed by a stochastic approximation procedure. This connection highlights an important feature of the PhiBE framework: most existing RL algorithms developed for the BE can be readily adapted to the PhiBE setting with minimal modification. Moreover, because the PhiBE formulation incurs a smaller discretization error than the BE, it achieves higher approximation accuracy given sufficient data (see Remark 18 for detailed discussion). From a theoretical perspective, directly applying the classical LSTD bound of Lazaric et al. (2012) to the continuous-time setting yields a model-misspecification bound with an $O(\Delta t^{-1/2})$ stability factor. By exploiting the ellipticity of the infinitesimal generator, our analysis instead yields a stability factor independent of Δt. Under consistent continuous-time scaling, the classical bound gives sample complexity of order $O(\Delta t^{-2})$, whereas our bound is $O(\Delta t^{-1})$. These comparisons concern the dependence of the bounds on Δt and do not imply that the LSTD estimator itself diverges (see the discussions following Theorems 16 and 19).

Organization The setting of the problem is specified in Section 2. Section 3 introduces the PDE-based Bellman equation, PhiBE, and establishes the discretization error of PhiBE and the (discrete-time) BE. In Section 4, a model-free algorithm for solving PhiBE is proposed, and the algorithmic and finite-sample errors are derived. Numerical experiments are conducted in Section 5.

Notation For a function $f(s) \in \mathbb{R}$, $\|f\|_{L^\infty} = \sup_s f(s)$; $\|f\|_\rho = \sqrt{\int f(s)^2 \rho(s) ds}$. For a 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; $\|f\|_{L^\infty} = \sqrt{\sum_{i=1}^d \|f_i(s)\|_{L^\infty}^2}$, and similarly for $\|f\|_\rho$. For a 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 policy-evaluation (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 discount 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 $\mu(s)$ and $\sigma(s)$ are Lipschitz continuous and that the reward function $\|r\|_{L^\infty}$ is bounded. This ensures that (2) has a unique strong solution (Oksendal, 2013) and that the infinite-horizon integral is bounded.

Remark 1 *Continuous-time reinforcement learning (Wang et al., 2020) can be formulated as follows:*

$$\begin{aligned} \max_{a_t = \pi(s_t)} \quad & V^\pi(s) = \mathbb{E}\left[\int_0^\infty e^{-\beta t} r(s_t, a_t) dt | s_0 = s\right] \\ s.t. \quad & ds_t = \mu(s_t, a_t) + \sigma(s_t, a_t) dB_t, \quad a_t = \pi(s_t). \end{aligned}$$

where the goal is to find the optimal policy $\pi(s)$ that maximizes the value function under unknown drift and diffusion. In this paper, we focus on the unregularized optimal control problem and therefore restrict attention to deterministic feedback policies $a_t = \pi(s_t)$. For stochastic policies (see, e.g., (Jia and Zhou, 2022a; Wang et al., 2020)), additional challenges may arise (see, e.g., (Jia et al., 2025; Szpruch et al., 2024)), which we leave for future study.

The policy evaluation problem is a fundamental step in reinforcement learning algorithms, such as actor-critic and PPO (Konda and Tsitsiklis, 1999; Schulman et al., 2017). Its goal is to assess the quality of a fixed policy π by evaluating the value function under that policy. The continuous-time PE problem (1)–(2) considered in this paper arises by denoting $r^\pi(s) = r(s, \pi(s))$, $\mu^\pi(s) = \mu(s, \pi(s))$, $\sigma^\pi(s) = \sigma(s, \pi(s))$. For simplicity, we omit the subscript π throughout, as the policy is fixed in the entire paper.

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 $\rho(s', \Delta t|s)$ over the discrete time interval 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 times $j\Delta t$ are given. Here the trajectory data $s = \{s_0^l, s_{\Delta t}^l, \cdots, s_{m\Delta t}^l\}_{l=1}^I$ contain I independent trajectories, and the initial state s_0^l of each trajectory is sampled from a distribution $\rho_0(s)$.

The error analysis in Section 3 is conducted under Case 1. We demonstrate that the (discrete-time) Bellman equation is not always the optimal equation for solving continuous-time PE problems, even when the discrete-time transition dynamics are known, and consequently, that 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 provides a superior approximation to the true value function compared with the classical Bellman equation. When only trajectory data are available (Case 2), one can also use the data to solve PhiBE through model-free RL, as discussed in Section 4.

3. A PDE-Based Bellman Equation (PhiBE)

In Section 3.1, we first introduce the (discrete-time) 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 the deterministic case (Section 3.2.1) and the stochastic case (Section 3.2.2). The error analysis provides guidance on when PhiBE is a better approximation than the BE.

3.1 (Discrete-Time) Bellman Equation

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

$$\tilde{V}(s) = \mathbb{E}\left[\sum_{j=0}^{\infty} e^{-\beta\Delta t j} r(s_{j\Delta t})\Delta t | s_0 = s\right]. \tag{3}$$

In this way, it can be viewed as a policy-evaluation problem in a Markov decision process, where the state is $s \in \mathbb{S}$, the reward is $r(s)\Delta t$, the discount factor is $e^{-\beta\Delta t}$, and the transition kernel is $\rho(s', \Delta t|s)$. Therefore, the approximate value function $\tilde{V}(s)$ satisfies the following (discrete-time) Bellman equation (Sutton and Barto, 2018). In this paper, we will simply refer to it as the Bellman equation (abbreviated as BE).

Definition 2 (Definition of BE)

$$\tilde{V}(s) = r(s)\Delta t + e^{-\beta\Delta t}\mathbb{E}_{s_{\Delta t}\sim\rho(s',\Delta t|s)}[\tilde{V}(s_{\Delta t})|s_0 = s]. \tag{4}$$

When the discrete-time transition distribution is not given, one can utilize various RL algorithms to solve the Bellman equation (4) using the trajectory data. Note that some continuous-time RL papers approximate the value function by $V(s) \approx \sum_{i=0}^{T} e^{-\beta i \Delta t} r(s_{i\Delta t})$, which in fact is also approximating the solution of the Bellman equation, owing to the equivalence between (4) and (3). 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 (Discretization error for BE) *Assume that $\|r\|_{L^\infty}$ and $\|\mathcal{L}_{\mu,\Sigma} r\|_{L^\infty}$ are bounded. Then the solution $\tilde{V}(s)$ to the BE (4) 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

$$\mathcal{L}_{\mu,\Sigma} = \mu(s) \cdot \nabla + \frac{1}{2} \Sigma : \nabla^2, \tag{5}$$

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

Remark 4 (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, $\|s \cdot \nabla r(s)\|_{L^\infty} \leq C$.*

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 see more clearly where the error in the BE comes from. Note that $V(s)$, as defined in (1), can be equivalently written as

$$\begin{aligned} V(s) =& \mathbb{E}\left[\int_0^{\Delta t} e^{-\beta t} r(s) dt + \int_{\Delta t}^{\infty} e^{-\beta t} r(s_t) dt | s_0 = s \right] \\ =& \mathbb{E}\left[\int_0^{\Delta t} e^{-\beta t} r(s) dt | s_0 = s \right] + e^{-\beta \Delta t} \mathbb{E}\left[V(s_{t+\Delta t}) | s_0 = s \right] \end{aligned} \tag{6}$$

One can interpret the Bellman equation defined in (4) 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 (6) by $r(s)$. The error between these two terms can be bounded by:

$$\left| \mathbb{E}\left[\left(\frac{1}{\Delta t} \int_0^{\Delta t} e^{-\beta t} r(s_t) dt \right) - r(s_0) \middle| s_0 = s \right] \right| \leq \frac{1}{2} \left(\beta \|r\|_{L^\infty} + \|\mathcal{L}_{\mu,\Sigma} r\|_{L^\infty} \right) \Delta t + o(\Delta t), \tag{7}$$

characterizes the error of $\left\| V - \tilde{V} \right\|_{L^\infty}$ in Theorem 3.

Theorem 3 indicates that the solution $\tilde{V}$ to the Bellman equation (4) 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.

Fundamentally, the Bellman equation is designed for discrete-time decision-making processes and depends only on the discrete-time transition kernel $\rho(s', \Delta t|s)$. This kernel may arise from any underlying mechanism. As a result, the Bellman equation does not utilize or encode any assumption about how $\rho(s', \Delta t \mid s)$ is generated. In particular, it does not exploit the assumption made in this paper that the discrete-time dynamics are induced by a continuous-time SDE. To incorporate this continuous-time structure into the approximation, one must go beyond the standard MDP framework. 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 combine the differential equation structure with the discrete-time transition dynamics and estimate the value function more accurately.

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 and can be described by the following ODE:

$$\frac{ds_t}{dt} = \mu(s_t). \tag{8}$$

If the discrete-time transition dynamics $p(s', \Delta t|s) = p_{\Delta t}(s)$ are 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 for 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 the solution to the following PDE,

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

The above equation is known as the Hamilton–Jacobi–Bellman (HJB) equation, often referred to as the continuous-time Bellman equation. In this paper, we will consistently refer to it as the HJB equation. A detailed derivation can be found in Appendix A or in classical control textbooks such as Fleming and Rishel (1975). Applying a finite-difference scheme, one can approximate $\mu(s_t)$ by

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

and substituting it back into (9) 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 Bellman equation, which is blind to the underlying SDE structure, PhiBE explicitly exploits this structure by involving the gradient of the value function, thereby embedding the smoothness induced by the continuous-time dynamics. At the same time, unlike the true PDE (9), which requires direct access to the continuous-time dynamics and is therefore non-identifiable from discrete-time information, PhiBE depends only on the discrete-time transition kernel. This makes PhiBE amenable to model-free algorithm design while still leveraging continuous-time structure.

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

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

results 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 the ith-order PhiBE for deterministic dynamics in the following definition.

Definition 5 (i-th order PhiBE in deterministic dynamics) *When the underlying dynamics are deterministic, the ith-order PhiBE is defined as follows:*

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

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

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

Remark 6 *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

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

Discussion on the relation to numerical schemes for the HJB equation. The PhiBE formulation is fundamentally different from classical numerical schemes for the HJB equation in three key aspects.

First of all, the goals of the two approaches are different. Classical numerical schemes aim to approximate the solution of the PDE $\mathcal{L}_{\mu,\sigma} V(s) = 0$, where the operator $\mathcal{L}_{\mu,\sigma}$ is explicitly given. In contrast, our goal is not only to approximate the solution to the PDE but also to do so without full knowledge of the PDE operator. We tackle the overall goal in two steps. The first step is to approximate the PDE operator $\mathcal{L}_{\mu,\sigma}$, and the second step is to solve the resulting equation. The PhiBE formulation (11) is constructed to address the first task, namely approximating the unknown operator $\mathcal{L}_{\mu,\sigma} V(s) = 0$ by $\mathcal{L}_{\hat{\mu},\hat{\sigma}} V(s) = 0$ using only partial information in the form of the discrete-time transition kernel $\rho(s', \Delta t \mid s)$, with no direct access to $\mu(s)$ or $\sigma(s)$. Consequently, the "discretization error" analyzed in this section refers to the error incurred from observing $\rho(s', \Delta t \mid s)$ at $\Delta t > 0$ rather than the infinitesimal kernel $\lim_{\Delta t \to 0} \rho(s', \Delta t \mid s)$. It does not refer to the numerical error of solving the PDE.

Second, the discretization formulations differ. Classical schemes discretize the state space over a uniform mesh h, e.g., $V'(s) \approx \frac{1}{h}(V(s+h) - V(s))$, while such uniform meshing cannot be directly applied in RL because trajectory data are irregularly distributed. Consequently, numerical error in classical schemes arises from spatial discretization, whereas in PhiBE it originates from time discretization through the approximation of $\hat{\mu}$ and $\hat{\Sigma}$. Moreover, although the definition of $\hat{\mu}(s)$ superficially resembles a finite-difference expression, its meaning is fundamentally different. In classical methods, one starts from known dynamics $\mu(s)$ and uses numerical schemes to approximate the trajectory $s_{j\Delta t}$. As a result, numerical error propagates and accumulates over time. In contrast, PhiBE takes the opposite approach: the trajectories $s_{j\Delta t}$ are observed from data and used to approximate the underlying dynamics $\mu(s)$. This inversion of roles leads to a distinct analytical treatment of convergence and convergence rates for $\hat{V}_i(s)$ that does not accumulate over time, in sharp contrast to classical numerical discretization errors.

Finally, the error analysis frameworks and emphasis are distinct. Classical analyses establish convergence by verifying that the discrete operator is monotone, stable, and consistent, then applying comparison principles and perturbation arguments to derive convergence

rates. In contrast, our analysis first quantifies the discrepancy between the data-driven dynamics $(\hat{\mu}, \hat{\Sigma})$ and the true dynamics (μ, Σ), a component never addressed in traditional numerical analyses, and then propagates this discrepancy to the value function via PDE stability. Furthermore, while classical numerical analyses focus on the order of accuracy in the mesh size h and typically omit constants, our comparison with discrete-time RL requires explicit characterization of these constants, as both standard RL and PhiBE are first-order approximations. The leading constant, which depends on the reward and dynamics, ultimately determines the practical performance of continuous-time RL algorithms.

In summary, PhiBE differs fundamentally from classical numerical schemes for HJB equations. Nevertheless, because the continuous-time RL problem is connected to a PDE through the PhiBE formulation, one could, in principle, estimate ρ_Δ from data, compute $(\hat{\mu}, \hat{\Sigma})$ explicitly, and then apply standard numerical schemes to solve the resulting PhiBE equation. The classical convergence results would then apply to this *model-based* approach. However, this paper will focus on developing a *model-free* algorithm for solving the PhiBE equation directly from data without explicitly estimating $\hat{\mu}$ or $\hat{\Sigma}$.

The error analysis of PhiBE for deterministic dynamics is presented in the following theorem.

Theorem 7 (Discretization error for deterministic PhiBE) *Assume that* $\|\nabla r(s)\|_{L^\infty}$ *and* $\left\|\mathcal{L}_\mu^i \mu(s)\right\|_{L^\infty}$ *are bounded. Additionally, assume that* $\|\nabla \mu(s)\|_{L^\infty} < \beta$. *Then the solution* $\hat{V}_i(s)$ *to the PhiBE* (11) *is an* i*th-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* (50) *that only depends on the order* i.

See Section 6.2 for the proof of Theorem 7. Several remarks regarding this theorem are in order.

1st-order PhiBE vs. BE By Theorem 7, the distance between the first-order PhiBE solution and the true value function can be bounded by

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

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

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

one observes that when the change of the reward is rapid, i.e., $\|\nabla r\|_{L^\infty}$ is large, 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.

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 7, 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.

Dependence on β When β is large, PhiBE always has a smaller discretization error than BE. However, when β is small, the BE error scales as $O(\beta^{-1})$, whereas the PhiBE error scales as $O(\beta^{-2})$. Thus, when β is small, the gap in discretization error between the two methods naturally narrows. However, as shown later in Section 4, the finite-sample error of the model-free algorithms for both BE and PhiBE scales as $O(\beta^{-2})$. As a result, in data-driven settings, the potential advantage of BE for small β is dominated by the finite-sample error, which scales as $O(\beta^{-2})$. In contrast, PhiBE retains an advantage in regimes with slowly varying dynamics and rapidly changing reward functions.

In addition, in standard RL settings, the discount factor is $\gamma = e^{-\beta \Delta t} \in [0.9, 0.99]$, so the resulting range of β depends on the sampling frequency. When the data are collected at $\Delta t = 0.01$, we expect $\beta = O(1)$, i.e., $\beta \in [1, 10]$. For data collected at $\Delta t = 0.1$, one obtains a smaller $\beta \in [0.1, 1]$. In the experiments in Section 5, we therefore evaluate both methods across a broad range $\beta \in [0.1, 10]$.

When does PhiBE outperform BE? In summary, PhiBE outperforms BE in terms of discretization error whenever one of the following conditions is present: (i) the dynamics evolve slowly; (ii) the reward function varies sharply; or (iii) the discount rate is not extremely small. In summary, these features arise frequently in continuous-time RL applications. First, slow dynamics are common in physical and biomedical systems, where state variables evolve gradually and smoothly. For instance, in glucose–insulin regulation, blood glucose levels follow physiological ODE models with inherently slow drift dynamics. Second, the reward function often needs to sharply penalize excursions outside a healthy range, so that the optimal policy strongly encourages keeping the patient within a safe region. This leads to a highly varying reward landscape, i.e., a large $\|\nabla_s r(s)\|$. Third, discount coefficients used in practice are rarely associated with extremely small continuous-time rates. In OpenAI Gym environments with $\Delta t \sim 10^{-2}$, the typical choice $\gamma \in [0.9, 0.99]$ corresponds to $\beta \in [1, 10]$. In clinical settings with larger sampling intervals, objectives often focus on maintaining stability over short or moderate horizons, so the effective discount rate is similarly not small. Taken together, the three features are common in real-world applications, making the settings where PhiBE outperforms BE not exceptional but the typical regime for continuous-time reinforcement learning.

Assumptions on $\left\|\mathcal{L}_\mu^i \mu(s)\right\|_{L^\infty}$ 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 $\mu(s)$. A sufficient condition ensuring that $\left\|\mathcal{L}_\mu^i \mu(s)\right\|_{L^\infty}$ is bounded is that $\left\|\nabla^k \mu(s)\right\|_{L^\infty}$ is bounded for all $0 \le k \le i$. Note that the linear dynamics $\mu(s) = As$ do not satisfy this condition. We lose some sharpness in the upper bound to make the theorem applicable to general

dynamics. However, we prove in Theorem 8 that PhiBE works when $\mu(s) = As$, and one can derive a sharper error estimate for this case.

Theorem 8 *When the underlying dynamics follow $\frac{d}{dt}s_t = As_t$, where $s_t \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$, the solution to the i-th order PhiBE in deterministic dynamics approximates the true value function with an error*

$$\left\|\hat{V}_i - V\right\|_{L^\infty} \le \frac{C_i}{\beta^2} \left\|s \cdot \nabla r(s)\right\|_{L^\infty} \|A\|_2 D_A^i \Delta t^i, \quad D_A = e^{\|A\|_2 \Delta t} \|A\|_2$$

where C_i is a constant defined in (50) *that only depends on the order i.*

The proof of the above theorem is provided in Section 6.3. According to Theorem 8, even when the dynamics $s_t = e^{At}s_0$ evolve exponentially fast due to positive real parts in the eigenvalues of A, the PhiBE solution remains a good approximation to the true value function provided that $D_A\Delta t$ is small. In particular, when $D_A\Delta t < 1$, higher-order PhiBE achieves a smaller approximation error than the first-order PhiBE.

3.2.2 Stochastic Dynamics

When $\sigma(s) \not\equiv 0$ is a non-degenerate matrix, the dynamics are stochastic and driven by the SDE in (2). By the Feynman–Kac theorem (Stroock and Varadhan, 1997), the value function $V(s)$ satisfies the following HJB equation:

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

where $\mathcal{L}_{\mu,\Sigma}$ is an operator defined in (5). 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 for stochastic dynamics to approximate the true value function $V(s)$.

