OPTIMAL-PHIBE: A PDE-BASED MODEL-FREE FRAMEWORK FOR CONTINUOUS-TIME REINFORCEMENT LEARNING

YUHUA ZHU[1], YUMING ZHANG[2], AND HAOYU ZHANG[3]

Abstract. This paper addresses continuous-time reinforcement learning (CTRL) where the system dynamics are governed by a stochastic differential equation but are unknown, and only discrete-time observations are available. Existing approaches face limitations: model-based PDE methods suffer from non-identifiability, while model-free methods based on the optimal Bellman equation (Optimal-BE) are prone to large discretization errors sensitive to both the dynamics and reward structure. To overcome these challenges, we introduce Optimal-PhiBE, a PDE-based formulation that directly incorporates discrete-time transition information, combining the strengths of both PDE and RL frameworks. Optimal-PhiBE avoids explicit dynamics estimation and exhibits lower sensitivity to reward oscillations, with smaller discretization errors when the uncontrolled system evolves slowly. In the linear-quadratic regulator (LQR) setting, we derive sharp error bounds for both Optimal-PhiBE and Optimal-BE, showing that Optimal-PhiBE exactly recovers the optimal policy in the undiscounted case and significantly outperforms Optimal-BE when the problem is weakly discounted or control-dominant. We further extend Optimal-PhiBE to achieve higher-order accuracy and propose a model-free policy iteration algorithm that solves the equation directly from trajectory data. Numerical experiments confirm the accuracy and efficiency of the proposed method.

1. Introdunction

Reinforcement learning (RL) has achieved remarkable success in artificial intelligence, with applications such as AlphaGo [39], strategic gameplay [24], and fine-tuning large language models [51]. However, these achievements are primarily in discrete-time sequential decision-making settings, where the system state changes only after an action is taken. In contrast, in many real-world decision-making problems, the state evolves continuously in time, regardless of whether actions are taken in continuous or discrete time. Examples arise in healthcare [14, 29], robotics [18, 19, 38], autonomous driving [37], and financial markets [23, 26]. Although these systems are fundamentally continuous in nature, the available data are typically collected at discrete time points. For instance, in dynamic treatment regimes, a patient's blood pressure evolves continuously, but measurements are recorded only when tests are taken. One of the key challenges in continuous-time reinforcement learning (CTRL) is addressing the mismatch between the continuous-time dynamics and the discrete-time data.

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 an optimal control problem with known dynamics [15, 47, 25]. Once the dynamics are estimated, various optimal control algorithms can be applied to solve the CTRL problem. 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 most cases, even ill-posed. We will discuss in Section 2.3 that infinitely many continuous-time dynamics 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 approach is to discretize continuous time and reformulate the CTRL problem as a discrete-time RL problem, i.e., a Markov decision process (MDP) [9, 8, 2]. This transformation enables the use of standard RL algorithms directly on discrete-time data within the classical RL framework. This approach has several advantages. First, it relies only on the discrete-time transition dynamics, thereby avoiding the identifiability

[1] (Corresponding author) Department of Statistics and Data Science, University of California, Los Angeles, USA. (yuhuazhu@ucla.edu).
[2] Department of Mathematics & Statistics, Auburn University, USA. (yzhangpaul@auburn.edu).
[3] Department of Mathematics, University of California, San Diego, USA (haz053@ucsd.edu).

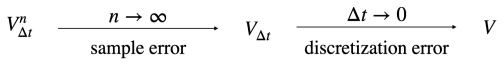


FIGURE 1. Error decomposition for continuous-time RL

issues inherent in the first approach. Second, many RL algorithms are model-free, eliminating the need to explicitly learn the transition dynamics. These plug-and-play algorithms are convenient to implement in practice. However, the discretization error introduced by the MDP framework can be significant and senstitive to all the elements in the system. Additionally, when observation data are sparse in time, the stability of the method may not be guaranteed, as illustrated in [42].

If one decomposes the CTRL error into two components, the first corresponds to the *discretization error*, which arises from the mismatch between the continuous-time dynamics and the discrete-time data. The second is the *finite-sample error*, stemming from the limited number of observations and characterized by sample complexity. We denote the approximation obtained from finite data n as $V^n_{\Delta t}$, see Figure 1. As the number of samples tends to infinity, the algorithm converges to $V_{\Delta t}$, which still differs from the true objective V due to discretization error. In the MDP framework, the discretization error refers to the difference between the optimal feedback policy derived from the Optimal-Bellman equation (Optimal-BE) and the true optimal policy, measured by the value function evaluated under the true dynamics. Since many RL algorithms, such as Q-learning [45], actor-critic [20], TRPO [35], and PPO [36], are derived from the Optimal-BE, their performance is fundamentally limited by this discretization error. In other words, the discretization error associated with the Optimal-BE represents the best approximation that any classical RL algorithm can achieve. As demonstrated in Figure 2, even for the deterministic linear-quadratic regulator (LQR) problem, the discretization error from Optimal-BE is sensitive to all system parameters.

In currently RL literature, most work has focused on analyzing the finite-sample error $V^n_{\Delta t} - V_{\Delta t}$ [28, 46, 17, 6, 49]. In contrast, comparatively little work has focused on analyzing the discretization error. Prior work, such as [11, 31], establishes that as $\Delta t \to 0$, the MDP formulation converges to the original CTRL problem, but without quantifying the rate or the structure of the discretization error. Only recently have some studies begun to investigate the discretization error $V_{\Delta t} - V$. In [3], the authors show that the optimal policy derived from the RL framework achieves only a 1/4-order approximation with weaker assumptions. For the policy evaluation problem, [50, 27] shows the MDP framework provides a first-order approximation with smoothness assumption in the dynamics. Moreover, [50] highlights that when the reward function exhibits large oscillations, the discretization error becomes particularly pronounced. In RL, reward functions often tend to have large oscillations to effectively differentiate between rewards and punishments, which are essential for learning the optimal policy. This characteristic suggests that the MDP framework may not always be ideal for solving CTRL problems, a limitation that has also been observed empirically [8].