Definition 9 (i-th order PhiBE in stochastic dynamics) *When the underlying dynamics are stochastic, the ith-order PhiBE is defined as follows:*

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

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 (5)*, and $a^{(i)} = (a_0^{(i)}, \cdots, a_i^{(i)})^\top$ is defined in* (13)*.*

Remark 10 (Model-based and model-free implementations of PhiBE) *Suppose that trajectory data are used to estimate the transition densities $\widehat{\rho}_{j\Delta t}(s' \mid s)$, $j = 1, \ldots, i$. Replacing $\rho(\cdot, j\Delta t \mid s)$ in* (18) *with $\widehat{\rho}_{j\Delta t}(\cdot \mid s)$ yields estimates $(\widehat{\mu}_i^\rho, \widehat{\Sigma}_i^\rho)$. Substituting these*

estimated coefficients into (17) *gives a model-based PhiBE that can be solved by a PDE or Galerkin method. The estimated one-step transition operator can likewise be inserted directly into the Bellman equation.*

These plug-in operations introduce a second error stage after transition estimation. For example, writing $\delta\rho_j(s' \mid s) = \widehat{\rho}_{j\Delta t}(s' \mid s) - \rho(s', j\Delta t \mid s)$*, the induced drift error satisfies*

$$\widehat{\mu}_i^{\rho}(s) - \hat{\mu}_i(s) = \frac{1}{\Delta t}\sum_{j=1}^{i} a_j^{(i)} \int_{\mathbb{S}} (s'-s)\delta\rho_j(s' \mid s)\, ds'.$$

The corresponding expression for the error in $\widehat{\Sigma}_i^{\rho}$ *is analogous. Thus, transition-density error must first be converted into weighted moment error, with the factor* $1/\Delta t$*, and the resulting coefficient perturbation must then be propagated through the PhiBE. The transition-estimation results cited in Section 1.1.2 do not provide the corresponding transition-to-final-solution bound for PhiBE. Such a bound requires norms strong enough to control the weighted moments above, their spatial regularity, and the stability of the perturbed PDE. We therefore leave a complete end-to-end analysis of this model-based PhiBE procedure for future work. Section 4.2 instead studies our model-free implementation, which bypasses a global estimate of* $\rho_{\Delta t}$ *and directly estimates the Galerkin system; the finite-sample theory in Section 4.3 applies only to that proposed estimator.*

Remark 11 *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 $\tilde{\Sigma}(s)$

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

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 (Zhu and Ying, 2020; Zhu et al., 2022) to alleviate the double-sampling problem when the underlying dynamics are smooth, that is, by approximating* $s'_{j\Delta t} \approx s_{(j-1)\Delta t} + (s_{(j+1)\Delta t} - s_{j\Delta t})$*. However, this would introduce additional bias into the approximation. We leave the study of this approximation* $\tilde{\Sigma}_i(s)$ *or the application of BFF to* $\tilde{\Sigma}_i(s)$ *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 that the solution $\hat{V}_i(s)$ to the ith-order PhiBE provides an ith-order approximation to the true value function $V(s)$ under the following assumptions.

Assumption 1 *Assumptions on the dynamics:*

(a) $\lambda_{min}(\Sigma(s)) > \lambda_{min} > 0$ *for all* $s \in \mathbb{S}$, *where* $\lambda_{min}(A)$ *is the smallest eigenvalue of* A.

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

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 ith-order approximations to μ and Σ, respectively.

To establish the error analysis, we define a weighted L^2 inner product and norm under a probability density $\rho(s)$ by

$$\langle f, g \rangle_\rho = \int f(s)g(s)\rho(s)ds, \quad \|f\|_\rho^2 = \int f^2(s)\rho(s)ds. \tag{19}$$

In addition, the density ρ satisfies

$$L_\rho = \|\nabla \log \rho\|_\rho \text{ is bounded.} \tag{20}$$

One may take $\rho(s)$ to be the stationary distribution (when it exists), namely a density satisfying

$$\int \mathcal{L}_{\mu,\Sigma}\phi(s)\rho(s)ds = 0. \quad \text{for all } \phi(s) \in C_c^\infty. \tag{21}$$

If a stationary distribution does not exist, one may select a density ρ such that

$$\left\| -\mu + \frac{1}{2}\nabla \cdot \Sigma + \frac{1}{2}\Sigma \nabla \log \rho \right\|_{L^\infty(\Omega)}^2 \le \frac{1}{2}\beta\lambda_{\min}, \tag{22}$$

where $\Omega = \{s : \rho(s) \neq 0\}$. If $\Omega \neq \mathbb{R}^d$, then all assumptions and results hold on Ω, rather than on the entire space $\mathbb{R}^d$.

Remark 12 *Several remarks regarding the weight* $\rho(s)$ *are in order. First, classical results (e.g., (Huang et al., 2015)) imply that when the drift and diffusion satisfy a suitable dissipation condition, a stationary distribution exists. Furthermore, once a stationary distribution exists, Theorem 1.1 of (Bogachev et al., 1996) yields the bound* $L_\rho \le \frac{1}{\lambda_{\min}} \|\mu + \nabla \cdot \Sigma\|_{L^\infty}$.

Second, if no stationary distribution exists, one can still choose a density ρ *such that* (22) *holds. For example, in one of our numerical experiments* (36)*, the system admits no stationary solution when* $\lambda > 0$. *In this case, one may take*

$$\rho(s) = \begin{cases} \dfrac{1}{2c}, & s \in [-c, c], \\ 0, & s \notin [-c, c], \end{cases} \qquad |c| \le \frac{\beta\sigma^2}{2\lambda},$$

which ensures (22). *The trade-off is that under this choice of weighted norm, one can guarantee convergence of the approximation error only on the domain* $\Omega = \{s : \rho(s) \neq 0\}$, *rather than on the entire space* $\mathbb{R}^d$. *At the same time, Assumption 1 on the drift and*

diffusion is also relaxed: instead of requiring it to hold globally, it only needs to be satisfied on the domain Ω.

Third, if one chooses to work with the Lebesgue measure on $\Omega \subset \mathbb{R}^d$*, then, in addition to Assumption 1 on the drift and diffusion terms, we require*

$$\left\| -\mu + \frac{1}{2}\nabla \cdot \Sigma \right\|^2_{L^\infty(\Omega)} \leq \frac{1}{2}\beta\lambda_{min}.$$

This ensures that the requirement (22) *remains valid when the density* ρ *is taken to be the (unnormalized) Lebesgue measure.*

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

Theorem 13 (Discretization error for BE in weighted L^2 norm) *Assume that* $\|r\|_\rho$ *and* $\|\mathcal{L}_{\mu,\Sigma} r\|_\rho$ *are bounded. Then the solution* $\tilde{V}(s)$ *to the BE* (4) *approximates the true value function* $V(s)$ *defined in* (1) *with an error*

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

The proof of Theorem 13 is given in Section 6.4. Next, the error analysis for PhiBE in stochastic dynamics is presented in the following theorem.

Theorem 14 (Discretization error for stochastic PhiBE) *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 $D_{\mu,\Sigma,\beta}$*,* $C_{r,\mu,\Sigma}$*, and* $\hat{C}_{\mu,\Sigma,\beta}$ *are constants defined in* (56) *and* (57) *that depend on* μ*,* Σ*, and* r*.*

The proof of Theorem 14 is given in Section 6.5.

Remark 15 (PhiBE vs. BE) *Here we discuss two cases. The first case is when the diffusion is known, that is,* $\hat{\Sigma}_i = \Sigma$*. In this case, 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 is proportional to the rate of change of the dynamics* $\mathbb{E}[\frac{d^{i+1}}{dt^{i+1}} s_t] = \mathcal{L}^i_{\mu,\Sigma}\mu$*. See the discussion under Theorem 7 for the benefits of the first- and higher-order PhiBE under different dynamics.*

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

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

where

$$\begin{aligned}
&L_\mu \lesssim \left\|\mathcal{L}\mu\right\|, \quad L_\Sigma \lesssim \left\|\mu\mu^\top + \Sigma\nabla\mu + \mathcal{L}\Sigma\right\|_{L^\infty} + \left\|\mu\right\|_{L^\infty}, \\
&L_{\nabla\cdot\Sigma} \lesssim \sqrt{\frac{C_\nabla}{\lambda_{min}}}\left\|\mu\mu^\top + \Sigma\nabla\mu + \mathcal{L}\Sigma\right\|_{L^\infty} + \left\|\nabla\cdot\left(\mu\mu^\top + \Sigma\nabla\mu + \mathcal{L}\Sigma\right)\right\|_{L^\infty} + \sqrt{\frac{C_\nabla}{\lambda_{min}}}\left\|\mu\right\|_{L^\infty} \\
&+ \left\|\nabla\mu\right\|_{L^\infty}, \\
&\sqrt{\frac{C_\nabla}{\lambda_{min}}} \lesssim \frac{\left\|\nabla\mu\right\|_{L^\infty}}{2\lambda} + \sqrt{\frac{\left\|\nabla\Sigma\right\|_{L^\infty}}{\lambda_{min}}}, \quad \left\|r\right\|_{\rho,L^\infty} = \left\|r\right\|_\rho + \left\|r\right\|_{L^\infty}.
\end{aligned}$$

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. Compared with the BE upper bound $\left\|\mathcal{L}r\right\|_\rho + \beta\left\|r\right\|_\rho$, which is more sensitive to large β and variations in the reward function, the PhiBE upper bound 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 than the BE.

4. Model-Free Algorithm for PhiBE

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. Finally, we conclude this section with an end-to-end error analysis for the model-free algorithm in Section 4.3.

4.1 Galerkin Method

Given p basis functions $\{\phi_i(s)\}_{i=1}^p$, 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 $\theta \in \mathbb{R}^p, \Phi(s) = (\phi_1(s), \cdots, \phi_p(s))^\top$, and $\mathcal{L}_{\mu,\Sigma}$ is defined in (5). The Galerkin method involves inserting the ansatz $\bar{V}$ into the PDE and then projecting it onto the finite basis,

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

where the same measure ρ defined in (19) is used here. This results in a linear system for 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 PE problems, one does not have access to the underlying dynamics μ, Σ. However, the approximate dynamics $\hat{\mu}_i, \hat{\Sigma}_i$ are 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 obtained by approximating A with $\hat{A}_i$:

$$\hat{A}_i = \left\langle \beta\Phi(s) - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i}\Phi(s), \Phi(s)\right\rangle_\rho, \tag{23}$$

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

$$\left\langle (\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i})\hat{V}_i^G(s), \Phi\right\rangle_\rho = \langle r(s), \Phi(s)\rangle_\rho. \tag{24}$$

For sufficiently small Δt satisfying $\Delta t^i \le \min\{\eta_{\mu,\Sigma,\beta}, D_{\mu,\Sigma,\beta}\}$, the Galerkin solution $\hat{V}_i^G(s)$ approximates the solution to the i-th order PhiBE defined in (17) *with an error*

$$\left\|\hat{V}_i^G(s) - \hat{V}_i(s)\right\|_\rho \le \frac{C_G}{\beta}\inf_{V^P=\theta^\top\Phi}\left\|\hat{V}_i - V^P\right\|_{H^1_\rho}.$$

Here, $\eta_{\mu,\Sigma,\beta}$, C_G, and $D_{\mu,\Sigma,\beta}$ are constants depending on μ, Σ, β, L^∞_ρ, and L^∞_Φ, defined in (60), (61), *and* (56), *respectively, and $\|f\|_{H^1_\rho} = \|f\|_\rho + \|\nabla f\|_\rho$.*

The proof of the theorem is given in Section 6.6.

Several remarks are in order. First, the above theorem establishes an approximation-error bound under model misspecification, that is, the error arising when the true value function cannot be exactly represented within the chosen linear bases $\Phi(s)$. More specifically, the bound in Theorem 16 can be interpreted as the limiting error of the data-driven algorithms (Algorithms 1 and 2) when infinite data are available.

Second, the Galerkin solver (24) is a continuous-time counterpart of LSTD (Bradtke and Barto, 1996): LSTD projects the discrete-time Bellman equation onto a linear function space, whereas our method projects the PhiBE equation. This connection motivates comparing the Galerkin error of Theorem 16 with the approximation error obtained by applying an LSTD bound to the same setting. The discrete-time value function $\tilde{V}$ is defined in (3), and its Galerkin approximation $\tilde{V}^G(s) = \theta_{\Delta t}^\top\Phi(s)$ satisfies the BE (4) in the following weak form:

$$\left\langle \tilde{V}^G(s) - e^{-\beta\Delta t}\mathbb{E}\left[\tilde{V}^G(s_{\Delta t}) \mid s_0 = s\right], \Phi(s)\right\rangle_\rho = \langle r(s)\Delta t, \Phi(s)\rangle_\rho.$$

The population LSTD approximation term in Theorem 1 of Lazaric et al. (2012) gives

$$\|\tilde{V} - \tilde{V}^G\|_\rho \le \frac{1}{\sqrt{1-e^{-2\beta\Delta t}}}\inf_{V^P=\theta^\top\Phi}\|\tilde{V} - V^P\|_\rho \lesssim \frac{1}{\sqrt{\beta\Delta t}}\inf_{V^P=\theta^\top\Phi}\|\tilde{V} - V^P\|_\rho. \tag{25}$$

For a fixed feature space with a nonvanishing best approximation error, the right-hand side of (25) diverges as $\Delta t \to 0$. This does not imply that the population Galerkin error $\|\tilde{V} - \tilde{V}^G\|_\rho$ itself diverges. Rather, it shows that directly applying the classical discrete-time RL bound to the continuous-time setting yields no uniform error guarantee as the data interval Δt vanishes.

In contrast, the PhiBE Galerkin bound in Theorem 16 is independent of Δt. Although this comparison does not establish that the PhiBE Galerkin error is always smaller than the population Galerkin error of LSTD, it demonstrates that our error analysis is sharper in its dependence on Δt when the transition kernel is induced by an SDE. This improvement arises because our analysis exploits the coercivity and elliptic structure of the infinitesimal generator, whereas the general contraction-based LSTD bound does not use this structure. The resulting Δt-uniform stability guarantee is the key novelty of Theorem 16.

The role of regularity in sharpening RL error bounds also appears in other settings. For example, Maran et al. (2023) use Lipschitz and Hölder regularity to analyze policy-performance mismatch in continuous-action behavioral cloning. Their result, however, does not address LSTD or policy-evaluation Galerkin error. Whether the regularity of SDE-induced transitions can also yield a Δt-uniform Galerkin bound for the discrete-time BE remains an interesting direction for future work.

Remark 17 *The Galerkin error remains bounded even when $\|\nabla \log \rho\|_{L^\infty}$ is unbounded. However, since the sample-complexity analysis in this paper assumes boundedness of $\|\nabla \log \rho\|_{L^\infty}$, we move the Galerkin error theorem under unbounded $\|\nabla \log \rho\|_{L^\infty}$ assumption to Section 6.6 to maintain internal consistency. The corresponding result is stated in Theorem 27.*

4.2 Model-Free Galerkin Method for PhiBE

When only discrete-time trajectory data are available, we first develop unbiased estimates $\bar{\mu}_i, \bar{\Sigma}_i$ of $\hat{\mu}_i, \hat{\Sigma}_i$ from the trajectory data:

$$\begin{aligned} \bar{\mu}_i(s^l_{j\Delta t}) &= \frac{1}{\Delta t}\sum_{k=1}^{i} a_k^{(i)}(s^l_{(j+k)\Delta t} - s^l_{j\Delta t}), \\ \bar{\Sigma}_i(s^l_{j\Delta t}) &= \frac{1}{\Delta t}\sum_{k=1}^{i} a_k^{(i)}(s^l_{(j+k)\Delta t} - s^l_{j\Delta t})(s^l_{(j+k)\Delta t} - s^l_{j\Delta t})^\top, \end{aligned} \tag{26}$$

with $a^{(i)}$ defined in (13). Then, using the above unbiased estimates, 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 approximate value function $\bar{V}(s) = \Phi(s)^\top\theta$ in terms of the finite basis. Note that our algorithm can also be applied to stochastic

or even unknown rewards, as only reward observations at discrete times are required. 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 basis set $\Phi(s) = (\phi_1(s), \cdots, \phi_p(s))^\top$.

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 (26).

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

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 basis set $\Phi(s) = (\phi_1(s), \cdots, \phi_p(s))^\top$.

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 (26).

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

Remark 18 (Comparison to the LSTD Algorithm) *Algorithms 1 and 2 serve as continuous-time counterparts of LSTD (Bradtke and Barto, 1996) by replacing the discrete-time Bell-*

man equation with the PhiBE equation. Both our algorithms and LSTD are model-free and employ linear basis functions. In addition, both methods can be derived by applying the Galerkin method, followed by a stochastic approximation of the resulting system (Szepesvári, 2011).

This connection, however, highlights a key advantage of our formulation. Owing to the structural similarity between PhiBE (11), (17) *and the discrete-time Bellman equation, both depend only on the current and next states. Thus, most existing RL algorithms developed for solving Bellman equations can be directly applied to PhiBE. Moreover, as shown in Section 3, our method enjoys a smaller discretization error, allowing one to achieve significantly improved approximation accuracy* for free *with a slight modification of the RL algorithm.*

4.3 Sample Complexity of the Model-Free Algorithm for First-Order PhiBE

We now establish the sample complexity of the model-free algorithm for solving the first-order PhiBE. A sample complexity analysis for higher-order PhiBE is deferred to future work. We assume that the data consist of n independent short trajectories $\{s_i, s_i'\}_{i=1}^n$. Each trajectory provides a single transition pair: an initial state $s_i \sim \rho(s)$ and its one-step successor $s_i' \sim \rho(s', \Delta t \mid s_i)$. Here the initial density $\rho(s)$ is the weight we defined in (19), and the discrete-time transition $\rho(s', \Delta t \mid s)$ is driven by the underlying SDE (2). The model-free algorithm approximates the value function by $\hat{V}_n(s) = \Phi(s)^\top \theta_n$, where $\theta_n \in \mathbb{R}^p$ is obtained by solving the empirical linear system

$$A_n \theta_n = b_n, \quad A_n = \frac{1}{n}\sum_{i=1}^n \Phi(s_i)\left[\beta\Phi(s_i) - \mathcal{L}_{\bar{\mu}_1^i, \bar{\Sigma}_1^i}\Phi(s_i)\right]^\top, \quad b_n = \frac{1}{n}\sum_{i=1}^n r(s_i)\Phi(s_i), \tag{27}$$

where

$$\bar{\mu}_1^i = \frac{1}{\Delta t}(s_i' - s_i), \quad \bar{\Sigma}_1^i = \frac{1}{\Delta t}(s_i' - s_i)(s_i' - s_i)^\top.$$

This estimator targets the Galerkin solution $\hat{V}_1^G(s) = \Phi(s)^\top \theta$ of the first-order PhiBE defined in (24) when the discrete-time transition dynamics are known. Here θ solves

$$A\theta = b, \quad A = \int \Phi(s)\left[\beta\Phi(s) - \mathcal{L}_{\hat{\mu}_1, \hat{\Sigma}_1}\Phi(s)\right]^\top \rho(s)ds, \quad b = \int r(s)\Phi(s)\rho(s)ds. \tag{28}$$

The theorem below establishes both consistency and finite-sample guarantees.