In this paper, we focus on CTRL problems where the underlying dynamics are governed by a standard stochastic differential equation (SDE), but only discrete-time transition data are available. We propose a new Optimal-Bellman equation, termed Optimal-PhiBE, that integrates discrete-time information into a continuous-time partial differential equation (PDE). Our approach combines the advantages of both existing frameworks while mitigating their limitations. First, our formulation is a PDE that preserves the continuous-time properties throughout the learning process. At the same time, Optimal-PhiBE relies only on discrete-time transition distributions, which is similar to the Optimal-BE, making it directly compatible with discrete-time data and enabling model-free algorithm design for solving CTRL problems. Moreover, our approach 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, Optimal-PhiBE depends solely on discrete-time transitions and does not require explicit dynamic modeling. Unlike the MDP framework, where discretization errors can be highly sensitive to both the system dynamics and the reward function, Optimal-PhiBE exhibits greater stability with respect to variations in the environment and reward structure. As demonstrated in Figure 2, for the standard LQR problem, given the same discrete-time information, cases where the Optimal-BE suffers from large discretization errors, the first-order Optimal-PhiBE accurately recovers the exact optimal policy.

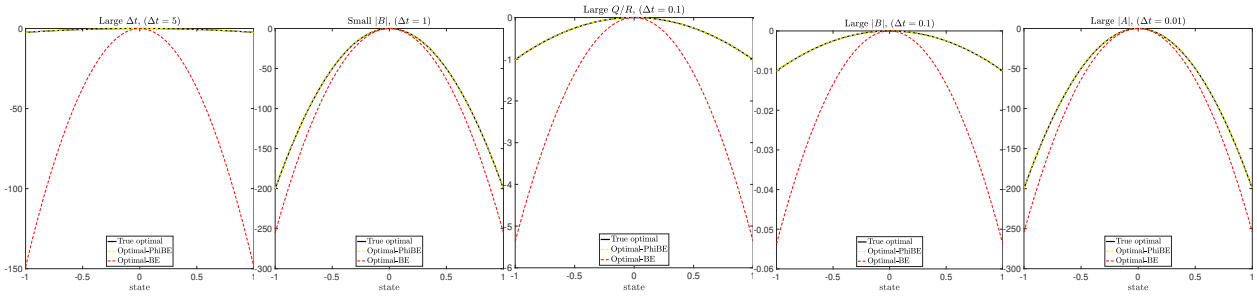


FIGURE 2. The value function under the optimal policy obtained from Optimal-PhiBE and Optimal-BE.

Related work. We review several existing works that aim to reduce discretization errors in continuous-time reinforcement learning (CTRL). In [8], the authors highlight that using the right-Riemann sum provides a more accurate approximation than the left-Riemann sum, particularly due to the effects of discounting. Both [27] and [50] introduce new Bellman equations that achieve smaller discretization errors for continuous-time policy evaluation compared to the standard Bellman equation. The method in [27] primarily focuses on achieving higher-order accuracy, while [50] exploits the structure of stochastic differential equations (SDEs) to obtain better accuracy. This paper builds upon [50] by extending the PhiBE framework from continuous-time policy evaluation to the Optimal-PhiBE formulation for CTRL. However, moving from policy evaluation to optimal control introduces several nontrivial challenges. From a PDE perspective, policy evaluation corresponds to solving a linear elliptic equation, whereas optimal control leads to a nonlinear elliptic equation, making the mathematical extension significantly more complex. From a reinforcement learning perspective, while the Bellman equation can be interpreted as a numerical discretization of a continuous-time integral, the presence of the maximum operator in optimal control complicates the analysis of discretization errors.

Contributions. We summarize the contributions of this paper as follows.

- We propose a PDE-based Optimal Bellman equation, termed Optimal-PhiBE, to approximate both the optimal policy and value function for CTRL, along with its high-order extension for improved discretization accuracy.
- We show that the i-th order Optimal-PhiBE yields an i-th order approximation to the CTRL problem. We characterize how the discretization error depends on the dynamics, reward, and discount coefficient, and show it remains small when the uncontrolled system evolves slowly.
- For the LQR problem, we derive sharp error estimates for Optimal-PhiBE and Optimal-BE in one dimension, and extend the Optimal-PhiBE error to higher dimensios.
- For the undiscounted LQR problem, Optimal-PhiBE exactly recovers the optimal policy using only discrete-time information. In the discounted setting, it outperforms Optimal-BE under any of the following conditions: (1) the problem is weakly discounted, (2) the reward exhibits significant variation in the state space, (3) the reward exhibits small variation in the action space, or (4) the system is control-dominant.
- We propose a model-free algorithm based on Optimal-PhiBE that solves CTRL problems directly from trajectory data, without explicitly estimating the system dynamics.

Organization. The problem setting is introduced in Section 2.1. We provide a detailed discussion of the two existing approaches in Sections 2.3 and 2.4. Our proposed Optimal-PhiBE is presented in Section 3, along with an analysis of its discretization error in terms of optimal value function and optimal policy. In Section 4, we examine the LQR problem and compare the discretization errors from Optimal-PhiBE and Optimal-BE. In Section 5, we introduce a model-free algorithm based on Optimal-PhiBE, and Section 6 conducts numerical experiments validating our approach. The complete proof of the theorems are given in Section 7.

2. The Problem Setting and two existing approaches

2.1. **Classical Optimal Control Setting.** For completeness, we first recall the classical stochastic optimal control setting. Let d be the dimension of the state space and $B = \{B_t\}_{t\geq 0}$ denote a standard Brownian motion in $\mathbb{R}^n$ on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}^B; \{\mathcal{F}_t^B\}_{t\geq 0})$. The drift $b : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^d$ and diffusion $\sigma : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^{d\times n}$ are time-homogeneous, with $\mathcal{A} \subseteq \mathbb{R}^m$ as the action space. Under a *feedback policy* $\pi : \mathbb{R}^d \to \mathcal{A}$, the *state* s_t evolves via the following stochastic differential equation (SDE),

$$ds_t = b(s_t, a_t)\, dt + \sigma(s_t, a_t)\, dB_t, \quad a_t = \pi(s_t). \tag{1}$$

The goal is to find an optimal policy $\pi^* : \mathbb{R}^d \to \mathcal{A}$ that maximizes the expected discounted reward

$$\pi^*(s) = \arg\sup_{\pi} V^\pi(s), \tag{2}$$

where the *value function* $V^\pi(s)$ under a policy π is defined as,

$$V^\pi(s) = \mathbb{E}\left[\int_0^\infty e^{-\beta t} r(s_t, a_t)\, dt \middle| s_0 = s\right] \text{ with } a_t = \pi(s_t). \tag{3}$$

Here, $r : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}$ is the instantaneous reward function, and $\beta > 0$ is the discount coefficient. The optimal value function satisfies,

$$V^*(s) = \max_{\pi} V^\pi(s). \tag{4}$$

We assume the following to ensure the well-posedness of the above stochastic control problem (1)–(4).

Assumption 1. *(i) $\mathcal{A}$ is compact; b, σ and r are continuous in a and are locally uniformly Lipschitz continuous in s.*
(ii) b and σ are uniformly bounded; r has polynomial growth in s, i.e., there exist a constant $C > 0$ and $\mu \geq 1$ such that

$$|r(s, a)| \leq C(1 + |s|^\mu)$$

holds for all $(s, a) \in \mathbb{R}^d \times \mathcal{A}$.

In the classical control setting, where b, σ and r are known, the above stochastic control problem (1)–(4) has been well studied. Under Assumption 1, the optimal value function $V^*(s)$ is the unique viscosity solution of the following Hamilton-Jacobi-Bellman (HJB) equation

$$\beta V^*(s) = \sup_{a\in\mathcal{A}} \left(r(s, a) + (\mathcal{L}_{b,\Sigma} V^*)(s, a)\right), \quad \text{where} \quad \mathcal{L}_{b,\Sigma} = b(s, a) \cdot \nabla_s + \frac{1}{2}\Sigma(s, a) : \nabla_s^2 \tag{5}$$

with $\Sigma(s, a) = \sigma(s, a)\sigma^\top(s, a) \in \mathbb{R}^{d\times d}$ and $\Sigma(s, a) : \nabla_s^2 = \sum_{i,j} \Sigma(s, a)_{ij} \partial_{s_i} \partial_{s_j}$. We refer readers to [48, Theorem 6.1] and the proof of [44, Proposition 4.1]. Here, and throughout the paper, the gradient and Hessian operators ∇ and ∇^2 are in the state space s unless specified. Moreover, the optimal feedback policy π^* is given by:

$$\pi^*(s) = \arg\sup_{a\in\mathcal{A}} (r(s, a) + (\mathcal{L}_{b,\Sigma} V^*)(s, a)).$$

Notably, $r(s, a) + \mathcal{L}_{b,\Sigma} V^\pi$ is the same as $H(s, a, \nabla V^\pi, \nabla^2 V^\pi)$, where the function $H : \mathbb{R}^d \times \mathcal{A} \times \mathbb{R}^d \times \mathbb{R}^{d\times d} \to \mathbb{R}$ is the *Hamiltonian* associated with the original problem (1)-(2) (see e.g. [48]), and is defined as follows.

$$H(s, a, p, q) = b(s, a) \cdot p + \frac{1}{2}\Sigma(s, a) : q + r(s, a).$$

2.2. **Continuous-time Reinforcement Learning (CTRL) Setting.** In this section, we describe the CTRL setting in this paper. In contrast to the classical stochastic control setting discussed previously, we assume that the dynamics of the system are unknown, meaning that the functions $b(s, a)$ and $\sigma(s, a)$ are not accessible to us. Instead, one only has access to the discrete-time trajectory data,

$$\{s^l_{j\Delta t}, a^l_{j\Delta t}, r^l_{j\Delta t}\}_{j=0,l=1}^{j=I,l=L} \tag{6}$$

Here, $s^l_{j\Delta t}$ denotes the state of the l-th trajectory at the time $j\Delta t$, and $a^l_{j\Delta t}$ is the action taken over the time interval $\tau \in [j\Delta t, (j + 1)\Delta t)$ for the l-th trajectory. The reward of the l-th trajectory at time $j\Delta t$ is observed as $r^l_{j\delta t} = r(s^l_{j\Delta t}, a^l_{j\Delta t})$. After the agent interacts with the environment, the next state $s^l_{(j+1)\Delta t}$

of the l-th trajectory is observed at time $(j+1)\Delta t$. We assume that the actions generating the trajectory data are piecewise constant in time, which reflects a more realistic setting. Similarly, the assumption that observations can only be collected at discrete time points also aligns with real-world scenarios. Nevertheless, the goal remains to learn an optimal policy that varies continuously in time, i.e., the optimal policy defined in (2); only the data collection process assumes piecewise-constant actions. The data may originate from a single trajectory or from multiple independent trajectories. The actions may be generated according to a policy or may come from off-policy data. In our setting, we do not assume access to an explicit form of the reward function $r(s,a)$; instead, we rely on the observed reward values associated with sampled state-action pairs.

As shown in Figure 1, we break the problem into two parts. In the first part of the paper, from Section 2.3 to Section 4, we focus on *discretization error*. That is, given the following discrete-time transition distribution,

$$\rho_{\Delta t}(s'|s,a) : \text{the distribution of } s_{(j+1)\Delta t} \text{ given } s_{j\Delta t} = s,\ a_\tau = a, \forall \tau \in [j\Delta t, (j+1)\Delta t)). \tag{7}$$

we investigate how to approximate the solution to the optimal control problem in (1)–(4), or equivalently, the viscosity solution to the HJB equation (5). When the system is governed by a standard SDE, this discrete transition depends on the unknown drift $b(s,a)$, diffusion $\sigma(s,a)$, and the time step Δt. The conditional distribution $\rho_{\Delta t}(s' \mid s,a)$ can also be viewed as the solution to the following PDE,

$$\partial_t \rho_t(s'|s,a) = \nabla_{s'} \cdot \left[-b(s',a)\rho_t(s'|s,a) + \nabla_{s'} \cdot \left[\frac{1}{2}\Sigma(s',a)\rho_t(s'|s,a) \right] \right],$$

at $t = \Delta t$ with initial density $\rho_0(s'|s,a) = \delta_s(s')$ for $\delta_{s_{j\Delta t}}$ being the Dirac measure at s.