Theorem 19 (Sample Complexity) *Assume that* $\left\|\nabla^k \Phi\right\| \le L_\Phi$ *ρ-a.s. for* $k = 0, 1, 2$ *and that the Gram matrix for the basis functions satisfies*

$$G_\Phi = \int \Phi(s)\Phi(s)^\top \rho(s)\, ds \succeq \lambda_\Phi I. \tag{29}$$

Suppose further that $\Delta t \le \eta_{\mu,\Sigma,\beta}$ *(the same* $\eta_{\mu,\Sigma,\beta}$ *as in Theorem 16). Then, for any sufficiently small* $\epsilon > 0$*, if*

$$n \gtrsim \frac{\log(2p/\delta) R^2 L_\Phi^8 \sigma_{\max}^2}{\beta^4 \epsilon^2 \Delta t \lambda_\Phi^4},$$

where $R = \|r\|_{L^\infty}$ *and* $\sigma_{\max} = \|\sigma\|_{L^\infty}$*, then, with probability at least* $1 - \delta$*, we have*

$$\|\hat{V}_n - \hat{V}_1^G\|_{L^\infty} \le \varepsilon.$$

Equivalently, if $n \gtrsim \frac{L_\Phi^4 \log(2p/\delta)\sigma_{\max}^2 \lambda_\Phi^{-2}}{\beta^2 \Delta t}$*, then, with probability* $1-\delta$*,*

$$\left\| V_n - \hat{V}_1^G \right\|_{L^\infty} \leq \left(L_\Phi^4 R \sigma_{\max} \lambda_\Phi^{-2} \sqrt{\log(2p/\delta)} \right) \frac{1}{\beta^2 \sqrt{n \Delta t}}.$$

The proof of the above theorem is given in Section 6.7. Several remarks are in order. The model-free Galerkin estimator analyzed in Theorem 19 is a continuous-time counterpart of LSTD, as discussed in Remark 18. This connection motivates a comparison of their finite-sample estimation errors. To express the estimation term in Theorem 1 of Lazaric et al. (2012) under the same continuous-time scaling, let $\gamma_{\Delta t} = e^{-\beta \Delta t}$ and $\|r_{\Delta t}\|_\infty = O(R\Delta t)$. Then $V_{\max} = \|r_{\Delta t}\|_\infty / (1-\gamma_{\Delta t}) = O(R/\beta)$, and their term $O((1-\gamma_{\Delta t})^{-1} V_{\max} n^{-1/2})$ becomes $O(R\beta^{-2}\Delta t^{-1} n^{-1/2})$. The corresponding sample-complexity bound is therefore of order $O(\epsilon^{-2}\beta^{-4}\Delta t^{-2})$. This bound need not be sharp when the transition kernel is induced by an SDE; rather, it shows that a direct application of the general LSTD sample complexity analysis to the continuous-time setting yields a Δt^{-2} dependence.

By comparison, Theorem 19 gives a sample-complexity bound of order $O(\epsilon^{-2}\beta^{-4}\Delta t^{-1})$, improving the dependence on the data interval from Δt^{-2} to Δt^{-1}. Although this comparison does not establish that our estimator always has a smaller finite-sample error than LSTD, it demonstrates that our analysis is sharper in its dependence on Δt when the transition kernel is induced by an SDE. This improved dependence on the data interval is the key finite-sample contribution of Theorem 19.

Corollary 20 (Overall Error Bound) *Under the same assumptions as in Theorem 19, with probability at least* $1-\delta$*,*

$$\|V_n - V(s)\|_\rho \lesssim \frac{C_{r,\mu,\Sigma}}{\beta^2}\Delta t + \frac{C_G}{\beta} \inf_{V^P = \theta^\top \Phi} \left\| \hat{V}_i - V^P \right\|_{H^1_{\rho,\infty}} + \frac{C_{r,\Sigma,\Phi}}{\beta^2 \sqrt{n\Delta t}},$$

where $C_{r,\Sigma,\Phi} = L_\Phi^4 R \sigma_{\max} \lambda_\Phi^{-2} \sqrt{\log(2p/\delta)}$*, and* $C_{r,\mu,\Sigma}, C_G$ *are the same constants as in Theorems 14 and 16.*

Several remarks are in order. The total error consists of three components: a discretization error, a model approximation error, and a finite-sample error. Among these terms, only the model error is independent of the time step Δt. The discretization and sample errors exhibit opposite dependencies on the data-collection frequency Δt, i.e.,

$$\|\hat{V}_n - V\| = O\left(\frac{\Delta t}{\beta^2} + \frac{1}{\beta^2 \sqrt{n\Delta t}} \right).$$

This bound highlights a fundamental trade-off inherent in continuous-time PE. Decreasing Δt reduces the discretization bias but increases the variance of the estimator, thereby requiring a larger sample size to maintain the same level of accuracy. As a result, excessively frequent sampling does not necessarily improve overall performance.

This phenomenon is not specific to the PhiBE framework but is intrinsic to continuous-time reinforcement learning. As illustrated in Figure 6 of our numerical experiments, the model-free RL algorithm (LSTD) exhibits a similar trade-off. Similar behavior was observed in (Zhang et al., 2023), which studied a Monte Carlo sampling algorithm for linear–quadratic

(LQ) policy evaluation. Their theoretical analysis is restricted to the LQ setting, with nonlinear dynamics examined only through numerical experiments. In contrast, our analysis applies to general controlled diffusion processes and provides explicit non-asymptotic error bounds. Moreover, their approach evaluates the value function only at discrete states and does not involve function approximation, whereas our algorithm produces a global approximation of the value function via function approximation, leading to a substantially more involved analysis. Finally, our results explicitly characterize the dependence on the discount parameter β, which is not addressed in Zhang et al. (2023).

5. Numerical Experiments

All MATLAB code used to generate the numerical results is available.[1]

5.1 Deterministic Dynamics

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

$$\frac{d}{dt}s_t = \lambda s_t, \tag{30}$$

and the other is nonlinear,

$$\frac{d}{dt}s_t = \lambda \sin^2(s_t). \tag{31}$$

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, for which the exact value function is $V(s) = \cos^3(ks)$ in both cases. We use periodic basis functions $\{\phi_k(s_1)\}_{k=1}^{2M+1} = \frac{1}{\sqrt{\pi}}\{\frac{1}{\sqrt{2}}, \cos(ms_1), \sin(ms_1)\}_{m=1}^{M}$ with M large enough so that the solution can be accurately represented by this finite basis ($M = 4$ for the low-frequency value function and $M = 30$ for the high-frequency value function).

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

$$\tilde{V}(s) = r(s)\Delta t + e^{-\beta\Delta t}\tilde{V}(e^{\lambda\Delta t}s), \tag{32}$$

and the ith-order PhiBE as

$$\beta\hat{V}_i(s) = r(s) + \frac{1}{\Delta t}\left[\sum_{k=1}^{i} a_k^{(i)}(e^{\lambda k\Delta t}s - s)\right]\cdot\nabla\hat{V}_i(s), \tag{33}$$

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

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

1. The repository is available at `https://github.com/yuhuazhumath/phibe-matlab`.

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 (4),

$$\tilde{A} = \sum_{l=1}^{I} \sum_{j=0}^{m-2} \Phi(s^l_{j\Delta t}) \left[\Phi(s^l_{j\Delta t}) - e^{-\beta \Delta t} \Phi(s^l_{(j+1)\Delta t})\right]^\top \tag{34}$$

instead of $\bar{A}_i$.

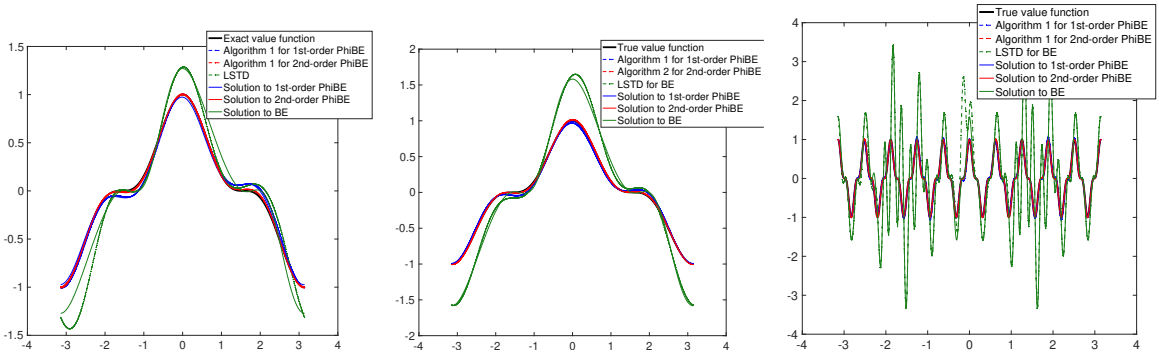
In Figure 1, the data are generated from the linear dynamics (30) with $\lambda = 0.05$ and collected at different Δt. For the low-frequency value function (Figure 1/(a), (c)), we use 40 data points with $J = 10, m = 4$, whereas for the high-frequency value function (Figure 1/(b)), we use 400 data points with $J = 100, m = 4$. We compare the second-order PhiBE solution with the BE solution (when the discrete-time transition dynamics are known) and the performance of LSTD with that of the proposed Algorithm 1 (when only trajectory data are available) under different data-collection intervals Δt, discount coefficients β, and reward oscillations 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 $I = 500/\Delta t$ large enough, and the exact solution to PhiBE is calculated by applying the Galerkin method to (33).

In Figure 3, the data are generated from the nonlinear dynamics (31) and collected at different Δt. For the low-frequency value function (Figure 3/(a), (c)), we use 80 data points with $J = 20, m = 4$, whereas for the high-frequency value function (Figure 3/(b)), we use 400 data points with $J = 100, m = 4$. We compare the first- and second-order PhiBE solutions with the BE solution (when the discrete-time transition dynamics are known) and the performance of LSTD with that of the proposed Algorithm 1 (when only trajectory data are available) under different values of Δt, β, k, and λ.

In Figure 4, the distances from the PhiBE and BE solutions to the true value function are plotted as $\Delta t \to 0$; the distances from the solutions approximated by Algorithm 1 and LSTD to the true value function are plotted as the amount of data increases. We set $J = [10, 10^2, 10^3, 10^3]$ and $m = 4$ for the data size. Here, the distance is measured using the L^2 norm

$$D(V, \hat{V}) = \sqrt{\int_{-\pi}^{\pi} (V(s) - \hat{V}(s))^2 ds}. \tag{35}$$

In Figures 1 and 3, when the discrete-time transition dynamics are known, the PhiBE solution is much closer to the true value function than the BE solution in all the experiments. In particular, the second-order PhiBE solution is almost identical to the exact value function. Additionally, when only trajectory data are available, one can approximate the solutions to PhiBE very well with only 40 or 400 data points. In particular, 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 using PhiBE. Moreover, the solution to PhiBE is not sensitive to oscillations in the reward function, which implies greater flexibility in designing the reward function in the RL problem. The BE error increases when β is too small or too large, whereas the PhiBE error decays as β increases. Furthermore, it is noteworthy that in Figure 3/(b)



(a) Nonlinear dynamics (31) with $\Delta t = 5, \beta = 0.1, k = 1, \lambda = 0.1$

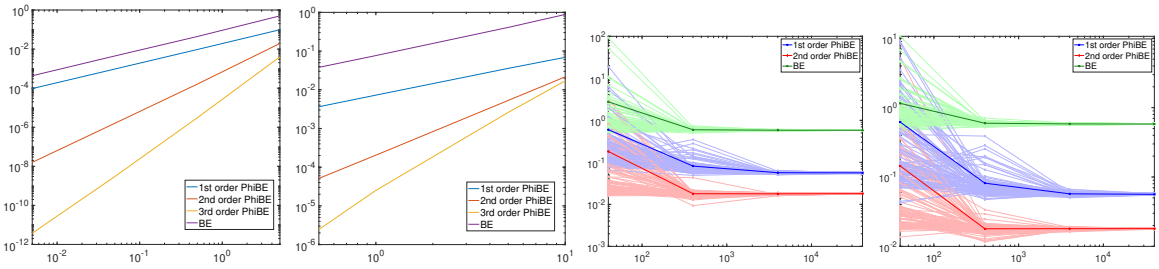
(b) Nonlinear dynamics (31) with $\Delta t = 0.1, \beta = 10, k = 1, \lambda = 5$

(c) Nonlinear dynamics (31) with $\Delta t = 0.1, \beta = 10, k = 10, \lambda = 2$

Figure 3: The PhiBE solution and the BE solution, when the discrete-time transition dynamics are given, are plotted in solid lines. The approximate PhiBE solution based on Algorithm 1 and the approximate BE solution based on LSTD, when discrete-time data are given, are plotted in dashed lines. Both algorithms utilize the same data points.

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 4/(a) and (b), one can observe that the solution for BE approximates the true solution with first-order accuracy, while the solution for the ith-order PhiBE approximates the true solution with ith-order accuracy. In Figure 4/(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 reach $O(\Delta t^i)$.



(a) Linear dynamics (31). (b) Nonlinear dynamics (31) (c) Linear dynamics (31). (d) Nonlinear dynamics (31).

Figure 4: The L^2 errors (35) of the PhiBE and BE solutions with decreasing Δt are plotted in the two left panels. The L^2 errors (35) of the approximate PhiBE and BE solutions with increasing amounts of data collected every $\Delta t = 5$ units of time are plotted in the two right panels. The solid lines are averages over 100 simulations for all data sizes. We set $\lambda = 0.05, \beta = 0.1, k = 1$ in both the linear and nonlinear cases.

5.2 Stochastic Dynamics

We consider the Ornstein–Uhlenbeck process,

$$ds(t) = \lambda s dt + \sigma dB_t, \tag{36}$$

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

$$\begin{aligned} \beta\hat{V}_i(s) =& r(s) + \frac{1}{\Delta t}\sum_{k=1}^{i} a_k^{(i)}(e^{\lambda k\Delta t} - 1)s\nabla\hat{V}(s) \\ &+ \frac{1}{2\Delta t}\sum_{k=1}^{i} a_k^{(i)}\left[\frac{\sigma^2}{2\lambda}(e^{2\lambda k\Delta t} - 1) + (e^{\lambda k\Delta t} - 1)^2 s^2\right]\Delta\hat{V}(s); \end{aligned} \tag{37}$$

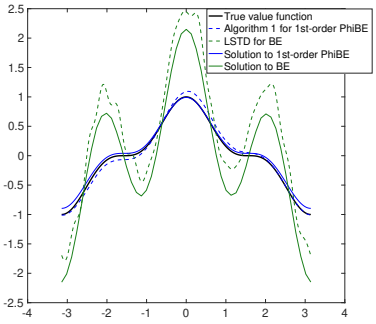
and BE as

$$\begin{aligned} \tilde{V}(s) =& r(s)\Delta t + e^{-\beta\Delta t}\mathbb{E}\left[\tilde{V}(s_{t+1})|s_t = s\right] \\ =& r(s)\Delta t + e^{-\beta\Delta t}\int_{\mathbb{S}} \tilde{V}(s')\rho_{\Delta t}(s', s)ds', \end{aligned} \tag{38}$$

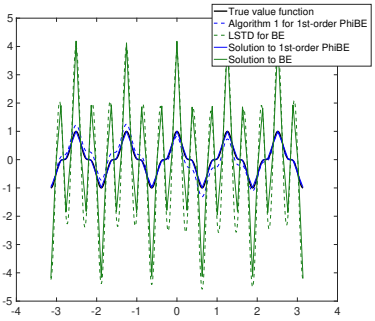
where

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

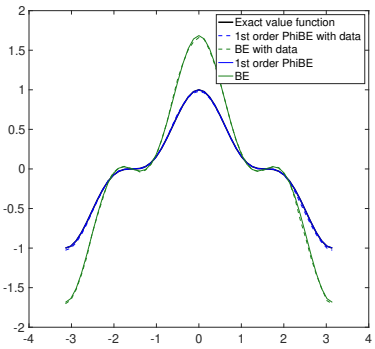
In Figure 5, we compare the exact and approximate solutions to PhiBE and BE, respectively, for different values of Δt, β, and k. For the approximate solution, we set the data size $m = 4$, and $J = 10^4$ for Figure 5/(a), $J = 10^5$ for Figure 5/(b), and $J = 100$



(a) Stochastic dynamics (36) with $\Delta t = 1, \beta = 0.1, k = 1$.



(b) Stochastic dynamics (36) with $\Delta t = 0.1, \beta = 0.1, k = 5$.



(c) Stochastic dynamics (36) with $\Delta t = 0.1, \beta = 10, k = 1$.

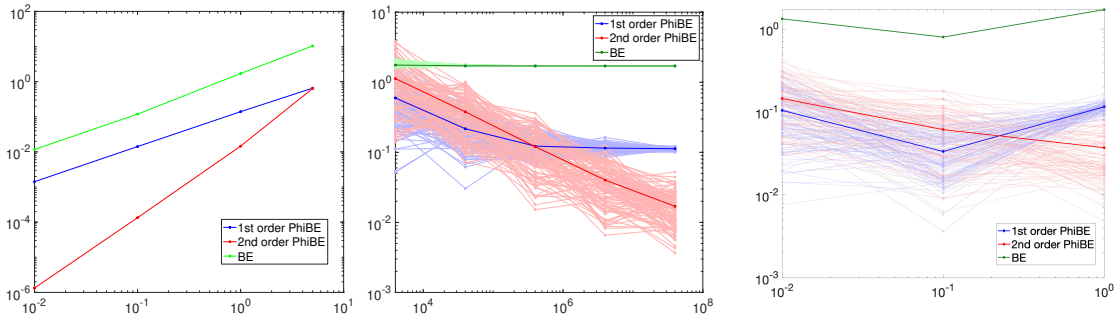
Figure 5: The PhiBE solution and the BE solution, when the discrete-time transition dynamics are given, are plotted in solid lines. The approximate PhiBE solution based on Algorithm 2 and the approximate BE solution based on LSTD, when discrete-time data are given, are plotted in dashed lines. Both algorithms utilize the same data points.

for Figure 5/(c). In Figure 6/(a), the errors of the exact PhiBE and BE solutions are plotted as $\Delta t \to 0$. In Figure 6/(b), the errors of the approximate PhiBE and BE solutions based on Algorithm 2 and LSTD are plotted as the amount of data increases. We set $J = [10^3, 10^4, 10^5, 10^6, 10^7]$ and $m = 4$ for the data size. In Figure 6/(c), the errors of the approximate PhiBE and BE solutions based on Algorithm 2 and LSTD are plotted as Δt decreases. We set $J = 10^6$ and $m = 4$ for the data size.

We observe similar performance under stochastic and deterministic dynamics, as shown in Figures 5 and 6. In Figure 6, 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. In particular, when the amount of data exceeds 10^6, the variance is smaller than 10^{-1}. Figure 6/(c) illustrates a non-monotone dependence of the error on the time discretization Δt when the total data budget is fixed. For the data-driven RL algorithm and the first-order PhiBE method, the error initially decreases as Δt is reduced from 1 to 0.1, reflecting a regime in which discretization error dominates. As Δt is further reduced from 0.1 to 0.01, the error increases, indicating that finite-sample error becomes dominant. The second-order PhiBE method, which exhibits higher variance, enters the variance-dominated regime earlier: its error already increases as Δt decreases from 1 to 0.1.

5.3 Stabilization Problem

In the previous section, we compared the PhiBE framework with the standard RL formulation and validated our theoretical results using a synthetic example. We now examine their performance on a real control problem.



(a) Discretization error w.r.t. data collection frequency Δt.

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

(c) Approximation error w.r.t data collection frequencyΔt.