In the remainder of this section, we review two existing approaches: the PDE framework and the MDP framework. In Section 2.3, we give an example of the identifiability issue for the PDE framework. In Section 2.4, we state the Optimal-Bellman equation for the MDP framework, and we defer to Section 4 to show why and when it is not the best approximation in this setting.

In Section 3, we introduce a novel approach to addressing the CTRL problem. This alternative, referred to as the *Optimal-PhiBE*, is based on the PDE framework but only requires the discrete transition dynamics in (7). The Optimal-PhiBE provides a solution that closely approximates the optimal value function V^* (Theorem 3.4). Moreover, the policy derived from this solution serves as a good approximation of the optimal policy π^* (Theorem 3.8). We make an explicit comparison to the MDP framework for the LQR problem in Section 4.

In the second part of the paper, Section 5, we present a model-free algorithm that uses discrete data to directly solve the Optimal-PhiBE (7), thereby approximating both the optimal value function and the optimal policy. In Section 6, we evaluate the performance of the proposed algorithm on Linear Quadratic Regulator (LQR) problems and Merton's Portfolio Optimization Problems, demonstrating its effectiveness and practical applicability.

2.3. **Model-based optimal control.** A natural idea for solving equation (5) is to estimate the dynamics, i.e., b and σ, and then solve the PDE using the estimated dynamics. However, the key challenge lies in the mismatch between discrete-time data and continuous-time dynamics. In particular, there is often an *unidentifiability* issue: given the discrete-time transition dynamics, there exist infinitely many continuous-time dynamics that yield the same discrete-time behavior.

Here we give an explicit example. Assume that the underlying true dynamics are deterministic and linear,

$$ds_t = (As_t + Ba_t)dt, \quad A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{8}$$

In addition, assume that we know a priori that the underlying system is linear, and that we are given the discrete-time transition induced by the linear system.

$$p_{\Delta t}(s,a) = e^{A\Delta t}s + A^{-1}(e^{A\Delta t} - I)Ba\Delta t \tag{9}$$

where $p_{\Delta t}(s,a)$ represents the state at time $t+\Delta t$ after taking action a during the time interval $[t, t+\Delta t)$ when the state at time t is s . Note that, given the discrete-time transition dynamics, it is equivalent to say that we are given an infinite set of trajectory data and that there is no model error in estimating the

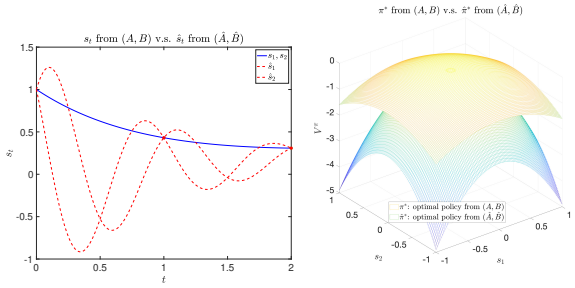


FIGURE 3. Unidentifiability issue for model-based optimal control given discrete-time information. The left plot show the trajectory $s_t, \hat{s}_t$ driven by the true (A, B) and the estimated $(\hat{A}, \hat{B})$. The right figure compares the optimal policy obtained from the estimated dynamics with the true optimal policy, and they are measured in terms of the value function under the true dynamics.

transition dynamics. However, even with such accurate information, one can still find infinitely many pairs of $(\hat{A}, \hat{B})$ that produce the same discrete dynamics $p_{\Delta t}(s, a)$ as given in (9). For example, the following pair is one of them,

$$\hat{A} = \begin{bmatrix} -1 & \frac{2\pi}{\Delta t} \\ -\frac{2\pi}{\Delta t} & -1 \end{bmatrix}, \quad \hat{B} = C^{-1}\hat{A}A^{-1}CB, \quad C = \begin{bmatrix} e^{\Delta t} - 1 & 0 \\ 0 & e^{\Delta t} - 1 \end{bmatrix}.$$

As shown in Figure 3, the continuous-time trajectory driven by the estimated dynamics $(\hat{A}, \hat{B})$ and a constant action $a = [1, 1]^\top$ differs from the true dynamics (A, B), although they coincide at $i\Delta t$. Since the final optimal policy is related to the entire continuous-time dynamics, the optimal policy derived from the incorrect model $(\hat{A}, \hat{B})$ is much worse than the true optimal policy, as shown in the right plot of Figure 3.

2.4. **Optimal-Bellman equation.** Reinforcement learning treats the continuous-time optimal control problem using the MDP framework, which is a framework for discrete-time decision process. Therefore, to apply standard reinforcement learning algorithms, we first need to discretize the original problem in time. Fixing a time discretization scale Δt, and with a slight abuse of notation by omitting the explicit dependence on Δt here, one can then approximate the original continuous-time problem (4) using the following Markov decision process (MDP) framework,

$$\tilde{V}^*(s) = \max_{a_j = \pi(s_j)} \tilde{V}^\pi(s) = \mathbb{E}\left[\sum_{j=0}^{\infty} \gamma^j \tilde{r}(s_j, a_j) \,\middle|\, s_0 = s \right], \tag{10}$$
$$\text{s.t.} \quad s_{j+1} \sim \rho_{\Delta t}(s' | s_j, a_j),$$

where the transition probability $\rho_{\Delta t}(s'|s, a)$ is defined in (7), the corresponding discounted factor is $\gamma = e^{-\beta \Delta t}$, and the discrete-time reward function is $\tilde{r}(s, a) = r(s, a)\Delta t$, where $r(s, a)$ is the original reward function from the continuous-time problem (4). One can view the definition of $\tilde{V}^\pi(s)$ as a numerical integral approximation to the continuous-time value function defined in (3).

The optimal value function defined in (10) can be equivalently written as the solution to the following Optimal-BE [41],

$$\text{Optimal-BE:} \quad \tilde{V}^*(s) = \max_a \left\{ \tilde{r}(s, a) + \gamma \mathbb{E}\left[\tilde{V}^*(s_1) \,\middle|\, s_0 = s, a \right] \right\}. \tag{11}$$

where the condition expectation can be equivalently written as

$$\mathbb{E}\left[\tilde{V}^*(s_1) \,\middle|\, s_0 = s, a \right] = \int \tilde{V}^*(s') \rho_{\Delta t}(s'|s, a) ds'.$$

Then the corresponding optimal policy $\tilde{\pi}^*(s)$ from MDP/Optimal-BE is

$$\tilde{\pi}^*(s) = \arg\max_{\pi(s)} \tilde{V}^\pi(s). \tag{12}$$

Most of the popular model-free RL algorithms (e.g., [2, 21, 35, 20, 45]) are built upon the *Optimal-Bellman equation (Optimal-BE)* mentioned above. A model-free algorithm means that one can obtain the optimal policy $\tilde{\pi}(s)$ without explicitly identifying the dynamics. While this approach is conceptually straightforward and leads to efficient model-free RL algorithms, the discretization error could be large when the reward function or the dynamics change rapidly in state space but slowly in action space. We will characterize the discretization error from the Optimal-BE for the LQR problem in Section 4 and compare it with our proposed framework.

Fundamentally, this issue arises because the MDP framework is designed for discrete-time decision-making processes, and only the discrete-time transition dynamics are used in Optimal-BE (11) or the MDP framework (10). However, our discrete-time transition dynamics are induced by the stochastic differential equation 1, which embeds smoothness information and a specific continuous-time noise. These details, however, are ignored in the standard MDP formulation. The problem is how to embed this continuous-time information into the equation while still only using the discrete-time transition dynamics. We will introduce the Optimal-PhiBE framework for this purpose.

Remark 2.1. *The choice of the discount factor γ and the rescaled reward $\tilde{r}(s,a)$ is not unique. For example, in* [2]*, they approximate $\int_0^\Delta te^{-\beta t}r(s_t)dt \approx r(s_0)\int_0^\Delta te^{-\beta t}dt$; in* [8]*, they use $e^{-\beta\Delta t}r(s_{\Delta t})\Delta t$ instead; in* [9, 27]*, they use a higher-order approximation. However, only adjusting γ and $\tilde{r}$ without using the continuous-time structure of the problem will not solve the large discretization error introduced by the MDP framework.*

We show in Section 4 that there is an optimal choice of γ and $\tilde{r}$ in the LQR problem for the MDP framework, and our error analysis is based on this optimal choice. However, the error is still sensitive to all the elements in the system.

3. Optimal-PhiBE

In this section, as opposed to the Optimal-BE (11), we introduce a new equation called *Optimal-PhiBE*, which uses the intrinsic structure of the problem, particularly the SDE dynamics (1), and combines it with the discrete-time data to approximate the continuous-time optimal value function (4). Given the discrete transition dynamics (7), we demonstrate that the solution to the Optimal-PhiBE provides an better approximation of the optimal value function V^* defined in (4) when the natural dynamics change slowly. Additionally, the policy derived from this solution closely approximates the optimal policy π^* defined in (2).

Optimal-PhiBE builds upon the *PhiBE* (Physics-informed Bellman Equation) framework proposed in [50], which was originally developed as a method for policy evaluation. To ensure a clear understanding, we first review the PhiBE framework before introducing Optimal-PhiBE.

3.1. Review of PhiBE.

In [50], a PDE-based model-free approach is introduced to address the continuous-time policy evaluation problem using discrete-time data. The goal is to approximate the value function under a given policy $\pi(s)$

$$V^\pi(s) = \mathbb{E}\left[\int_0^\infty e^{-\beta t}r^\pi(s_t)\,dt \,\middle|\, s_0 = s\right],$$
$$s.t. \quad ds_t = b^\pi(s_t)\,dt + \sigma^\pi(s_t)\,dB_t,$$

where $r^\pi(s) = r(s,\pi(s)), b^\pi(s) = b(s,\pi(s))$ and $\sigma^\pi(s) = \sigma(s,\pi(s))$ represent the reward, the drift and diffusion under the policy π. The value function $V^\pi(s)$ can be written equivalently as the solution to the following PDE [10],

$$\beta V^\pi(s) = r^\pi(s) + b^\pi(s)\cdot\nabla V^\pi(s) + \frac{1}{2}\Sigma^\pi(s) : \nabla^2 V^\pi(s), \tag{13}$$

where $\Sigma^\pi(s) = \sigma^\pi(s)\sigma(s)^{\pi^\top}$.

Similarly, in the RL setting we consider: the drift and diffusion terms are unknown. We assume that one only has access to the discrete-time transition dynamics $\rho^\pi_{\Delta t}(s'|s)$, which represents the distribution of the state after Δt given that the current state is s and policy $\pi(s_t)$ is applied in $t \in [0,\Delta t)$. Note that [50] assumes the policy is continuously applied, which is different from the setting in this paper, where we assume that the discrete-time trajectory data are obtained by piecewise constant action. [50] proposes a new PDE-based Bellman equation that simultaneously utilizes both the continuous-time PDE structure of the problem and the discrete-time information. The new equation replaces the continuous-time drift and

diffusion terms in (13) with an approximation using the discrete-time information. Specifically, the i-th order approximations for $b^\pi(s)$ and $\Sigma^\pi(s)$ are defined as:

$$\hat{b}_i^\pi(s) = \mathbb{E}\left[\frac{1}{\Delta t}\sum_{j=1}^{i} a_j^i (s_{j\Delta t} - s_0) \,\middle|\, s_0 = s\right], \quad \hat{\Sigma}_i^\pi(s) = \mathbb{E}\left[\frac{1}{\Delta t}\sum_{j=1}^{i} a_j^i (s_{j\Delta t} - s_0)(s_{j\Delta t} - s_0)^\top \,\middle|\, s_0 = s\right], \tag{14}$$

where the conditional expectation is taken over the discrete-time transition dynamics $\rho_{\Delta t}^\pi(s'|s)$. The coefficients a_j^i are obtained by the Taylor expansion, and can be determined by solving,

$$(a_1^i, \cdots, a_i^i)^\top = (A^{(i)})^{-1} b^{(i)}, \quad \text{with} \quad A_{kj}^{(i)} = j^k, \quad b_k^{(i)} = \begin{cases} 1, & k = 1, \\ 0, & k \neq 1, \end{cases} \quad \text{for } 1 \leq j, k \leq i. \tag{15}$$