Figure 6: The L^2 error (35) of the PhiBE and BE solutions as a function of the time discretization Δt is shown in (a). Plot (b) reports the L^2 error of the data-driven approximations as the amount of data increases, with samples collected every $\Delta t = 1$ unit of time. Plot (c) shows the L^2 error of the data-driven approximations with a fixed budget of 4×10^6 data points as Δt decreases. Solid lines denote averages over 100 independent simulations. In all plots, we set $\beta = 0.1$ and $k = 1$.

Problem setup. We consider the continuous-time PE problem

$$\begin{aligned} V(s) &= \mathbb{E}\left[\int_0^\infty e^{-\beta t}(qs_t^2 + ru(s_t)^2)|s_0 = s\right] \\ dS_t &= \left(-\kappa s_t^3 + (\alpha s_t - bu(s_t)\right)\, dt + \sigma\, dW_t, \end{aligned} \tag{39}$$

with parameters $\kappa > 0$, $\alpha \in \mathbb{R}$, $b > 0$, and $\sigma > 0$. We focus on the linear feedback policy

$$u(s) = Ks.$$

Under this policy, the true value function V solves the elliptic equation

$$\beta V(s) - \mathcal{L}V(s) = r(s), \qquad \mathcal{L}f(s) = \Big(-\kappa s^3 + (\alpha - bK)s\Big) f'(s) + \frac{\sigma^2}{2} f''(s), \tag{40}$$

where $r(s) = qs^2 + rK^2 s^2$.

This model represents a standard stabilization problem, where the quadratic reward balances state deviation and control effort, and the linear policy $u(s) = -Ks$ acts as a proportional controller. When $\kappa = 0$, the dynamics reduce to a linear system, corresponding to the classical continuous-time LQR formulation widely used in robotics, aerospace, and process control. In this regime, the value function admits a closed-form solution, providing a baseline against which the approximation accuracy of RL and PhiBE can be directly evaluated. When $\kappa \neq 0$, the cubic drift introduces strong nonlinear restoring forces, modeling phenomena such as soft-spring mechanical systems or the stabilization of an inverted pendulum near its upright position.

5.3.1 Linear Dynamics

In the linear-dynamics setting, the true value function admits a closed-form solution,

$$V(s) = a_1 s^2 + a_2, \quad a_1 = \frac{R}{\beta - 2\lambda}, \quad a_2 = \frac{\sigma^2}{\beta} a_1, \tag{41}$$

where $R = q + rK^2$ and $\lambda = \alpha - bK$. Moreover, for this linear system, the Bellman equation (BE) and PhiBE both admit exact analytical solutions. For the discrete-time BE, we have

$$\text{BE} : \tilde{V}(s) = Rs^2 \Delta t + \gamma \int V(s')\rho(s'|s)ds', \quad R = q + rK^2, \quad \gamma = e^{-\beta \Delta t},$$

where

$$\rho(s'|s) \sim \mathcal{N}(e^{\lambda \Delta t} s, \hat{\sigma}^2 \Delta t), \qquad \hat{\sigma}^2 = \frac{\sigma^2}{2\lambda \Delta t}(e^{2\lambda \Delta t} - 1).$$

This yields the quadratic solution for BE,

$$\tilde{V}(s) = a_1^R s^2 + a_0^R, \quad a_1^R = \frac{R\Delta t}{1 - \gamma e^{2\lambda \Delta t}}, \quad a_0^R = \frac{\gamma \Delta t}{1 - \gamma} \hat{\sigma}^2 a_1^R. \tag{42}$$

Similarly, the PhiBE equation can be evaluated exactly:

$$\begin{aligned} \text{PhiBE} : \hat{V}(s) =& Rs^2 + \hat{V}'(s) \int \frac{s' - s}{\Delta t} \rho(s'|s)ds' + \frac{1}{2}\hat{V}''(s) \int \frac{(s' - s)^2}{\Delta t} \rho(s'|s)ds' \\ =& Rs^2 + \hat{\lambda} s \hat{V}'(s) + \frac{\hat{\sigma}^2 + \hat{\lambda}^2 s^2 \Delta t}{2} \hat{V}''(s), \quad \hat{\lambda} = \frac{e^{\lambda \Delta t} - 1}{\Delta t}. \end{aligned}$$

This yields another quadratic solution,

$$\hat{V}(s) = a_1^P s^2 + a_0^P, \quad a_1^P = \frac{R}{\beta - 2\hat{\lambda} - \eta}, \quad a_0^P = \frac{\hat{\sigma}^2}{\beta} a_1^P, \tag{43}$$

where $\eta = \frac{1}{\Delta t}(e^{2\lambda \Delta t} - 2e^{\lambda \Delta t} + 1)$.

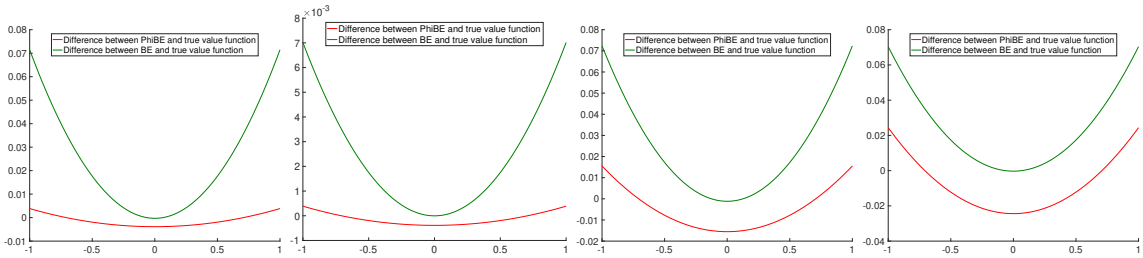
Figure 7 compares the analytical BE and PhiBE solutions, $\widetilde{V}$ and $\widehat{V}$, with the true value function $V(s)$ across four test cases. Throughout all experiments we set $q = 1$, $r = 0.1$, and $K = 2$. The four configurations are:

$$\begin{aligned} &\text{case 1} \quad \text{Baseline: } \alpha = b = \frac{1}{4}, \sigma = 0.5, \beta = 1, \Delta t = 0.1. \\ &\text{case 2} \quad \text{More frequent observations (smaller } \Delta t\text{): } \alpha = b = \frac{1}{4}, \sigma = 0.5, \beta = 1, \Delta t = 0.01. \\ &\text{case 3} \quad \text{Quicker dynamics: } \alpha = b = 1, \sigma = 1, \beta = 1, \Delta t = 0.1. \\ &\text{case 4} \quad \text{Less discounted (smaller } \beta\text{): } \alpha = b = \frac{1}{4}, \sigma = 0.5, \beta = 0.1, \Delta t = 0.1. \end{aligned} \tag{44}$$

From Figure 7, we observe that reducing Δt by one order decreases the approximation error of both formulations by approximately one order, consistent with the theoretical convergence rates. Moreover, as the system dynamics become faster or the discount factor β becomes smaller, the advantage of PhiBE over BE diminishes, although PhiBE still

	mean (PhiBE)	mean (BE)	variance (PhiBE)	variance (BE)
Baseline ($n = 10^6$)	4.76×10^{-3}	4.54×10^{-2}	2.7×10^{-6}	1.02×10^{-5}
Smaller Δt ($n = 10^6$)	9.64×10^{-3}	1.04×10^{-2}	3.84×10^{-5}	4.3×10^{-5}
Smaller Δt ($n = 10^7$)	2.91×10^{-3}	4.7×10^{-3}	3.6×10^{-6}	8.1×10^{-6}
Quick dynamics ($n = 10^6$)	3.45×10^{-2}	5.5×10^{-2}	3.87×10^{-4}	8.07×10^{-4}
Smaller β ($n = 10^6$)	7.48×10^{-2}	7.81×10^{-2}	2.11×10^{-3}	2.93×10^{-3}
Smaller β ($n = 10^7$)	3.41×10^{-2}	4.97×10^{-2}	1.97×10^{-4}	3.89×10^{-4}

Table 1: Comparison of the approximation error for the model-free PhiBE solver and the LSTD method under linear dynamics.



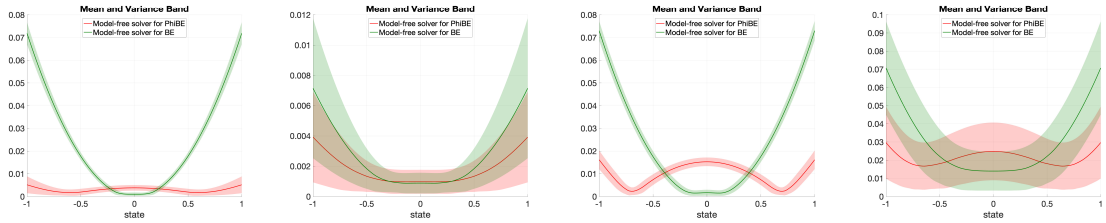
(a) Baseline (b) Smaller Δt (c) Quicker dynamics (d) Smaller β

Figure 7: Stabilization control under linear dynamics. The error $\hat{V}(s) - V(s)$ from the PhiBE solution (red) and $\tilde{V}(s) - V(s)$ from the BE solution (green) are plotted, where $V(s)$, $\widehat{V}(s)$, and $\widetilde{V}(s)$ are given in (41), (43), and (42). Each subplot corresponds to one scenario in (44).

performs better overall. This observation aligns with our theoretical analysis: (1) PhiBE achieves greater improvement when the dynamics evolve more slowly, and (2) the approximation error of PhiBE scales as $O(1/\beta^2)$, whereas that of BE scales as $O(1/\beta)$ when β is small.

We next evaluate the data-driven algorithms. To solve the PhiBE formulation, we use the first-order method in Algorithm 2; for the BE formulation, we apply LSTD as in (34). We generate $\{s_i\}_{i=1}^n$ from $[-1, 1]$ on the uniform mesh, and $s_i' \sim \mathcal{N}(e^{\lambda \Delta t} s_i, \hat{\sigma}^2 \Delta t)$ for all $1 \le i \le n$, consistent with the linear system dynamics. The mean and variance of the approximation error, computed over 100 Monte Carlo repetitions for all four cases, are shown in Figure 8.

Figure 8 demonstrates that the data-driven estimators concentrate around the corresponding analytic solutions in Figure 7 when the sample size is sufficiently large. However, the variability differs across scenarios. To examine this more closely, Table 1 reports the mean and variance of the approximation error. Two observations stand out from Table 1. First, across all test cases, PhiBE consistently attains smaller mean error and smaller variance compared with BE given the same amount of data. Second, for a fixed sample size, the variance increases for both methods as Δt decreases, β decreases, or the dynamics become faster.



(a) Baseline: $n = 10^6$ (b) Smaller Δt: $n = 10^7$ (c) Quicker dynamics: $n = 10^6$ (d) Smaller β: $n = 10^7$

Figure 8: Stabilization control under linear dynamics. The mean and variance of the approximation error $|\hat{V}(s) - V(s)|$ from the PhiBE algorithm (Algorithm 2) are shown in red, and those of $\tilde{V}(s) - V(s)$ from the LSTD method (34) are shown in green. Each subplot corresponds to one scenario in (44), illustrating how the error behaves under different values of Δt, β, and the system dynamics.

5.3.2 Nonlinear Dynamics

In this section, we demonstrate that our approach also applies to nonlinear dynamics. In this setting, the true value function is no longer available in closed form, so we approximate the ground-truth solution using a high-accuracy Galerkin method with sufficiently many polynomial basis functions. The data collection procedure follows the linear case, except that we generate each s_i' by simulating the SDE using the Euler-Maruyama scheme with step size $\delta_t = 10^{-3}$.

We consider the following four cases, setting $q = 1$, $r = 0.1$, $K = 2$, and $n = 10^6$ for all experiments:

$$\begin{aligned}
&\text{case 1} \quad \text{Baseline: } \alpha = b = \kappa = 0.1, \sigma = 0.05, \beta = 1, \Delta t = 0.1;\\
&\text{case 2} \quad \text{More frequent observations (smaller } \Delta t\text{): } \alpha = b = \kappa = 0.1, \sigma = 0.05, \beta = 1, \Delta t = 0.01;\\
&\text{case 3} \quad \text{Quicker dynamics: } \alpha = b = \kappa = 1/2, \sigma = 1/4, \beta = 1, \Delta t = 0.1;\\
&\text{case 4} \quad \text{Less discounted (smaller } \beta\text{): } \alpha = b = \kappa = 0.1, \sigma = 0.05, \beta = 0.1, \Delta t = 0.1.
\end{aligned} \tag{45}$$

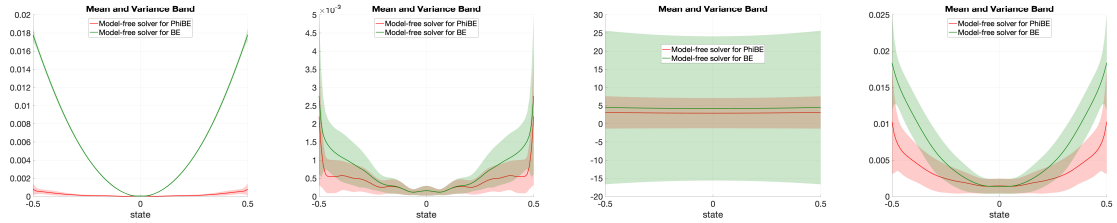
The resulting approximation errors are shown in Figure 9. The qualitative behavior is similar to the linear-dynamics experiments: PhiBE exhibits smaller mean error and more stable performance across all cases.

One caveat appears in Case 3, where both methods exhibit a few significant outliers. These outliers inflate the empirical variance, especially for LSTD, which in one run produced an error nearly two orders of magnitude larger than the average. This behavior is sensitive to the random sample, and a systematic analysis of such instability phenomena is left for future work.

5.4 High-Dimensional Case

We next evaluate the algorithms on a 10-dimensional linear-quadratic PE problem:

$$\begin{aligned}
V(s) &= \mathbb{E}\left[\int_0^\infty e^{-\beta t}(s_t^\top Q s_t)dt | s_0 = s\right]\\
ds_t &= As_t\,dt + \sigma\,dW_t,
\end{aligned}$$



(a) Baseline (b) Smaller Δt (c) Quicker dynamics (d) Smaller β

Figure 9: Stabilization control under nonlinear dynamics. The mean and variance of the approximation error $|\hat{V}(s) - V(s)|$ from the PhiBE algorithm (Algorithm 2) are shown in red, and those of $\tilde{V}(s) - V(s)$ from the LSTD method (34) are shown in green. Each subplot corresponds to one scenario in (45), illustrating how the error behaves under different values of Δt, β, and the system dynamics.

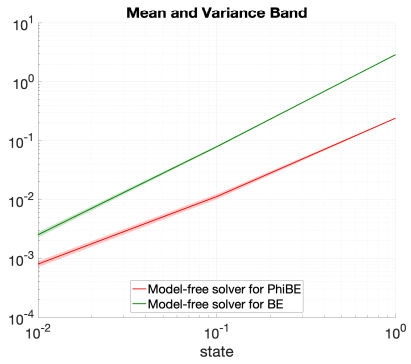


Figure 10: Stabilization control under linear dynamics with high-dimensional state space $s \in \mathbb{R}^{10}$. The mean and variance of the approximation error $\left\|\hat{V}(s) - V(s)\right\|_\rho$ from the PhiBE algorithm (Algorithm 2) are shown in red, and those of $\left\|\tilde{V}(s) - V(s)\right\|_\rho$ from the LSTD method (34) are shown in green.

where $Q \in \mathbb{R}^{10\times 10}$ is constructed to have eigenvalues $(1, \ldots, 10)$ by generating a random orthonormal matrix O and setting $Q = O^\top \mathrm{diag}(1, \ldots, 10)\, O$. The drift matrix $A \in \mathbb{R}^{10\times 10}$ is generated analogously with eigenvalues $-(1, \ldots, 10) \times 0.1$. The diffusion matrix $\Sigma \in \mathbb{R}^{10\times 10}$ is a diagonal matrix with all diagonal entries equal to 0.3, and we set $\beta = 1$.

We use polynomial basis functions up to second order. The dataset consists of samples $(s_i, s_i')_{i=1}^n$, where $s_i \sim \mathrm{Unif}([-1, 1]^{10})$, and s_i' is obtained by simulating the SDE for one step of size Δt starting from s_i. Since achieving a smaller discretization error at smaller Δt requires more samples, we choose $n = 10^5,\ 10^6,\ 10^8$ for $\Delta t = 1,\ 0.1,\ 0.01$, respectively.

Figure 10 reports the results for LSTD and the model-free PhiBE algorithm. The findings closely parallel the one-dimensional case: both error curves decay approximately at first order in Δt, confirming that the approach scales effectively to higher dimensions and that the theoretical predictions remain valid in the high-dimensional setting.