Notably, the approximations $\hat{b}_i^\pi(s)$ and $\hat{\Sigma}_i^\pi(s)$ only rely on the discrete transition dynamics. Using these approximations, [50] defines the i-th order PhiBE as follows.

Definition 3.1 (PhiBE, [50])**.** *The i-th order PhiBE is defined as:*

$$\beta \hat{V}_i^\pi(s) = r^\pi(s) + \hat{b}_i^\pi(s) \cdot \nabla \hat{V}_i^\pi(s) + \frac{1}{2}\hat{\Sigma}_i^\pi(s) : \nabla^2 \hat{V}_i^\pi(s), \tag{16}$$

where $\hat{b}_i^\pi(s)$ is given by (14), and $\hat{\Sigma}_i^\pi(s)$ is either zero (if $\Sigma^\pi = 0$) or given by (14) (if $\Sigma^\pi \neq 0$).

It is proven in [50] that the solutions $\hat{V}_i^\pi(s)$ to PhiBE provide more accurate approximations of the solution $V^\pi(s)$ to (13) than the MDP framework, particularly when the underlying dynamics change slowly and the reward function changes quickly.

3.2. **Optimal-PhiBE.** In this section, we introduce *Optimal-PhiBE*, an extension of the PhiBE framework. While PhiBE focuses on evaluating value functions for a given policy π, Optimal-PhiBE uses discrete-time transition dynamics (7) to approximate the optimal value function $V^*(s)$, as defined in (4), which is also the solution to the HJB equation (5). In addition to the value function, Optimal-PhiBE also approximates the optimal policy π^*, as defined in (2).

Given the discrete-time transition dynamics (7), one can derive approximations similar to those in (14). For instance, the first-order approximation of $b(s, a)$ is given by,

$$\hat{b}_1(s,a) = \mathbb{E}\left[\frac{1}{\Delta t}(s_{\Delta t} - s_0) \,\middle|\, s_0 = s, a_\tau = a \text{ for } \tau \in [0, \Delta t)\right] = \frac{1}{\Delta t}\int (s' - s)\rho_{\Delta t}(s'|s,a)ds',$$

and the first-order approximation of $\Sigma(s, a)$ is,

$$\hat{\Sigma}_1(s,a) = \mathbb{E}\left[\frac{1}{\Delta t}(s_{\Delta t} - s_0)(s_{\Delta t} - s_0)^\top \,\middle|\, s_0 = s, a_\tau = a \text{ for } \tau \in [0, \Delta t)\right] = \frac{1}{\Delta t}\int (s'-s)(s'-s)^\top \rho_{\Delta t}(s'|s,a)ds'.$$

Note that these approximations rely solely on discrete-time transition dynamics as described in (7). Based on the above approximations, we can approximate the HJB equation (5) as,

$$\beta \hat{V}_1^*(s) = \sup_{a \in \mathcal{A}} \left\{ r(s,a) + \hat{b}_1(s,a) \cdot \nabla \hat{V}_1^*(s) + \frac{1}{2}\hat{\Sigma}_1(s,a) : \nabla^2 \hat{V}_1^*(s) \right\}. \tag{17}$$

We will demonstrate in this section that the solution to this equation provides a good approximation to the optimal value function V^*.

Furthermore, this approach can be extended to higher-order approximations of $b(s, a)$ and $\Sigma(s, a)$. By substituting these higher-order approximations into the HJB equation, we obtain more accurate approximations to the optimal value function when Δt is small. This leads to the formulation of i-th order Optimal-PhiBE, which provides precise approximations of the solutions to the HJB equation. The formal definition is as follows.

Definition 3.2 (Optimal-PhiBE)**.** *The i-th order Optimal-PhiBE is defined as follows:*

$$\beta \hat{V}_i^*(s) = \sup_{a \in \mathcal{A}} \left\{ r(s,a) + \hat{b}_i(s,a) \cdot \nabla \hat{V}_i^*(s) + \frac{1}{2}\hat{\Sigma}_i(s,a) : \nabla^2 \hat{V}_i^*(s) \right\}, \tag{18}$$

where

$$\hat{b}_i(s,a) = \mathbb{E}\left[\frac{1}{\Delta t}\sum_{j=1}^{i} a_j^{(i)}(s_{j\Delta t} - s_0)\middle| s_0 = s, a_\tau = a \text{ for } \tau \in [0, i\Delta t)\right], \tag{19}$$

and

$$\hat{\Sigma}_i(s,a) = \begin{cases} 0, & \text{if } \Sigma \equiv 0, \\ \mathbb{E}\left[\frac{1}{\Delta t}\sum_{j=1}^{i} a_j^{(i)}(s_{j\Delta t} - s_0)(s_{j\Delta t} - s_0)^\top \middle| s_0 = s, a_\tau = a \text{ for } \tau \in [0, i\Delta t)\right], & \text{if } \Sigma \neq 0. \end{cases} \tag{20}$$

Here the coefficient $a^{(i)}$ are given by (15)*.*

When $i = 1$, the Optimal-PhiBE reduces to the first-order case, which we refer (17) *as the standard Optimal-PhiBE.*

Remark 3.3. *An alternative equivalent definition of the coefficients $a^{(i)}$ is given by the following system of equations:*

$$\sum_{j=1}^{i} a_j^{(i)} j^k = \begin{cases} 0, & k \neq 1, \\ 1, & k = 1, \end{cases} \quad \text{for } 1 \leq j, k \leq i. \tag{21}$$

If one knows that the underlying dynamics are deterministic, it is best to set $\hat{\Sigma} = 0$ based on this prior knowledge. However, if the nature of the dynamics, whether deterministic or stochastic, is unclear, one can still use $\hat{\Sigma}$ as defined in the second case of (20). The i-th order approximation remains valid in this case, as proven in Theorem 3.4. In some specific cases, such as LQR, using the deterministic version of the Optimal-PhiBE might outperform the stochastic version, even if the dynamics are stochastic. We will elaborate on this in Section 4.