6. Proofs

6.1 Proof of Theorem 3.

Let $\rho(s', t|s)$ be the probability density function of s_t starting from $s_0 = s$. It satisfies the following PDE

$$\partial_t \rho(s', t|s) = \nabla \cdot [\mu(s')\rho(s', t|s)] + \frac{1}{2}\sum_{i,j} \partial_{s_i}\partial_{s_j}[\Sigma_{ij}(s')\rho(s', t|s)]. \tag{46}$$

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

$$\begin{aligned} V(s) - \tilde{V}(s) &= \mathbb{E}\left[\sum_{i=0}^{\infty} \int_{\Delta t i}^{\Delta t(i+1)} f(t, s_t) - f(\Delta t i, s_{\Delta t i}) dt | s_0 = s\right] \\ &= \sum_{i=0}^{\infty} \int_{\Delta t i}^{\Delta t(i+1)} \left(\int_{\mathbb{S}} f(t, s')\rho(s', t|s) - f(\Delta t i, s')\rho(\Delta t i, s') ds' \right) dt. \end{aligned} \tag{47}$$

Since

$$\begin{aligned} &\int_{\mathbb{S}} f(t, s')\rho(s', t|s) - f(\Delta t i, s')\rho(\Delta t i, s') ds' \\ =& \int_{\mathbb{S}} f(t, s')(\rho(s', t|s) - \rho(s', \Delta t i|s)) + (f(t, s') - f(\Delta t i, s'))\rho(s', \Delta t i|s) ds' \\ =& \int_{\mathbb{S}} f(t, s')\partial_t \rho(s', \xi_1|s)(t - \Delta t i) + \partial_t f(\xi_2, s')(t - \Delta t i)\rho(s', \Delta t i|s) ds' \\ &\text{where } \xi_1, \xi_2 \in (\Delta t i, \Delta t(i+1)) \\ =& \int_{\mathbb{S}} \mathcal{L}_{\mu,\Sigma} f(t, s'), \rho(s', \xi_1|s)(t - \Delta t i) ds' - \int_{\mathbb{S}} \beta e^{-\beta \xi_2} r(s')\rho(s', \Delta t i|s)(t - \Delta t i) ds' \\ =& \left(e^{-\beta t} \int_{\mathbb{S}} \mathcal{L}_{\mu,\Sigma} r(s')\rho(s', \xi_1|s) ds' - \beta e^{-\beta \xi_2} \int_{\mathbb{S}} r(s')\rho(s', \Delta t i|s) ds' \right)(t - \Delta t i), \end{aligned} \tag{48}$$

where the second equality is due to the mean value theorem, and the third equality is obtained by inserting the equation (46) for $\rho(s', t|s)$ and integrating by parts. Therefore, for $t \in [\Delta t i, \Delta t(i+1)]$,

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

Therefore, one has

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

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 7

Note that the true value function V and the ith-order PhiBE solution $\hat{V}$ satisfy

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

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 21 *For functions V and $\hat{V}$ satisfying*

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

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 B.1 for the proof of the above lemma.) Therefore, one has

$$\left\|V-\hat{V}_i\right\|_{L^\infty}\leq\frac{\left\|\mu-\hat{\mu}_i\right\|_{L^\infty}\left\|\nabla r\right\|_{L^\infty}}{(\beta-\left\|\nabla\mu\right\|_{L^\infty})^2}.\tag{49}$$

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

Lemma 22 *The distance between $\hat{\mu}_i(s)$ defined in* (12) *and the true dynamics can be bounded by*

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

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

$$C_i=\frac{\sum_{j=0}^{i}|a_j^{(i)}|j^{i+1}}{(i+1)!}.\tag{50}$$

(See Section B.2 for the proof of the above lemma.) Hence, applying the above lemma to (49) completes the proof:

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

6.3 Proof of Theorem 8

For the linear dynamics $ds_t = As_t dt$, one has $s_t = e^{At} s_0$; therefore,

$$\hat{\mu}(s) = \frac{1}{\Delta t} \sum_{j=1}^{i} a_j^{(i)} (s_{jdt} - s_0) = \frac{1}{\Delta t} \sum_{j=1}^{i} a_j^{(i)} (e^{Aj\Delta t} - I)s.$$

Let $\hat{A}_i = a_j^{(i)}(e^{Aj\Delta t} - I)$. Then the ith-order PhiBE reads

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

which can be equivalently written as

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

Now define $\phi(\theta) = e^{At} e^{\theta \Delta_i t} s$, where $\Delta_i = \hat{A} - A$, then $e^{At}s = \phi(0), \quad e^{\hat{A}t}s = \phi(1)$, which implies

$$\begin{aligned} &r(e^{At}s) - r(e^{\hat{A}t}s) = r(\phi(0)) - r(\phi(1)) = \int_0^1 \frac{d}{d\theta} r(\phi(\theta)) d\theta = \int_0^1 \nabla (r(\phi(\theta)))^\top \phi'(\theta) d\theta \\ =& \left[\int_0^1 (\nabla r(\phi(\theta)))^\top \phi(\theta) d\theta\right] \Delta_i t. \end{aligned}$$

Therefore, one has

$$\begin{aligned} &\left|V(s) - \hat{V}_1(s)\right| = \left|\int_0^\infty e^{-\beta t} \left(r(e^{At}s) - r(e^{\hat{A}t}s)\right) dt\right| \\ =& \left|\int_0^1 \left(\int_0^\infty (\nabla r(\phi(\theta)))^\top \phi(\theta) e^{-\beta t} t dt\right) d\theta \Delta_i\right| \leq \|u \cdot r(u)\|_{L^\infty} \int_0^1 \left(\int_0^\infty e^{-\beta t} t dt\right) d\theta \, |\Delta_A| \\ =& \frac{\|u \cdot r(u)\|_{L^\infty}}{\beta^2} \|\Delta_i\|. \end{aligned}$$

Next we estimate $\|\Delta_i\|$. First note that

$$\hat{A}_i = \frac{1}{\Delta t} \sum_{j=1}^{i} a_j^{(i)} (e^{Aj\Delta t} - I) = \frac{1}{\Delta t} \sum_{j=1}^{i} a_j^{(i)} \left(\sum_{k=1}^{i} \frac{1}{k!} (Aj\Delta t)^k + R_{ij}\right)$$

where

$$R_{ij} = e^{Aj\Delta t} - \sum_{k=0}^{i} \frac{1}{k!} (Aj\Delta t)^k = \frac{A^{i+1}(j\Delta t)^{i+1}}{(i+1)!} e^{A\xi}, \quad \text{for } \xi \in [0, j\Delta t),$$

and therefore

$$\|R_{ij}\| \leq \frac{\|A\|^{i+1}(j\Delta t)^{i+1}}{(i+1)!} e^{\|A\| j\Delta t} \leq \frac{\|A\|^{i+1}(j\Delta t)^{i+1}}{(i+1)!} e^{\|A\| i\Delta t}.$$

By the definition of $a_j^{(i)}$, one has

$$\hat{A}_i = \sum_{k=1}^{i} \frac{1}{k!} A^k \Delta t^{k-1} \left(\sum_{j=1}^{i} a_j^{(i)} j^k \right) + \frac{1}{\Delta t} \sum_{j=1}^{i} a_j^{(i)} R_{ij} = A + \frac{1}{\Delta t} \sum_{j=1}^{i} a_j^{(i)} R_{ij},$$

which leads to

$$\|\Delta_i\| = \left\| \hat{A}_i - A \right\| \leq \frac{1}{\Delta t} \sum_{j=1}^{i} \left| a_j^{(i)} \right| \|R_{ij}\| \leq C_i \|A\| D_A^i \Delta t^i, \quad \text{with } D_A = e^{\|A\| \Delta t} \|A\| \tag{51}$$

where C_i is defined in (50).

6.4 Proof of Theorem 13

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

Proposition 23 *For the operator $\mathcal{L}_{\mu,\Sigma}$ defined in* (5)*, under Assumption 1/(a), one has*

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

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

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

where $C_{\nabla\mu,\nabla\Sigma}$ is defined in (53) *depending on the first derivatives of μ, Σ.*

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

$$\begin{aligned}
&\langle \mathcal{L}_{\mu,\Sigma} V(s), V(s)\rangle_\rho = \langle \mu\cdot\nabla V, V\rangle_\rho - \frac{1}{2}\sum_{i,j}\left\langle \partial_{s_j}(\Sigma_{ij}V\rho), \partial_{s_i}V\right\rangle \\
=&\sum_i \left\langle \mu_i\rho, \partial_{s_i}\left(\frac{1}{2}V^2\right)\right\rangle - \frac{1}{2}\sum_{i,j}\left\langle \partial_{s_j}(\Sigma_{ij}\rho), \partial_{s_i}\left(\frac{1}{2}V^2\right)\right\rangle - \frac{1}{2}\sum_{i,j}\left\langle (\partial_{s_j}V)\Sigma_{ij}, \partial_{s_i}V\right\rangle_\rho \\
=&-\sum_i \left\langle \partial_{s_i}(\mu_i\rho), \frac{1}{2}V^2\right\rangle + \frac{1}{2}\sum_{i,j}\left\langle \partial_{s_i}\partial_{s_j}(\Sigma_{ij}\rho), \frac{1}{2}V^2\right\rangle - \frac{1}{2}\int (\nabla V)^\top \Sigma(\nabla V)\rho\, ds \\
=&\left\langle \nabla\cdot\left(-\mu\rho + \frac{1}{2}\nabla\cdot(\Sigma\rho)\right), \frac{1}{2}V^2\right\rangle - \frac{1}{2}\int (\nabla V)^\top \Sigma(\nabla V)\rho\, ds \\
\le&\left\langle \nabla\cdot\left(-\mu\rho + \frac{1}{2}\nabla\cdot(\Sigma\rho)\right), \frac{1}{2}V^2\right\rangle - \frac{\lambda_{\min}}{2}\|\nabla V\|_\rho^2,
\end{aligned} \tag{52}$$

where the last inequality is because of the positivity of the matrix $\Sigma(s)$ in Assumption 1.

If ρ is the stationary solution, then

$$\left\langle \nabla\cdot\left(-\mu\rho + \frac{1}{2}\nabla\cdot(\Sigma\rho)\right), \frac{1}{2}V^2\right\rangle = 0.$$

If ρ satisfies $\frac{1}{\lambda_{\min}}\left\|-\mu + \frac{1}{2}\nabla\cdot\Sigma + \frac{1}{2}\Sigma\nabla\log\rho\right\|_{L^\infty}^2 \le \frac{\beta}{2}$, then

$$\begin{aligned}
&\left\langle \nabla\cdot\left(-\mu\rho + \frac{1}{2}\nabla\cdot(\Sigma\rho)\right), \frac{1}{2}V^2\right\rangle = \left\langle \left(\mu - \frac{1}{2}\nabla\cdot\Sigma - \frac{1}{2}\Sigma\nabla\log\rho\right)\rho, \nabla\left(\frac{1}{2}V^2\right)\right\rangle \\
\le&\left\|\mu - \frac{1}{2}\nabla\cdot\Sigma - \frac{1}{2}\Sigma\nabla\log\rho\right\|_{L^\infty}\|V\|_\rho\|\nabla V\|_\rho \\
\le&\frac{1}{\lambda_{\min}}\left\|\mu - \frac{1}{2}\nabla\cdot\Sigma - \frac{1}{2}\Sigma\nabla\log\rho\right\|_{L^\infty}^2\|V\|_\rho^2 + \frac{\lambda_{\min}}{4}\|\nabla V\|_\rho^2 \le \frac{\beta}{2}\|V\|_\rho^2 + \frac{\lambda_{\min}}{4}\|\nabla V\|_\rho^2.
\end{aligned}$$

Substituting this into (52) completes the proof of the first inequality.

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

$$\begin{aligned}
&\sum_i \langle \partial_{s_i}\mathcal{L}_{\mu,\Sigma}V(s), V(s)\rangle_\rho \\
\le&\sum_i\left(\langle \partial_{s_i}\mu\cdot\nabla V, \partial_{s_i}V\rangle_\rho + \frac{1}{2}\langle \partial_{s_i}\Sigma : \nabla^2 V, \partial_{s_i}V\rangle_\rho\right) - \frac{\lambda_{\min}}{4}\sum_i\|\nabla\partial_{s_i}V\|_\rho^2 + \frac{\beta}{2}\|\nabla V\|_\rho^2 \\
\le&\|\nabla\mu\|_{L^\infty}\|\nabla V\|_\rho^2 + \frac{1}{2}\|\nabla\Sigma\|_{L^\infty}\left\|\nabla^2 V\right\|_\rho\|\nabla V\|_\rho - \frac{\lambda_{\min}}{4}\left\|\nabla^2 V\right\|_\rho^2 + \frac{\beta}{2}\|\nabla V\|_\rho^2 \\
=&-\frac{\lambda_{\min}}{4}\left(\left\|\nabla^2 V\right\|_\rho - \frac{\|\nabla\Sigma\|_{L^\infty}}{\lambda_{\min}}\|\nabla V\|_\rho\right)^2 + \left(\frac{\|\nabla\Sigma\|_{L^\infty}^2}{4\lambda_{\min}} + \|\nabla\mu\|_{L^\infty} + \frac{\beta}{2}\right)\|\nabla V\|_\rho^2 \\
\le&\frac{C_{\nabla\mu,\nabla\Sigma}}{2}\|\nabla V\|_\rho^2,
\end{aligned}$$

where

$$C_{\nabla\mu,\nabla\Sigma} = \frac{\|\nabla\Sigma\|_{L^\infty}^2}{2\lambda_{\min}} + 2\|\nabla\mu\|_{L^\infty} + \beta. \tag{53}$$

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 obtains the last two inequalities of the lemma.

■

Proof of Theorem 13 Now we are ready to prove Theorem 13.
Proof By (47), one has

$$\begin{aligned}
&\left\|V(s) - \tilde{V}(s)\right\|_\rho \\
\le&\sum_{i=0}^{\infty}\sqrt{\Delta t\int_{\Delta ti}^{\Delta t(i+1)}\left\|\int_{\mathbb{S}} f(t,s')\rho(s',t|s) - f(\Delta ti,s')\rho(\Delta ti,s')ds'\right\|_\rho^2 dt},
\end{aligned} \tag{54}$$

where Jensen's inequality is used. By (48), 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 $p_2(s,t)$ satisfy

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

By Proposition 23, one has

$$\partial_t\left(\frac{1}{2}\|p_i(t)\|_\rho^2\right) \le -\frac{\lambda_{\min}}{4}\|\nabla p_i(t)\|_\rho^2 + \frac{\beta}{2}\|p_i(t)\|_\rho^2,$$

which implies,

$$\|p_i(t)\|_\rho \le e^{\frac{\beta}{2}t}\|p_i(0)\|_\rho .$$

Therefore, one has

$$\begin{aligned}
&\left\|\int_{\mathbb{S}} f(t,s')\rho(s',t|s) - f(\Delta ti,s')\rho(\Delta ti,s')ds'\right\|_\rho \\
\le& e^{-\beta\Delta ti}(t-\Delta ti)e^{\frac{\beta}{2}\Delta ti}\left(e^{\frac{\beta}{2}(\xi_1-\Delta ti)}\|\mathcal{L}_{\mu,\Sigma}r(s)\|_\rho + \beta\|r(s)\|_\rho\right), \quad \xi_1\in[\Delta ti,\Delta t(i+1)].
\end{aligned}$$

Substituting this into (54) yields

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

where the last equality comes from

$$e^{\frac{\beta}{2}\Delta t}\Delta t^2\sum_{i=0}^{\infty}e^{-\frac{\beta}{2}\Delta ti} = \frac{2\Delta t}{\beta} + o(\Delta t),$$

which completes the proof. ■

6.5 Proof of Theorem 14

First note that V and $\hat{V}_i$ satisfy

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

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 24 *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 $C_\mu+\frac{1}{2}C_\Sigma\le\sqrt{\frac{\beta\lambda_{min}}{8}}, C_\Sigma\le\frac{\lambda_{min}}{4}$, one has

$$\left\|V-\hat{V}\right\|_\rho \le \left[\frac{4C_\mu+2C_{\nabla\cdot\Sigma}}{\beta}\left(1+\frac{2C_\Sigma}{\lambda_{min}}\right)+\frac{2C_\Sigma}{\sqrt{\beta\lambda_{min}}}\right]\|\nabla V\|_\rho + \frac{2C_\Sigma L_\rho}{\beta}\left\|\nabla\hat{V}\right\|_{L^\infty}.$$

(See Section B.3 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 25 *Under Assumption 1, for $\hat{\mu}(s), \hat{\Sigma}(s)$ defined in (18), one has*

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

and

$$\left\|\nabla \cdot (\hat{\Sigma} - \Sigma)\right\|_{L^\infty}^2 \leq L_{\nabla\cdot\Sigma} \Delta t^i + o(\Delta t^i), \tag{55}$$

where $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}$ are constants depending on μ, Σ, i defined in (76), (79), (81), respectively.

(See Section B.4 for the proof of the above lemma.) Combining the above two lemmas yields

$$\left\|V - \hat{V}_i\right\| \leq \left[\left(\frac{3L_\mu + 2L_{\nabla\cdot\Sigma}}{\beta}\left(1 + \frac{L_\Sigma \Delta t^i}{\lambda_{\min}}\right) + \frac{L_\Sigma}{\sqrt{\beta\lambda_{\min}}}\right)\|\nabla V\|_\rho + \frac{2L_\Sigma L_\rho}{\beta}\left\|\nabla \hat{V}\right\|_{L^\infty}\right]\Delta t^i + o(\Delta t^i)$$

for

$$\Delta t^i \leq D_{\mu,\Sigma,\beta}, \quad D_{\mu,\Sigma,\beta} = \min\left\{\frac{\lambda_{\min}}{2L_\Sigma}, \frac{\sqrt{2\beta\lambda_{\min}}}{2L_\mu + L_{\nabla\cdot\Sigma}}\right\}, \tag{56}$$

with $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}$ defined in (76), (79), (94). Furthermore, the following lemma provides upper bounds for $\|\nabla V\|_\rho$ and $\left\|\nabla \hat{V}\right\|_{L^\infty}$, completing the proof of Theorem 14.

Lemma 26 *Under Assumption 1/(a), for $V(s)$ satisfying (16), one has*

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

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

where $C_{\nabla\mu,\nabla\Sigma}$ is a constant defined in (53) that depends on $\nabla\mu(s), \nabla\Sigma(s)$.

(See Section B.5 for the proof of the above lemma.) Therefore, one has

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

where

$$
\begin{aligned}
C_{r,\mu,\Sigma} =& (3L_\mu + 2L_{\nabla\cdot\Sigma}) \left(\sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}} \|r\|_\rho + \|\nabla r\|_\rho \right) \\
&+ 2L_\Sigma L_\rho \left(\sqrt{\frac{2C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}} \|r\|_{L^\infty} + \|\nabla r\|_{L^\infty} \right), \\
\hat{C}_{r,\mu,\Sigma} =& \frac{L_\Sigma}{\sqrt{\lambda_{\min}}} \left(\sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}} \|r\|_\rho + \|\nabla r\|_\rho \right),
\end{aligned} \tag{57}
$$

with $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}, L_\rho, C_{\nabla\mu,\nabla\Sigma}$ defined in (76), (79), (94), (20), (53).

6.6 Proof of Theorem 16 and Galerkin Error with Unbounded $\|\nabla \log \rho\|_{L^\infty}$

The ith-order approximation $\hat{V}_i(s)$ can be decomposed into two parts:

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

where $\hat{V}_i^P(s)$ can be any function 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 decomposed into two parts:

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

Subtracting the above equation from the Galerkin equation (24) 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 $e_i^G(s) = \hat{V}_i^G(s) - \hat{V}_i^P(s) = \sum_{k=1}^p e_k \phi_k(s)$. Then multiplying the above equation by $(e_1, \cdots, e_p)$ yields

$$
\begin{aligned}
\left\langle (\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i}) e_i^G, e_i^G \right\rangle_\rho =& \left\langle (\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i}) e_i^P, e_i^G \right\rangle_\rho \\
\left\langle (\beta - \mathcal{L}_{\mu,\Sigma}) e_i^G, e_i^G \right\rangle_\rho =& \left\langle \mathcal{L}_{\hat{\mu}_i-\mu,\hat{\Sigma}_i-\Sigma} e_i^G, e_i^G \right\rangle_\rho + \left\langle (\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i}) e_i^P, e_i^G \right\rangle_\rho .
\end{aligned} \tag{58}
$$

When $L_\rho^\infty = \|\nabla \log \rho\|_{L^\infty}$ is bounded, by applying the last inequality of Proposition 23 and Lemma 25, one has

$$
\begin{aligned}
\frac{\beta}{2} \left\|e_i^G\right\|_\rho^2 + \frac{\lambda_{\min}}{4} \left\|\nabla e_i^G\right\|_\rho^2 \leq& c_1 \left\|\nabla e_i^G\right\|_\rho^2 + c_2 \left\|e_i^G\right\|_\rho \left\|\nabla e_i^G\right\|_\rho + \beta \left\|e_i^P\right\|_\rho \left\|e_i^G\right\|_\rho \\
&+ c_3 \left\|\nabla e_i^P\right\|_\rho \left\|\nabla e_i^G\right\|_\rho + c_4 \left\|\nabla e_i^P\right\|_\rho \left\|e_i^G\right\|_\rho ,
\end{aligned}
$$

where

$$
\begin{aligned}
c_1 &= \frac{L_\Sigma}{2}\Delta t^i, \quad c_2 = \left(L_\mu + \frac{L_{\nabla\cdot\Sigma}}{2} + \frac{L_\Sigma L_\rho^\infty}{2}\right)\Delta t^i, \\
c_3 &= \frac{1}{2}\left(\|\Sigma\|_{L^\infty} + L_\Sigma \Delta t^i\right), \\
c_4 &= \left(\|\mu\|_{L^\infty} + \frac{\|\nabla\cdot\Sigma\|_{L^\infty}}{2} + \frac{\|\Sigma\|_{L^\infty} L_\rho^\infty}{2}\right) + \left(L_\mu + \frac{L_{\nabla\cdot\Sigma}}{2} + \frac{L_\Sigma L_\rho^\infty}{2}\right)\Delta t^i
\end{aligned} \tag{59}
$$

with $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}$ defined in (76), (79), (94). Under the assumption that $c_1 \le \frac{\lambda_{\min}}{16}, \frac{c_2}{2} \le \min\left\{\frac{\lambda_{\min}}{16}, \frac{\beta}{4}\right\}$, i.e.,

$$
\Delta t^i \le \eta_{\mu,\Sigma,\beta}, \quad \eta_{\mu,\Sigma,\beta} = \frac{\min\left\{\frac{\lambda_{\min}}{4}, \beta\right\}}{2L_\mu + L_{\nabla\cdot\Sigma} + L_\Sigma \max\{L_\rho^\infty, 4\}} \tag{60}
$$

with $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}$ defined in (76), (79), (94), one has

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

which implies that

$$
\left\|\hat{V}_i - \hat{V}_i^G\right\|_\rho \le \left\|e_i^P\right\|_\rho + \left\|e_i^G\right\|_\rho \le 5\left\|e_i^P\right\|_\rho + \sqrt{\frac{8}{\beta}\left(\frac{2c_3^2}{\lambda_{\min}} + \frac{4c_4^2}{\beta}\right)}\left\|\nabla e_i^P\right\|_\rho.
$$

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

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

where

$$
C_G = \max\left\{5\beta, \frac{4c_3\sqrt{\beta}}{\sqrt{\lambda_{\min}}} + 6c_4\right\}, \quad \|f\|_{H_\rho^1} = \|f\|_\rho + \|\nabla f\|_\rho. \tag{61}
$$

with c_3, c_4 defined in (59).

Theorem 27 (Galerkin Error with unbounded $\|\nabla\log\rho\|_{L^\infty}$) *The Galerkin solution $\hat{V}_i^G(s) = \theta^\top\Phi(s)$ satisfies*

$$
\left\langle(\beta - \mathcal{L}_{\hat{\mu}_i,\hat{\Sigma}_i})\hat{V}_i^G(s), \Phi\right\rangle_\rho = \langle r(s), \Phi(s)\rangle_\rho. \tag{62}
$$

When $\|\nabla\log\rho\|_{L^\infty}$ is unbounded, assume that the basis vector $L_\Phi^\infty = \|\Phi\|_{L^\infty}$ is bounded. Then, as long as $\Delta t^i \le \min\{\hat{\eta}_{\mu,\Sigma,\beta}, D_{\mu,\Sigma,\beta}\}$, the Galerkin solution $\hat{V}_i^G(s)$ approximates the solution to the ith-order PhiBE defined in (17) *with an error*

$$
\left\|\hat{V}_i^G(s) - \hat{V}_i(s)\right\|_\rho \le \frac{\hat{C}_G}{\beta}\inf_{V^P=\theta^\top\Phi}\left\|\hat{V}_i - V^P\right\|_{H_{\rho,\infty}^1}.
$$

where $\hat{\eta}_{\mu,\Sigma,\beta}, \hat{C}_G, D_{\mu,\Sigma,\beta}$ *are constants depending on* $\mu, \Sigma, \beta, L_\rho^\infty, L_\Phi^\infty$ *defined in* (65), (66), (56) *respectively, and* $\|f\|_{H^1_{\rho,\infty}} = \|f\|_{H^1_\rho} + \|\nabla f\|_{L^\infty}$.

Proof When $L_\rho^\infty = \|\nabla \log \rho\|_{L^\infty}$ is not bounded, applying the second-to-last inequality of Proposition 23 and Lemma 25 to (58) gives

$$\begin{aligned}\frac{\beta}{2}\left\|e_i^G\right\|_\rho^2 + \frac{\lambda_{\min}}{4}\left\|\nabla e_i^G\right\|_\rho^2 \le c_1\left\|\nabla e_i^G\right\|_\rho^2 + c_5\left\|e_i^G\right\|_\rho\left\|\nabla e_i^G\right\|_\rho + c_6\left\|e_i^G\right\|_{L^\infty}\left\|\nabla e_i^G\right\|_\rho \\ + \beta\left\|e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho + c_3\left\|\nabla e_i^P\right\|_\rho\left\|\nabla e_i^G\right\|_\rho + c_7\left\|\nabla e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho + c_8\left\|\nabla e_i^P\right\|_{L^\infty}\left\|e_i^G\right\|_\rho,\end{aligned} \tag{63}$$

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 (76), (79), (94), (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 linearly independent basis functions with respect to the weighted L^2 norm. Let $L_\Phi^\infty = \|\Phi\|_{L^\infty}$; then (63) can be rewritten as

$$\begin{aligned}\frac{\beta}{2}\left\|e_i^G\right\|_\rho^2 + \frac{\lambda_{\min}}{4}\left\|\nabla e_i^G\right\|_\rho^2 \le c_1\left\|\nabla e_i^G\right\|_\rho^2 + c_9\left\|e_i^G\right\|_\rho\left\|\nabla e_i^G\right\|_\rho + \beta\left\|e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho \\ + c_3\left\|\nabla e_i^P\right\|_\rho\left\|\nabla e_i^G\right\|_\rho + c_7\left\|\nabla e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho + c_8\left\|\nabla e_i^P\right\|_{L^\infty}\left\|e_i^G\right\|_\rho,\end{aligned}$$

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 $c_1 \le \frac{\lambda_{\min}}{16}$, $\frac{c_9}{2} \le \min\{\frac{\lambda_{\min}}{16}, \frac{\beta}{4}\}$, i.e.,

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

with $L_\mu, L_\Sigma, L_{\nabla\cdot\Sigma}$ defined in (76), (79), (94), then one has

$$\begin{aligned}\frac{\beta}{4}\left\|e_i^G\right\|_\rho^2 + \frac{\lambda_{\min}}{8}\left\|\nabla e_i^G\right\|_\rho^2 \le& \beta\left\|e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho + c_3\left\|\nabla e_i^P\right\|_\rho\left\|\nabla e_i^G\right\|_\rho \\ &+ c_7\left\|\nabla e_i^P\right\|_\rho\left\|e_i^G\right\|_\rho + c_8\left\|\nabla e_i^P\right\|_{L^\infty}\left\|e_i^G\right\|_\rho, \\ \left\|e_i^G\right\|_\rho \le 7\left\|e_i^P\right\|_\rho &+ \sqrt{\frac{16c_3^2}{\beta\lambda_{\min}} + \frac{48c_7^2}{\beta^2}}\left\|\nabla e_i^P\right\|_\rho + \frac{7c_8}{\beta}\left\|\nabla e_i^P\right\|_{L^\infty},\end{aligned}$$

which implies that

$$\left\|\hat{V}_i - \hat{V}_i^G\right\|_\rho \leq 8\left\|e_i^P\right\|_\rho + \sqrt{\frac{16c_3^2}{\beta\lambda_{\min}} + \frac{48c_7^2}{\beta^2}}\left\|\nabla e_i^P\right\|_\rho + \frac{7c_8}{\beta}\left\|\nabla e_i^P\right\|_{L^\infty}.$$

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

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

where

$$\hat{C}_G = \max\left\{8\beta, \frac{4c_3\sqrt{\beta}}{\sqrt{\lambda_{\min}}} + 7c_7, 7c_8\right\}, \quad \|f\|_{H^1_{\rho,\infty}} = \|f\|_{H^1_\rho} + \|\nabla f\|_{L^\infty}. \tag{66}$$

with c_3, c_7, c_8 defined in (59), (64). ■

6.7 Proof of Theorem 19

Note that in the proof, we refer to the vector norm $\|b\|$ as the Euclidean norm and the matrix norm $\|A\|$ as the operator norm.

We first prove that the matrix A defined in (27) is positive definite. By applying the first and last inequalities of Proposition 23, one has, for all $\theta \in \mathbb{R}^p$,

$$\begin{aligned}
&\theta^\top A\theta = \left\langle \beta V(s) - \mathcal{L}_{\hat{\mu}_1,\hat{\Sigma}_1}V(s), V(s)\right\rangle_\rho, \quad \text{with } V = \Phi(s)^\top\theta \\
=&\beta\|V(s)\|_\rho^2 - \langle \mathcal{L}_{\mu,\Sigma}V(s), V(s)\rangle_\rho - \left\langle \mathcal{L}_{\hat{\mu}_1-\mu,\hat{\Sigma}_1-\Sigma}V(s), V(s)\right\rangle_\rho \\
\geq&\beta\|V\|_\rho^2 + \frac{\lambda_{\min}}{2}\|\nabla V\|_\rho^2 \\
&- \left[\left(\|\epsilon_\mu\|_\infty + \frac{1}{2}\|\nabla\cdot\epsilon_\Sigma\|_\infty + \frac{1}{2}\|\epsilon_\Sigma\|_\infty\|\nabla\log\rho\|_\infty\right)\|V(s)\|_\rho\|\nabla V\|_\rho + \frac{1}{2}\|\epsilon_\Sigma\|_\infty\|\nabla V\|_\rho^2\right]
\end{aligned}$$

where $\epsilon_\mu = \hat{\mu}_1 - \mu, \epsilon_\Sigma = \hat{\Sigma}_1 - \Sigma$. By Lemma 25, one has

$$\|\epsilon_\mu\|_\infty + \frac{1}{2}\|\nabla\cdot\epsilon_\Sigma\|_\infty + \frac{1}{2}\|\epsilon_\Sigma\|_\infty\|\nabla\log\rho\|_\infty \lesssim (L_\mu + L_\Sigma L_\rho^\infty + L_{\nabla\cdot\Sigma})\Delta t, \quad \|\epsilon_\Sigma\|_\infty \leq L_\Sigma\Delta t.$$

Under the assumption that $\Delta t \leq \eta_{\mu,\Sigma,\beta}$ defined in (60) is small enough, one can further bound

$$\theta^\top A\theta \geq \frac{\beta}{2}\|V\|_\rho^2 = \frac{\beta}{2}\theta^\top G_\Phi\theta \geq \frac{\beta\lambda_\Phi}{2}\|\theta\|^2,$$

where the last inequality comes from Assumption (29) on Φ.

In addition, one also has the following bound on $\|b\|$:

$$\|b\| = \left\|\int r(s)\Phi(s)\rho(s)ds\right\| \leq RL_\Phi, \quad R = \|r\|_{L^\infty}.$$

Next, one can decompose the error $\|\theta_n - \theta\|$ as

$$\begin{aligned}&\|\theta_n - \theta\| = \left\|A_n^{-1}b_n - A^{-1}b\right\| = \left\|A_n^{-1}(b_n - b) - (A_n^{-1} - A^{-1})b\right\| \\ &\leq \left\|A_n^{-1}\right\| \|b_n - b\| + \left\|A_n^{-1}\right\| \|A_n - A\| \left\|A^{-1}\right\| \|b\|.\end{aligned}$$

If one has

$$\|A_n - A\| \lesssim \min\left\{\frac{\beta\lambda_\Phi}{4}, \frac{(\beta\lambda_\Phi)^2\epsilon}{16RL_\Phi^2}\right\}, \quad \|b_n - b\| \leq \frac{\epsilon\beta\lambda_\Phi}{8L_\Phi}$$

then, by the positive definiteness of A and the boundedness of b, one has

$$\|V_n - V\| \leq \|\theta_n - \theta\| L_\Phi \leq \left(\frac{4}{\beta\lambda_\Phi}\|b_n - b\| + \frac{8L_\Phi R}{(\beta\lambda_\Phi)^2}\|A_n - A\|\right) L_\Phi \leq \epsilon$$

Since we assume that ϵ is small enough, the goal becomes determining the sample complexity required to bound

$$\|b_n - b\| \leq \frac{\epsilon\beta\lambda_\Phi}{8L_\Phi}, \quad \|A_n - A\| \lesssim \frac{(\beta\lambda_\Phi)^2\epsilon}{16RL_\Phi^2} \tag{67}$$

Next, we bound the first term. Let $b_i = r(s_i)\Phi(s_i)$ with $s_i \sim \rho$. By the boundedness of $\|r\|_\infty$ and the almost-sure boundedness of $\Phi(s)$, the random vectors b_i are uniformly bounded. Applying Bernstein's inequality for sums of bounded random vectors (Vershynin, 2018), we obtain that for $n \geq \log(1/\delta)$, with probability at least $1 - \delta$,

$$\|b_n - b\| = \left\|\frac{1}{n}\sum_{i=1}^n b_i - E[b_i]\right\| \lesssim L_\Phi R\sqrt{\frac{\log(1/\delta)}{n}}$$

Therefore, the sample complexity to achieve the first error in (67) is

$$n \gtrsim \frac{\log(1/\delta)L_\Phi^2 R^2}{\epsilon^2\beta^2\lambda_\Phi^2}$$

Now we bound the second term. Let $w_i = s_i' - s_i$. Under Assumption 1, the transition density $\rho(\Delta t, w \mid s_i)$ admits two-sided Gaussian bounds of Aronson type (Aronson, 1967). Specifically, there exist positive constants c_1, c_2, c_3, c_4, depending only on the dimension d, the ellipticity constants λ, Λ of $\sigma\sigma^\top$, and $\|b\|_\infty$, such that for all sufficiently small Δt,

$$c_1\,\Delta t^{-d/2}\exp\Big(-\frac{|w|^2}{c_2\,\Delta t}\Big) \;\leq\; \rho(\Delta t, w \mid s_i) \;\leq\; c_3\,\Delta t^{-d/2}\exp\Big(-\frac{|w|^2}{c_4\,\Delta t}\Big).$$

Define the norm $\|\cdot\|_{\psi_1}$ for a scalar random variable X as

$$\|X\|_{\psi_1} = \inf\left\{C > 0 : \mathbb{E}\left[e^{\frac{|X|}{C}}\right] \leq 2\right\},$$

and for a vector or matrix X, define it as

$$\|X\|_{\psi_1} = \sup_{\|u\|_2=1}\left\|u^\top X\right\|_{\psi_1}, \quad \|X\|_{\psi_1} = \sup_{\|u\|_2=\|v\|_2=1}\left\|u^\top Xv\right\|_{\psi_1}.$$

Then one has

$$\|w_i\|_{\psi_1|s_i} \lesssim \sigma_{\max}\sqrt{\Delta t}.$$

It is well known that sub-Gaussian random variables are sub-exponential and that

$$\|X\|_{\psi_1} \lesssim \|X\|_{\psi_2},$$

where the implicit constant is universal (Vershynin, 2018).

Let $M_i(t) := \int_0^t u^\top \sigma(s_\tau) dB_\tau$ with some $\|u\| = 1$, Since the quadratic variation of $M_i(t)$ can be bounded by (Karatzas and Shreve, 2014)

$$\langle M_i \rangle_t = \int_0^t u^\top \sigma(s_\tau)\sigma(s_\tau)^\top u \, d\tau \le \sigma_{\max}^2 t,$$

$\lambda M_i(t)$ is a continuous local martingale with $M_i(0) = 0$ and $\mathbb{E}[\exp(\frac{\lambda^2}{2}\langle M\rangle_t)] < \infty$ for any finite t. By the Doléans–Dade exponential martingale theorem (Karatzas and Shreve, 1998, Chapter 3),

$$Z_t := \exp\left(\lambda M_t - \frac{\lambda^2}{2}\langle M\rangle_t\right)$$

is a martingale, and therefore

$$\mathbb{E}[Z_t \mid s_i(0) = s_i] = Z_0 = 1.$$

It follows that

$$\mathbb{E}[\exp(\lambda M_{\Delta t}) \mid s_i(0) = s_i] \le \exp\left(\frac{\lambda^2}{2}\langle M\rangle_{\Delta t}\right) \le \exp\left(\frac{\lambda^2}{2}\sigma_{\max}^2 \Delta t\right).$$

This implies (Vershynin, 2018)

$$\|M_i(\Delta t)\|_{\psi_2|s_i} \le C\sigma_{\max}\sqrt{\Delta t}, \tag{68}$$

for a universal constant C. In addition, because the upper bound is independent of s_i, the relation between $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ (Vershynin, 2018) implies

$$\|M_i(\Delta t)\|_{\psi_1} \lesssim \sigma_{\max}\sqrt{\Delta t}. \tag{69}$$

Let $Y_i = \int_0^{\Delta t} \mu(s_i(t))dt$. Then $|Y_i| \le \|\mu\|_{L^\infty}\Delta t$. By (69) and $u^\top w_i = M_i(\Delta t) + u^\top Y_i$, one has

$$\|w_i - \mathbb{E}[w_i]\|_{\psi_1} \le 2\|w_i\|_{\psi_1} \lesssim \sigma_{\max}\sqrt{\Delta t} + \|\mu\|_{L^\infty}\Delta t.$$

Therefore,

$$\left\| \frac{w_i}{\Delta t}\cdot\nabla\Phi(s_i) - \mathbb{E}\left[\frac{w_i}{\Delta t}\cdot\nabla\Phi(s_i)\right]\right\|_{\psi_1} \lesssim L_\Phi\left(\frac{\sigma_{\max}}{\sqrt{\Delta t}} + \|\mu\|_{L^\infty}\right).$$

By (68) and the boundedness of $\left\|\nabla^2\Phi(s)\right\| \le L_\Phi$, $M_i^\top\nabla^2\Phi(s_i)M_i$ is sub-exponential; therefore,

$$\begin{aligned}\left\|w_i^\top\nabla^2\Phi(s_i)w_i\right\|_{\psi_1|s_i} &\le \left\|M_i^\top\nabla^2\Phi(s_i)M_i\right\|_{\psi_1|s_i} + 2\left\|M_i^\top\nabla^2\Phi(s_i)Y_i\right\|_{\psi_1|s_i} + \left\|Y_i^\top\nabla^2\Phi(s_i)Y_i\right\|_{\psi_1|s_i}\\ &\lesssim L_\Phi\left(\sigma_{\max}^2\Delta t + \|\mu\|_{L^\infty}\sigma_{\max}\Delta t\sqrt{\Delta t} + \|\mu\|_{L^\infty}^2\Delta t^2\right)\end{aligned}$$

Hence, one has

$$\begin{aligned}&\left\|\frac{1}{\Delta t}w_i^\top\nabla^2\Phi(s_i)w_i-\mathbb{E}\left[\frac{1}{\Delta t}w_i^\top\nabla^2\Phi(s_i)w_i\right]\right\|_{\psi_1}\le 2\left\|\frac{1}{\Delta t}w_i^\top\nabla^2\Phi(s_i)w_i\right\|_{\psi_1}\\ \lesssim& L_\Phi(\sigma_{\max}^2+\sigma_{\max}\|\mu\|_{L^\infty}\sqrt{\Delta t}+\|\mu\|_{L^\infty}^2\Delta t).\end{aligned}$$

Therefore, for $A_i=\Phi(s_i)\left(\beta\Phi(s_i)-\frac{w_i}{\Delta t}\cdot\nabla\Phi(s_i)-\frac{1}{2\Delta t}w_i^\top\nabla^2\Phi(s_i)w_i\right)^\top$

$$\|A_i-\mathbb{E}[A_i]\|_{\psi_1}\lesssim L_\Phi^2\left(\beta+\frac{\sigma_{\max}}{\sqrt{\Delta t}}+\|\mu\|_{L^\infty}+\sigma_{\max}^2+\sigma_{\max}\|\mu\|_{L^\infty}\sqrt{\Delta t}+\|\mu\|_{L^\infty}^2\Delta t\right).$$

Finally, by applying Bernstein's inequality (Tropp, 2012), one obtains, with probability $1-\delta$,

$$\|A_n-A\|\lesssim\|A_i-\mathbb{E}[A_i]\|_{\psi_1}\left(\sqrt{\frac{\log(2p/\delta)}{n}}+\frac{\log(2p/\delta)}{n}\right)$$

In addition, for $n\ge\log(2p/\delta)$ sufficiently large and Δt sufficiently small, one can further simplify the above inequality:

$$\|A_n-A\|\lesssim\frac{L_\Phi^2\sigma_{\max}}{\sqrt{\Delta t}}\sqrt{\frac{\log(2p/\delta)}{n}}$$

Therefore, the sample complexity to achieve the error in (67) is

$$n\gtrsim\frac{\log(2p/\delta)R^2L_\Phi^8\sigma_{\max}^2}{\beta^4\epsilon^2\Delta t\lambda_\Phi^4}.$$

To summarize, with probability $1-\delta$, if

$$n\gtrsim\max\left\{\frac{\log(2p/\delta)R^2L_\Phi^8\sigma_{\max}^2}{\beta^4\epsilon^2\Delta t\lambda_\Phi^4},\frac{\log(1/\delta)L_\Phi^2R^2}{\epsilon^2\beta^2\lambda_\Phi^2}\right\}$$

one has

$$\left\|V_n-\hat{V}_1^G\right\|_\rho\le\epsilon$$

This gives the first inequality.