The definition of Optimal-PhiBE is similar to that of PhiBE, with one key difference: instead of applying the policy $\pi(s)$ continuously over the interval $t \in [t, t + i\Delta t]$, the discrete-time transition dynamics used in Optimal-PhiBE are induced by a constant action a during this time interval. This assumption is more practical, as continuously adjusting actions is often challenging in real-world applications. Nevertheless, the goal remains to learn an optimal policy that varies continuously in time; only the data collection process assumes piecewise-constant actions.

Furthermore, if we expand the Optimal-BE to second-order terms around s, we obtain,

$$\hat{\beta}\tilde{V}^*(s) = \max_a \{r(s,a) + \frac{\gamma}{\Delta t}\mathbb{E}[(s_1 - s_0)\cdot\nabla\tilde{V}^*(s) + (s_1 - s_0)^\top\nabla^2\tilde{V}^*(s)(s_1 - s_0)|s_0 = s, a]\}$$

where $\hat{\beta} = \frac{1-\gamma}{\Delta t}$, and $\gamma = e^{-\beta\Delta t}$. This equation is very similar to Optimal-PhiBE (17), with the differences in the coefficients $|\hat{\beta} - \beta| \sim O(\beta^2\Delta t)$ and $|\gamma - 1| \sim O(\beta\Delta t)$ being small when β or Δt are small. Although the two formulations are similar, Optimal-PhiBE should not be viewed as an approximation to the Optimal-BE for the CTRL problem, since Optimal-PhiBE is directly derived from the HJB equation (7.3). In contrast, for discrete-time RL problems, Optimal-PhiBE can be viewed as an approximation to the Optimal-BE, with the potential advantage of reduced sample complexity.

3.3. **Error Analysis of Optimal-PhiBE.** In Definition 3.2, Optimal-PhiBE (18) was introduced as an approximation to the original HJB equation (5). In this section, we analyze the errors introduced by the Optimal-PhiBE framework from two perspectives: the approximation of the optimal value function and the approximation of the optimal feedback policy.

The solution of the Optimal-PhiBE (18), denoted by $\hat{V}_i^*$, provides an approximation to the optimal value function V^*, the solution to the HJB equation (5). To evaluate the accuracy of the Optimal-PhiBE framework, we quantify the error between $\hat{V}^*$ and V^*, as described in Theorem 3.4.

In addition to the value function, the Optimal-PhiBE framework also produces an approximate optimal feedback policy, $\hat{\pi}_i^*$, which is defined as:

$$\hat{\pi}_i^*(s) = \text{argmax}_{a\in\mathcal{A}}\left[r(s,a) + \hat{b}_i(s,a)\cdot\nabla_s\hat{V}_i^*(s) + \frac{1}{2}\hat{\Sigma}_i(s,a) : \nabla_s^2\hat{V}_i^*(s)\right], \tag{22}$$

where $\hat{V}_i^*$ is the solution to (18), $\hat{b}_i$ and $\hat{\Sigma}_i$ are defined in Definition 3.2. For the original problem (4), the true optimal feedback policy is denoted by π^* and defined in (2). Since the optimal control π^* is not necessarily unique, directly comparing π^* and $\hat{\pi}^*$ may not be feasible. Instead, we assess the difference by comparing their respective induced value functions under the true dynamics (1). Specifically, we consider $V^{\pi^*}(s)$ and $V^{\hat{\pi}^*}(s)$, where the value function under a policy π is defined as,

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

The above s_t satisfies (1) with $a_t = \pi(s_t)$.

The main assumption for the i-th order Optimal-PhiBE to serve as a good i-th order approximation of the original HJB (5) is as follows.

Assumption 2. a) *$r, b, \sigma, \nabla_s r, \nabla_s b, \nabla_s \sigma$ are uniformly bounded.*

b) *$\mathcal{L}^i_{b,\Sigma} b, \mathcal{L}^i_{b,\Sigma}\Sigma, \nabla_s(\mathcal{L}^i_{b,\Sigma} b), \nabla_s(\mathcal{L}^i_{b,\Sigma}\Sigma)$ are uniformly bounded, where $\mathcal{L}_{b,\Sigma}$ is defined in* (5)*. (We also write $\mathcal{L}_{b,\Sigma}$ as $\mathcal{L}_b$ when $\Sigma \equiv 0$ in the paper.)*

c) *Let $h_i(s) = \mathcal{L}^i_{b,\Sigma}(bs^\top) - (\mathcal{L}^i_{b,\Sigma} b)s^\top$, $\|h_i(s)\|_\infty, \|\nabla_s h_i(s)\|_\infty$ are uniformly bounded.*

Several remarks regarding the assumption are in order. Assumption (a) is used to prove the Lipschitz continuity of the solution. The boundedness conditions on $\mathcal{L}^i_{b,\Sigma} b$ and $\mathcal{L}^i_{b,\Sigma}\Sigma$ in Assumption (b) are crucial for estimating the distance between the true dynamics, $b(s,a)$ and $\Sigma(s,a)$, and the corresponding PhiBE dynamics, $\hat{b}(s,a)$ and $\hat{\Sigma}(s,a)$. Additionally, the terms $\nabla_s(\mathcal{L}^i_{b,\Sigma} b)$ and $\nabla_s(\mathcal{L}^i_{b,\Sigma}\Sigma)$ are used to prove the boundedness of $\nabla_s \hat{b}(s,a)$ and $\nabla_s \hat{\Sigma}(s,a)$. Assumption (c) is a technical assumption introduced to bound the distance between Σ and $\hat{\Sigma}$. A sufficient condition for both Assumptions (b) and (c) to hold is the uniform boundedness of $\nabla_s^j b$ and $\nabla_s^j \Sigma$ for $0 \le j \le 2i+1$.