On the other hand, if

$$n\gtrsim n_1=\frac{16L_\Phi^4\log(2p/\delta)\left(\frac{\sigma_{\max}}{\sqrt{\Delta t}}\right)^2}{\beta^2\lambda_\Phi^2}\approx\frac{L_\Phi^4\log(2p/\delta)\sigma_{\max}^2\lambda_\Phi^{-2}}{\beta^2\Delta t}$$

then w.p. $1-\delta$, one has

$$\|A_n-A\|\le\frac{\beta\lambda_\Phi}{4}.$$

Therefore, for all $n\ge n_1$, with probability $1-\delta$, one has

$$\begin{aligned}\|V_n-V\|&\le\left(\frac{4}{\beta\lambda_\Phi}\|b_n-b\|+\frac{8L_\Phi R}{(\beta\lambda_\Phi)^2}\|A_n-A\|\right)L_\Phi\\ &\lesssim\left(\frac{4}{\beta\lambda_\Phi}L_\Phi R\sqrt{\frac{\log(p/\delta)}{n}}+\frac{8L_\Phi R}{(\beta\lambda_\Phi)^2}L_\Phi^2\sigma_{\max}\sqrt{\frac{\log(2p/\delta)}{n\Delta t}}\right)L_\Phi\\ &\lesssim\left(L_\Phi^4R\sigma_{\max}\lambda_\Phi^{-2}\sqrt{\log(2p/\delta)}\right)\frac{1}{\beta^2\sqrt{n\Delta t}}\end{aligned}$$

7. Conclusion

In this paper, we introduce PhiBE, a PDE-based Bellman equation that integrates discrete-time information into a continuous-time PDE. PhiBE outperforms the classical Bellman equation in approximating the continuous-time PE 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. We further propose higher-order PhiBE, which yields improved approximations of the true value function and achieves the same accuracy with sparser data, thereby improving sample efficiency. In addition, we develop a model-free algorithm for solving PhiBE under linear function approximation and establish theoretical guarantees that sharpen classical RL convergence results, particularly in their dependence on the discretization step size Δt.

This work serves as the first step toward a systematic PDE-based approach to continuous-time reinforcement learning. We expect the PhiBE framework to extend naturally to more general RL settings, including richer function classes and control problems beyond policy evaluation. Another direction for future work is a systematic theoretical and empirical comparison among direct estimation of the continuous-time SDE coefficients, transition-model-based PhiBE, and the model-free PhiBE estimator proposed and analyzed here.

Acknowledgments

The author was supported in part by the U.S. National Science Foundation under Award No. 2529107.

Appendix A. Derivation of the HJB Equation Under Deterministic Dynamics

By the definition of $V(s)$ in (1), one has

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

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

$$\beta V(s_t) = r(s_t) + \frac{d}{dt} s_t \cdot \nabla V(s_t). \tag{70}$$

or equivalently,

$$\beta V(s) = r(s) + \mu(s) \cdot \nabla V(s). \tag{71}$$

Appendix B. Proof of Lemmas

B.1 Proof of Lemma 21

By the Feynman–Kac theorem, $\hat{V}$ can be equivalently written 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,

$$\begin{aligned} \left|V(s) - \hat{V}(s)\right| &= \left|\int_0^\infty e^{-\beta t}(r(s_t) - r(\hat{s}_t))dt\right| = \left|\int_0^\infty e^{-\beta t}\left(\int_{s_t}^{\hat{s}_t} \nabla r(s) ds\right) dt\right| \\ &\le \|\nabla r\|_{L^\infty} \int_0^\infty e^{-\beta t} \left|\hat{s}_t - s_t\right| dt, \end{aligned} \tag{72}$$

where

$$\frac{d}{dt}s_t = \mu(s_t), \quad \frac{d}{dt}\hat{s}_t = \hat{\mu}(\hat{s}_t), \quad s_0 = \hat{s}_0 = s. \tag{73}$$

Subtracting the two equations in (73) and multiplying by $(\hat{s}_t - s_t)^\top$ gives

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

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

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

Inserting the above inequality back to (72) gives

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

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

B.2 Proof of Lemma 22

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,

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

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

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

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

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

B.3 Proof of Lemma 24

Subtracting the second equation from the first and letting $e(s) = V(s) - \hat{V}(s)$ gives

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

Multiplying the above equation by $e(s)\rho(s)$ and integrating it over $s \in \mathbb{S}$ yields

$$\begin{aligned}
&\frac{\beta}{2}\|e\|_\rho^2 = \langle \mathcal{L}_{\mu,\Sigma} e, e\rangle_\rho + \left\langle \mathcal{L}_{\hat{\mu}-\mu,\hat{\Sigma}-\Sigma}\hat{V}, e\right\rangle_\rho \\
\le& -\frac{\lambda_{\min}}{4}\|\nabla e\|_\rho^2 + \left(C_\mu + \frac{1}{2}C_{\nabla\cdot\Sigma}\right)\|e\|_\rho\left\|\nabla\hat{V}\right\|_\rho + \frac{1}{2}C_\Sigma\|\nabla e\|_\rho\left\|\nabla\hat{V}\right\|_\rho \\
&+ \frac{1}{2}C_\Sigma L_\rho\|e\|_\rho\left\|\nabla\hat{V}\right\|_{L^\infty} \\
\le& -\left(\frac{\lambda_{\min}}{4} - c_2\right)\|\nabla e\|_\rho^2 + \left(c_1\|e\|_\rho + c_2\|\nabla V\|_\rho\right)\|\nabla e\|_\rho + \left(c_1\|\nabla V\|_\rho + c_3\left\|\nabla\hat{V}\right\|_{L^\infty}\right)\|e\|_\rho,
\end{aligned} \tag{74}$$

where the first and third equations in Proposition 23 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

$$\begin{aligned}\frac{\beta}{2}\|e\|_\rho^2 \le& -\frac{\lambda_{\min}}{8}\left(\|\nabla e\|_\rho - \frac{4}{\lambda_{\min}}(c_1\|e\|_\rho + c_2\|\nabla V\|_\rho)\right)^2\\ &+\frac{2}{\lambda_{\min}}(c_1\|e\|_\rho + c_2\|\nabla V\|_\rho)^2 + \left(c_1\|\nabla V\|_\rho + c_3\left\|\nabla\hat{V}\right\|_{L^\infty}\right)\|e\|_\rho\\ \le& \frac{2c_1^2}{\lambda_{\min}}\|e\|_\rho^2 + \left[\left(\frac{4c_1c_2}{\lambda_{\min}} + c_1\right)\|\nabla V\|_\rho + c_3\left\|\nabla\hat{V}\right\|_{L^\infty}\right]\|e\|_\rho + \frac{2c_2^2}{\lambda_{\min}}\|\nabla V\|_\rho^2.\end{aligned}$$

Under the assumption that $2c_1^2 \le \frac{1}{8}\beta\lambda_{\min}$, one has

$$\frac{\beta}{4}\|e\|_\rho^2 \le \left[\frac{4}{\beta}\left(\frac{4c_1c_2}{\lambda_{\min}} + c_1\right)^2 + \frac{2c_2^2}{\lambda_{\min}}\right]\|\nabla V\|_\rho^2 + \frac{4c_3^2}{\beta}\left\|\nabla\hat{V}\right\|_{L^\infty}^2 + \frac{\beta}{8}\|e\|_\rho^2,$$

which yields

$$\|e\|_\rho \le \left[\frac{4c_1}{\beta}\left(\frac{4c_2}{\lambda_{\min}} + 1\right) + \frac{3c_2}{\sqrt{\beta\lambda_{\min}}}\right]\|\nabla V\|_\rho + \frac{4c_3}{\beta}\left\|\nabla\hat{V}\right\|_{L^\infty}.$$

B.4 Proof of Lemma 25

The proof of Lemma 25 relies on the following two lemmas, which we prove below.

Lemma 28 *Define the 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 29 *For* $p(s,t) = \mathbb{E}[f(s_t)|s_0 = s]$ *with* s_t *driven by the SDE* (2), *under Assumption 1/(a), one has*

$$\|\nabla p(s,t)\|_{L^\infty} \le \sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{min}}}\|f\|_{L^\infty} + \|\nabla f\|_{L^\infty}, \quad \textit{with } C_{\nabla\mu,\nabla\Sigma} \textit{ defined in } (53).$$

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

$$\|p(s,t)\|_{L^\infty} \le C_1(\|\mu\|_{C^1}, \|\Sigma\|_{C^1}, \|f\|_{C^1})\sqrt{e^t - 1}.$$

$$\|\nabla p(s,t)\|_{L^\infty} \le C_2(\|\mu\|_{C^2}, \|\Sigma\|_{C^2}, \|f\|_{C^2})\sqrt{e^t - 1}.$$

Here, $C_1(\|\mu\|_{C^1}, \|\Sigma\|_{C^1}, \|f\|_{C^1})$ *and* $C_2(\|\mu\|_{C^2}, \|\Sigma\|_{C^2}, \|f\|_{C^2})$ *are defined in* (87) *and* (92).

Now we are ready to prove Lemma 25. By Lemma 28, one has

$$\hat{\mu}_i(s) = \mu(s) + \frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j^{(i)} \left(\int_0^{j\Delta t} \mathbb{E}[\mathcal{L}^i_{\mu,\Sigma}\mu(s_t)|s_0 = s]t^i dt \right), \tag{75}$$

which implies that

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

where

$$L_\mu = \hat{C}_i \left\|\mathcal{L}^i_{\mu,\Sigma}\mu(s)\right\|_{L^\infty} \quad \text{with } \hat{C}_i = \sum_{j=1}^{i}\frac{|a_j^{(i)}|j^{i+1}}{(i+1)!} \text{ defined in (50).} \tag{76}$$

To prove the second inequality in the lemma, we first apply Lemma 28 to obtain

$$\begin{aligned}\hat{\Sigma}_i(s) =& \Sigma(s) \\ &+ \frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j^{(i)} \int_0^{j\Delta t} \mathbb{E}[\mathcal{L}^i_{\mu,\Sigma}\left(\mu(s_t)(s_t - s_0)^\top + (s_t - s_0)\mu^\top(s_t) + \Sigma(s_t)\right)|s_0 = s]t^i dt.\end{aligned} \tag{77}$$

Note that

$$\begin{aligned}h(s_t) :=& \mathcal{L}^i_{\mu,\Sigma}\left(\mu(s_t)(s_t - s_0)^\top + (s_t - s_0)\mu^\top(s_t) + \Sigma(s_t)\right) \\ &- \left[\mathcal{L}^i_{\mu,\Sigma}\mu(s_t)(s_t - s_0)^\top + (s_t - s_0)(\mathcal{L}^i_{\mu,\Sigma}\mu(s_t))^\top\right]\end{aligned} \tag{78}$$

is a function that depends only on the derivatives $\nabla^j\Sigma$ and $\nabla^j\mu$ of order up to $2i$, which can be bounded under Assumption 1/(b). Thus applying the second inequality of Lemma 29 yields

$$\begin{aligned}&\left\|\mathbb{E}[\mathcal{L}^i_{\mu,\Sigma}\left(\mu(s_t)(s_t - s_0)^\top + (s_t - s_0)\mu^\top(s_t) + \Sigma(s_t)\right)|s_0 = s]\right\|_{L^\infty} \\ \leq& \|h(s)\|_{L^\infty} + 2C_1 e^{t/2},\end{aligned}$$

where $C_1 = C_1(\|\mu\|_{C^1}, \|\Sigma\|_{C^1}, \left\|\mathcal{L}^i_{\mu,\Sigma}\mu\right\|_{C^1})$ is defined in (87). Hence, one has

$$\begin{aligned}&\left\|\int_0^{j\Delta t} \mathbb{E}[\mathcal{L}^i_{\mu,\Sigma}\left(\mu(s_t)(s_t - s_0)^\top + (s_t - s_0)\mu^\top(s_t) + \Sigma(s_t)\right)|s_0 = s]t^i dt\right\|_{L^\infty} \\ \leq& \frac{1}{i+1}\|h\|_{L^\infty}(j\Delta t)^{i+1} + 2C_1\int_0^{j\Delta t}(2+t)t^i dt, \quad \text{for } j\Delta t \leq 3 \\ =& \frac{1}{i+1}\left(\|h\|_{L^\infty} + 4C_1\right)(j\Delta t)^{i+1} + 2C_1\frac{1}{i+2}(j\Delta t)^{i+2},\end{aligned}$$

where $e^{t/2} \leq 2 + t$ for $t \leq 3$ is used in the first inequality. Plugging the above inequality back to (77) implies

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

where

$$L_\Sigma = \hat{C}_i(\|h(s)\|_{L^\infty} + 4C_1(\|\mu\|_{C^1}, \|\Sigma\|_{C^1}, \left\|\mathcal{L}^i_{\mu,\Sigma}\mu\right\|_{C^1})) \tag{79}$$

with $\hat{C}_i, h(s), C_1(p,q,g)$ defined in (76), (78), (87).

To prove the third inequality, one first applies $\nabla\cdot$ to (77):

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

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 (78). Therefore, by the first and third inequalities in Lemma 29, and denoting $c_1 = \sqrt{C_{\nabla\mu,\nabla\Sigma}/\lambda_{\min}}\,\|h\|_{L^\infty} + \|\nabla\cdot h\|_{L^\infty}$, $c_2 = C_2(\|\mu\|_{C^2}, \|\Sigma\|_{C^2}, \left\|\mathcal{L}^i_{\mu,\Sigma}\mu\right\|_{C^2})$, one has

$$\begin{aligned}
&\left\|\nabla\cdot\hat{\Sigma}_i(s) - \nabla\cdot\Sigma(s)\right\|_{L^\infty} \le \frac{1}{\Delta t i!}\sum_{j=1}^{i} |a_j^{(i)}| \left(\int_0^{j\Delta t} c_1 t^i + 2c_2 e^{t/2} t^i dt\right) \\
\le &\frac{1}{\Delta t i!}\sum_{j=1}^{i} |a_j^{(i)}| \left(\int_0^{j\Delta t} (c_1 + 4c_2) t^i + 2c_2 t^{i+1} dt\right) = L_{\nabla\cdot\Sigma}\Delta t^i + o(\Delta t^i),
\end{aligned} \tag{80}$$

where

$$\begin{aligned}
L_{\nabla\cdot\Sigma} = &\hat{C}_i \Big[\sqrt{C_{\nabla\mu,\nabla\Sigma}/\lambda_{\min}}\,\|h\|_{L^\infty} + \|\nabla\cdot h\|_{L^\infty} \\
&+ 4C_2(\|\mu\|_{C^2}, \|\Sigma\|_{C^2}, \left\|\mathcal{L}^i_{\mu,\Sigma}\mu\right\|_{C^2})\Big],
\end{aligned} \tag{81}$$

with $\hat{C}_i, h(s), C_{\nabla\mu,\nabla\Sigma}, C_2(p,q,g)$ defined in (76), (78), (53), (92).

Proof of Lemma 28 **Proof** First note that

$$\Pi_{i,\Delta t} f(s) = \frac{1}{\Delta t}\sum_{j=1}^{i} a_j^{(i)} \int_{\mathbb{S}} f(s' - s)\rho(s', j\Delta t|s) ds', \tag{82}$$

where $\rho(s', t|s)$ is defined in (46). 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 (82) 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 be simplified to

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

After 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 29 **Proof** Note that $p(s,t)$ satisfies the following forward Kolmogorov equation (Pavliotis, 2016),

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

Multiplying the above equation by p and letting $q_l = \partial_{s_l} p$, under Assumption 1/(a), one has

$$\partial_t(\frac{1}{2}p^2) = \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}p^2) - \frac{1}{2}q^\top \Sigma q, \quad \partial_t(\frac{1}{2}p^2) \le \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}p^2) - \frac{\lambda_{\min}}{2}\|q\|_2^2. \tag{83}$$

On the other hand, q_l satisfies

$$\partial_t q_l = \mathcal{L}_{\mu,\Sigma} q_l + \mathcal{L}_{\partial_{s_l}\mu,\partial_{s_l}\Sigma} p, \quad \text{with } q(s,0) = \partial_{s_l} f(s). \tag{84}$$

Multiplying the above equation by q_l and then summing over l gives

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

Adding the above inequality to $c\times$(83) with $c = \frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}$ and $C_{\nabla\mu,\nabla\Sigma}$ defined in (53), one has

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

Let $g(s,t) = \frac{c}{2}p^2 + \frac{1}{2}\|q\|_2^2$. 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 $g_1(s,t)$ satisfies $\partial_t g_1 = \mathcal{L}_{\mu,\Sigma} g_1$, 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 the comparison theorem, one has, for all $s \in \mathbb{S}$,

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

which completes the proof for the first inequality.

For the second inequality, let $p(t,s) = \mathbb{E}[f(s_t)(s_t - s_0)|s_0 = s]$. Let $p_1(t,s) = \mathbb{E}[f(s_t)s_t|s_0 = s]$ and $p_2(t,s) = \mathbb{E}[f(s_t)|s_0 = s]$. Then p satisfies the following PDE:

$$\partial_t p = \partial_t(p_1 - p_2 s) = \mathcal{L}_{\mu,\Sigma} p_1 - (\mathcal{L}_{\mu,\Sigma} p_2)s = \mathcal{L}_{\mu,\Sigma} p + \mu p_2 + \Sigma \nabla p_2, \quad \text{with } p(0,s) = 0.$$

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

$$\partial_t(\frac{1}{2}\|p\|_2^2) \le \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|p\|_2^2) - \frac{\lambda_{\min}}{2}\|q\|_2^2 + \frac{1}{2a}\|\mu p_2 + \Sigma \nabla p_2\|_{L^\infty}^2 + \frac{a}{2}\|p\|_2^2, \quad \text{for all } a > 0.$$

Let $c_2 = \|\mu p_2 + \Sigma \nabla p_2\|_{L^\infty}^2$, one has

$$\partial_t(\frac{1}{2}\|p\|_2^2) \le \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|p\|_2^2) - \frac{\lambda_{\min}}{2}\|q\|_2^2 + \frac{c_2}{2a} + \frac{a}{2}\|p\|_2^2, \quad \text{for all } a > 0. \tag{85}$$

Let $g(t,s) = \frac{1}{2}\|p\|_2^2 e^{-at} + \frac{c_2}{2a^2}e^{-at}$. Then

$$\partial_t g \le \mathcal{L}_{\mu,\Sigma} g, \quad \text{with } g(0,s) = \frac{c_2}{2a^2}. \tag{86}$$

Similarly, by the comparison theorem, $\|g(t,s)\|_{L^\infty} \le \|g(0,s)\|_{L^\infty} = \frac{c_2}{2a^2}$, which implies

$$\|p(t,\cdot)\|_{L^\infty} \le \frac{\sqrt{c_2}}{a}\sqrt{e^{at} - 1} = \sqrt{c_2}\sqrt{e^t - 1} \le C_1(\|\mu\|_{C^1}, \|\Sigma\|_{C^1}, \|f\|_{C^1})\sqrt{e^t - 1}$$

with

$$C_1(\|\mu\|_{C^1}, \|\Sigma\|_{C^1}, \|f\|_{C^1}) = \|\mu\|_{L^\infty}\|f\|_{L^\infty} + \|\Sigma\|_{L^\infty}\left(\sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}}\|f\|_{L^\infty} + \|\nabla f\|_{L^\infty}\right) \tag{87}$$

where the first equality is obtained by selecting $a = 1$. One obtains the last inequality by applying the first inequality in this lemma to $\|\nabla p_2\|_{L^\infty}$ and $\|p_2\|_{L^\infty} \le \|f\|_{L^\infty}$.

On the other hand, letting $h(t,s) = \mu p_2 + \Sigma \nabla p_2$, applying ∇ to (86), and multiplying it by $q^\top$ gives

$$\partial_t(\frac{1}{2}\|q\|_2^2) = \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|q\|_2^2) - \frac{1}{2}\sum_l (\nabla q_l)^\top \Sigma(\nabla q_l) + q^\top \nabla\mu \cdot q + \sum_l \frac{1}{2}(\partial_{s_l}\Sigma : \nabla q)q_l + q \cdot \nabla h, \tag{88}$$

$$\begin{aligned}\partial_t(\frac{1}{2}\|q\|_2^2) \le& \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|q\|_2^2) - \frac{\lambda_{\min}}{2}\|\nabla q\|_2^2 + \|\nabla\mu\|_2\|q\|_2^2 + \frac{1}{2}\|\nabla\Sigma\|_2\|\nabla q\|_2\|q\|_2 + \frac{1}{2a}\|\nabla h\|_{L^\infty}^2 \\ &+ \frac{a}{2}\|q\|_2^2 \\ \le& \mathcal{L}_{\mu,\Sigma}(\frac{1}{2}\|q\|_2^2) + \left(\|\nabla\mu\|_{L^\infty} + \frac{\|\nabla\Sigma\|_{L^\infty}^2}{8\lambda_{\min}}\right)\|q\|_2^2 + \frac{1}{2a}\|\nabla h\|_{L^\infty}^2 + \frac{a}{2}\|q\|_2^2.\end{aligned}$$

Adding the above inequality to $c\times$(85) with $c = \frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}$ and $C_{\nabla\mu,\nabla\Sigma}$ defined in (53), one has

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

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

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

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

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

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

Since one needs to bound $\nabla^2 p_2$, we estimate it first. Differentiating (84) with respect to s_k and letting $q_{lk} = \partial_{s_k}\partial_{s_l} p$ gives

$$\partial_t q_{lk} = \mathcal{L}_{\mu,\Sigma} q_{lk} + \mathcal{L}_{\partial_{s_l}\mu,\partial_{s_l}\Sigma} q_k + \mathcal{L}_{\partial_{s_k}\mu,\partial_{s_k}\Sigma} q_l + \mathcal{L}_{\partial_{s_l}\partial_{s_k}\mu,\partial_{s_l}\partial_{s_k}\Sigma} p$$

Multiplying both sides of the above equation by q_{lk} and summing over l, m gives

$$\begin{aligned}
&\partial_t\left(\frac{1}{2}\left\|\nabla^2 p\right\|_2^2\right)\\
\le&\mathcal{L}_{\mu,\Sigma}\left(\frac{1}{2}\left\|\nabla^2 p\right\|_2^2\right) + \underbrace{\left(\|\nabla\mu\|_{L^\infty}^2 + \frac{\|\nabla\Sigma\|_{L^\infty}^2}{2\lambda_{\min}} + \frac{1}{2}\left\|\nabla^2\Sigma\right\|_{L^\infty}^2\right)}_{c_3}\left\|\nabla^2 p\right\|_2^2 + \frac{1}{2}\left\|\nabla^2\mu\right\|_2^2\|q\|_2^2
\end{aligned} \tag{89}$$

The equation (88) gives

$$\partial_t\left(\frac{1}{2}\|\nabla p\|_2^2\right) \le \mathcal{L}_{\mu,\Sigma}\left(\frac{1}{2}\|\nabla p\|_2^2\right) - \frac{\lambda_{\min}}{4}\left\|\nabla^2 p\right\|_2^2 + \lambda_{\min} c\|q\|_2^2. \tag{90}$$

Now combining (83), (90), and (89) leads to

$$\begin{aligned}
&(83)\times\frac{2}{\lambda_{\min}}\left(\frac{1}{2}\left\|\nabla^2\mu\right\|_2^2 + 4cc_3\right) + (90)\times\frac{4c_3}{\lambda_{\min}} + (89)\\
=&\frac{1}{2}\partial_t\left[\frac{2}{\lambda_{\min}}\left(\frac{1}{2}\left\|\nabla^2\mu\right\|_2^2 + 4cc_3\right)\|p\|_2^2 + \frac{4c_3}{\lambda_{\min}}\|\nabla p\|_2^2 + \left\|\nabla^2 p\right\|_2^2\right]\\
\le&\frac{1}{2}\mathcal{L}_{\mu,\Sigma}\left[\frac{2}{\lambda_{\min}}\left(\frac{1}{2}\left\|\nabla^2\mu\right\|_2^2 + 4cc_3\right)\|p\|_2^2 + \frac{4c_3}{\lambda_{\min}}\|\nabla p\|_2^2 + \left\|\nabla^2 p\right\|_2^2\right]
\end{aligned} \tag{91}$$

By the comparison theorem, one has

$$\left\|\nabla^2 p\right\|_2 \le \frac{1}{\sqrt{\lambda_{\min}}}\left(\left\|\nabla^2\mu\right\|_2 + 2\sqrt{cc_3}\right)\|f\|_2^2 + \frac{2\sqrt{c_3}}{\sqrt{\lambda_{\min}}}\|\nabla f\|_2 + \left\|\nabla^2 f\right\|_2$$

By the first inequality of this lemma and the above inequality, one has

$$\sqrt{c}\|\mu\|_{L^\infty} + \|\nabla h\|_{L^\infty} \le C_2(\|\mu\|_{C^2}, \|\Sigma\|_{C^2}, \|f\|_{C^2})$$

where

$$\begin{aligned}
&C_2(\|\mu\|_{C^2}, \|\Sigma\|_{C^2}, \|f\|_{C^2}) \\
=&\sqrt{c}\,\|\mu\|_{L^\infty} + \|\nabla\mu\|_{L^\infty}\,\|f\|_{L^\infty} \\
&+ (\|\mu\|_{L^\infty} + \|\nabla\Sigma\|_{L^\infty})\,(\sqrt{c}\,\|f\|_{L^\infty} + \|\nabla f\|_{L^\infty}) \\
&+ \|\Sigma\|_{L^\infty}\left[\frac{1}{\sqrt{\lambda_{\min}}}\left(\left\|\nabla^2\mu\right\|_2 + 2\sqrt{cc_3}\right)\|f\|_2^2 + \frac{2\sqrt{c_3}}{\sqrt{\lambda_{\min}}}\,\|\nabla f\|_2 + \left\|\nabla^2 f\right\|_2\right]
\end{aligned} \tag{92}$$

which completes the proof. ■

B.5 Proof of Lemma 26

Based on Proposition 23, one has

$$\begin{aligned}
&\frac{\beta}{2}\,\|V\|_\rho^2 - \langle r(s), V(s)\rangle_\rho \le -\frac{\lambda_{\min}}{4}\,\|\nabla V(s)\|_\rho^2; \\
&\beta\,\|\nabla V\|_\rho^2 - \langle \nabla r(s), \nabla V(s)\rangle_\rho \le \frac{C_{\nabla\mu,\nabla\Sigma}}{2}\,\|\nabla V\|_\rho^2.
\end{aligned}$$

Multiplying the first inequality by $\frac{2C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}$ and adding it to the second gives

$$\begin{aligned}
&\beta\left(\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}\,\|V\|^2 + \|\nabla V(s)\|_\rho^2\right) \\
\le&\frac{1}{2\beta}\left(\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}\,\|r\|_\rho^2 + \|\nabla r\|_\rho^2\right) + \frac{\beta}{2}\left(\frac{2C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}\,\|V\|_\rho^2 + \|\nabla V\|_\rho^2\right),
\end{aligned} \tag{93}$$

which implies

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

For the second inequality, let $p(s,t) = \mathbb{E}[r(s_t)|s_0 = s]$. Then, by Lemma 29, one has

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

As in the estimate of $\left\|\nabla\cdot(\Sigma-\hat{\Sigma})\right\|_{L^\infty}$ in Lemma 22, one has

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

where

$$L_{\nabla\Sigma} = \hat{C}_i \left[\sqrt{C_{\nabla\mu,\nabla\Sigma}/\lambda_{\min}} \|h\|_{L^\infty} + \|\nabla h\|_{L^\infty} + 4 \left(\sqrt{C_{\nabla\mu,\nabla\Sigma}/\lambda_{\min}} \|\mu\|_{L^\infty} + \|\nabla\mu\|_{L^\infty} \right) \right], \tag{94}$$

with $\hat{C}_i, h(s), C_{\nabla\mu,\nabla\Sigma}$ defined in (76), (78), (53). Applying ∇ to (75) 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 29, one has

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

with

$$L_{\nabla\mu} = C_i \left(\sqrt{\frac{C_{\nabla\mu,\nabla\Sigma}}{\lambda_{\min}}} \left\| \mathcal{L}^i_{\mu,\Sigma}\mu(s) \right\|_{L^\infty} + \left\| \nabla \mathcal{L}^i_{\mu,\Sigma}\mu(s) \right\|_{L^\infty} \right). \tag{95}$$

Therefore,

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

References

Yacine Aït-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica*, 70(1):223–262, 2002.

Brandon Amos, Ivan Jimenez, Jacob Sacks, Byron Boots, and J Zico Kolter. Differentiable mpc for end-to-end planning and control. *Advances in neural information processing systems*, 31, 2018.

Donald G. Aronson. Bounds for the fundamental solution of a parabolic equation. *Bulletin of the American Mathematical Society*, 73, 1967. doi: 10.1090/S0002-9904-1967-11830-5.

Leemon C Baird. Reinforcement learning in continuous time: Advantage updating. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 4, pages 2448–2453. IEEE, 1994.

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.

Tadej Battelino, Thomas Danne, Richard M Bergenstal, Stephanie A Amiel, Roy Beck, Torben Biester, Emanuele Bosi, Bruce A Buckingham, William T Cefalu, Kelly L Close, et al. Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. *Diabetes care*, 42(8):1593–1603, 2019.

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.

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

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Claudio Cobelli, Chiara Dalla Man, Giovanni Sparacino, Lalo Magni, Giuseppe De Nicolao, and Boris P Kovatchev. Diabetes: models, signals, and control. *IEEE reviews in biomedical engineering*, 2:54–96, 2009.

Fabienne Comte and Nicolas Marie. Nonparametric estimation of the transition density function for diffusion processes. *Stochastic Processes and their Applications*, 188:104667, 2025. doi: 10.1016/j.spa.2025.104667.

Fabienne Comte, Valentine Genon-Catalot, and Yves Rozenholc. Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli*, 13(2):514–543, 2007. doi: 10.3150/07-BEJ5173.

Kris De Asis and Richard S Sutton. An idiosyncrasy of time-discretization in reinforcement learning. *arXiv preprint arXiv:2406.14951*, 2024.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602 (7897):414–419, 2022.

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.

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

Harry Emerson, Matthew Guy, and Ryan McConville. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. *Journal of Biomedical Informatics*, 142:104376, 2023.

Wendell H. Fleming and Raymond W. Rishel. *Deterministic and Stochastic Optimal Control*. Springer-Verlag, 1975.

Danielle Florens-Zmirou. On estimating the diffusion coefficient from discrete observations. *Journal of Applied Probability*, 30(4):790–804, 1993. doi: 10.2307/3214513.

Emmanuel Gobet, Marc Hoffmann, and Markus Reiß. Nonparametric estimation of scalar diffusions based on low frequency data. *The Annals of Statistics*, 32(5):2223–2253, 2004. doi: 10.1214/009053604000000797.

Xiaoqin Guo, Hung Vinh Tran, and Yuming Paul Zhang. Policy iteration for nonconvex viscous hamilton–jacobi equations. *arXiv preprint arXiv:2503.02159*, 2025.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. A general framework for learning mean-field games. *Mathematics of Operations Research*, 48(2):656–686, 2023.

Lars Peter Hansen and José Alexandre Scheinkman. Back to the future: Generating moment implications for continuous-time markov processes. *Econometrica*, 63(4):767–804, 1995. doi: 10.2307/2171800.

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.

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, 2022a.

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, 2022b.

Yanwei Jia and Xun Yu Zhou. q-learning in continuous time. *Journal of Machine Learning Research*, 24(161):1–61, 2023.

Yanwei Jia, Du Ouyang, and Yufei Zhang. Accuracy of discretely sampled stochastic policies in continuous-time reinforcement learning. *arXiv preprint arXiv:2503.09981*, 2025.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pages 2137–2143. PMLR, 2020.

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

Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*. springer, 2014.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 2nd edition, 1998.

Amirmohammad Karimi, Jun Jin, Jun Luo, A Rupam Mahmood, Martin Jagersand, and Samuele Tosatto. Dynamic decision frequency with continuous options. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7545–7552. IEEE, 2023.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3):209–232, 2002. doi: 10.1023/A:1017984413808.

Bekzhan Kerimkulov, David Siska, and Lukasz Szpruch. Exponential convergence and stability of Howard's policy improvement algorithm for controlled diffusions. *SIAM J. Control Optim.*, 58(3):1314–1340, 2020. ISSN 0363-0129. doi: 10.1137/19M1236758. URL `https://doi.org/10.1137/19M1236758`.

Mathieu Kessler. Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics*, 24(2):211–229, 1997. doi: 10.1111/1467-9469.00059.

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.

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

Claire Lacour. Adaptive estimation of the transition density of a markov chain. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 43(5):571–597, 2007. doi: 10.1016/j.anihpb.2006.09.003.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *The Journal of Machine Learning Research*, 13(1):3041–3074, 2012.

Jaeyoung Lee and Richard S Sutton. Policy iterations for reinforcement learning problems in continuous time and space—fundamental theory and methods. *Automatica*, 126:109421, 2021.

Matthias Löffler and Antoine Picard. Spectral thresholding for the estimation of Markov chain transition operators. *Electronic Journal of Statistics*, 15(2):6281–6310, 2021. doi: 10.1214/21-EJS1935.

Davide Maran, Alberto Maria Metelli, and Marcello Restelli. Tight performance guarantees of imitator policies with continuous actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9073–9080, 2023. doi: 10.1609/aaai.v37i8.26089.

Robert C Merton. Optimum consumption and portfolio rules in a continuous-time model. In *Stochastic optimization models in finance*, pages 621–661. Elsevier, 1975.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

John Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889, 2001.

Rémi Munos. Policy gradient in continuous time. *Journal of Machine Learning Research*, 7:771–791, 2006.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

Akihiro Oga and Yuta Koike. Drift estimation for a multi-dimensional diffusion process using deep neural networks. *Stochastic Processes and their Applications*, 170:104240, 2024. doi: 10.1016/j.spa.2023.104240.

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

Seohong Park, Jaekyeom Kim, and Gunhee Kim. Time discretization-invariant safe action repetition for policy gradient methods. *Advances in Neural Information Processing Systems*, 34:267–279, 2021.

Grigorios A Pavliotis. *Stochastic processes and applications*. Springer, 2016.

Asger Roer Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, 22(1):55–71, 1995.

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

Angel Romero, Yunlong Song, and Davide Scaramuzza. Actor-critic model predictive control. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14777–14784. IEEE, 2024.

Mathieu Sart. Estimation of the transition density of a Markov chain. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(3):1028–1068, 2014. doi: 10.1214/13-AIHP551.

Emeline Schmisser. Penalized nonparametric drift estimation for a multidimensional diffusion process. *Statistics*, 47(1):61–84, 2013. doi: 10.1080/02331888.2011.591931.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Antonio Sciarretta, Michael Back, and Lino Guzzella. Optimal control of parallel hybrid electric vehicles. *IEEE Transactions on control systems technology*, 12(3):352–363, 2004.

Bruno Siciliano and Luigi Villani. *Robot force control*. Springer Science & Business Media, 1999.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

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

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

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

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.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th annual international conference on machine learning*, pages 993–1000, 2009.

Csaba Szepesvári. Least squares temporal difference learning and galerkin's method. In *Mini-Workshop: Mathematics of Machine Learning, Oberwolfach Reports, European Mathematical Society, Oberwolfach*, pages 2385–2388, 2011.

Lukasz Szpruch, Tanut Treetanthiploet, and Yufei Zhang. Optimal scheduling of entropy regularizer for continuous-time linear-quadratic reinforcement learning. *SIAM Journal on Control and Optimization*, 62(1):135–166, 2024.

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.

Wenpin Tang, Yuming Paul Zhang, and Xun Yu Zhou. Exploratory hjb equations and their convergence. *SIAM Journal on Control and Optimization*, 60(6):3191–3216, 2022.

Wenpin Tang, Hung Vinh Tran, and Yuming Paul Zhang. Policy iteration for the deterministic control problems—a viscosity approach. *SIAM Journal on Control and Optimization*, 63(1):375–401, 2025.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Kyriakos G Vamvoudakis and Frank L Lewis. Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, 46(5):878–888, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Haoran Wang and Xun Yu Zhou. Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, 30(4):1273–1308, 2020.

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.

Peter Whittle. Risk-sensitive linear/quadratic/gaussian control. *Advances in Applied Probability*, 13(4):764–777, 1981.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International conference on machine learning*, pages 6995–7004. PMLR, 2019.

Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.

Zichen Vincent Zhang, Johannes Kirschner, Junxi Zhang, Francesco Zanini, Alex Ayoub, Masood Dehghan, and Dale Schuurmans. Managing temporal resolution in continuous value estimation: a fundamental trade-off. *Advances in Neural Information Processing Systems*, 36:62519–62548, 2023.

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.

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.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom Brown, Alec Radford, Dario Amodei, and Paul F. Christiano. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.