Compared to the original PhiBE work in [50], which focused on the non-degenerate case of Σ, this work extends the i-th order approximation results to handle degenerate Σ. Additionally, while the original work provided an error estimate in the weighted L^2 norm, we offer an estimate in the stronger L^∞ norm with fewer assumptions.

Now we are ready to carry out the main results from the two perspectives mentioned above.

3.3.1. *Perspective 1: Error analysis in terms of optimal value function.* Theorem 3.4 provides an estimate of the distance between the true optimal value function, V^*, and the solution to the Optimal-PhiBE, $\hat{V}_i^*$.

Theorem 3.4. *Under Assumption 2, and β is large enough such that L_β is positive, one has*

$$\left\|\hat{V}_i^* - V^*\right\|_\infty \le \frac{C_i \|\nabla_s r\|_\infty}{\beta L_\beta}\left[\left\|\mathcal{L}^i_{b,\Sigma} b\right\|_\infty + 6\sqrt{dL_\beta + d\|\nabla_x b\|_\infty + d^2\|\nabla_x \sigma\|^2_\infty}\left(\left\|\mathcal{L}^i_{b,\Sigma}\Sigma\right\|_\infty + \|h_i\|_\infty + 3\|b\|_\infty\right)\right]\Delta t^i$$

where $\mathcal{L}_{b,\Sigma}, h_i$ are defined in (5) *and Assumption 2/(c).*

$$L_\beta = \begin{cases} \beta - \|\nabla_s b\|_\infty, & \text{for constant } \sigma(s,a) \\ \beta - \ln 2/\tau, \text{ with } \tau \text{ defined in } (27), & \text{for non-constant } \sigma(s,a). \end{cases} \tag{23}$$

Specially, when $\Sigma \equiv 0$ (deterministic dynamics) with Assumption 2/(a), (b), and $\|\nabla_s b\|_\infty < \beta$, one has

$$\left\|\hat{V}_i^* - V^*\right\|_\infty \le \frac{C_i \|\nabla_s r\|_\infty \left\|\mathcal{L}^i_b b\right\|_\infty}{\beta(\beta - \|\nabla_s b\|_\infty)}\Delta t^i$$

where $\mathcal{L}_b$ is defined in (5)*, $C_i = 4\hat{C}_i$ is a constant only depending on the order i with $\hat{C}_i$ defined in* (53)*.*

Several remarks regarding Theorem 3.4 are in order. The error estimate in the above theorem is similar to that of the policy evaluation problem in [50] in two key aspects. First, although the error still depends on the oscillation of the reward function in the state space, $\|\nabla_s r\|_\infty$, the impact of reward variation can be mitigated if the dynamics change slowly in the state space. This effect is particularly evident in the case of deterministic dynamics. Specifically, in the deterministic case, $\mathcal{L}^i_b b = \frac{d^{i+1}}{dt^{i+1}} s_t$, which measures the rate of change of the dynamics over time. Second, for the i-th order Optimal-PhiBE, the error is $O(\Delta t^i)$.

On the other hand, the proof technique used to prove the above theorem differs from the one employed in [50]. This new framework introduces several improvements. First, it does not require smoothness in the action space or the optimal policy, making the theorem applicable to more general settings, including both continuous and discrete action spaces. Second, the theorem extends to degenerate diffusion terms, unifying deterministic and stochastic dynamics. As a result, there is no need to differentiate between deterministic and stochastic dynamics, as the stochastic Optimal-PhiBE framework can be used for all types of environments.

Proof. The proof of the above theorem is based on the following three Lemmas, which are proved in Section 7.1, 7.2 and 7.3, respectively. We will use the following notations: for $F : \mathbb{R}^d \times A \to \mathbb{R}^{n\times m}$ with $n, m \geq 1$,

$$\|F\|_\infty := \sup_{(s,a)} \|F(s,a)\|_2 \quad \text{and} \quad |F(s,a)| := \|F(s,a)\|_2. \tag{24}$$

where $\|\cdot\|_2$ is the spectral norm or 2-norm.

Lemma 3.5. *Under Assumption 2, one has $\hat{b}, \nabla_s \hat{b}$ are uniformly bounded, and*

$$\left\|\hat{b}_i - b\right\|_\infty \leq \hat{C}_i \left\|\mathcal{L}^i_{b,\Sigma} b\right\|_\infty \Delta t^i. \tag{25}$$

For the stochastic dynamics, when $i\Delta t \leq 3$, one has $\hat{\Sigma}, \nabla_s \hat{\Sigma}$ are uniformly bounded, and

$$\left\|\hat{\Sigma}_i - \Sigma\right\|_\infty \leq \hat{C}_i \left(\left\|\mathcal{L}^i_{b,\Sigma}\Sigma\right\|_\infty + \|h_i\|_\infty + 3\|b\|_\infty\right) \Delta t^i, \tag{26}$$

where $h_i(s)$ is a function defined in (55) that only depends on the derivative $\nabla^j \Sigma, \nabla^j \mu$ with $0 \leq j \leq 2i$.

Lemma 3.6. *Under Assumption 2/(a), and*

$$\rho > \tau^{-1} \ln 2 \quad \text{for some } \tau \text{ such that } \tau\|\nabla_s b\|_\infty + \tau^{\frac{1}{2}} C_d \|\nabla_s \sigma\|_\infty \leq 1/2 \tag{27}$$

where C_d is a dimensional constant, then

$$\|V^*\|_\infty \leq \|r\|_\infty/\rho, \quad \|\nabla V^*\|_\infty \leq \frac{2\|\nabla_s r\|_\infty}{\rho - (\ln 2)/\tau}.$$

In the case when $\|\nabla_s \sigma\|_\infty = 0$, it suffices to assume $\rho > \|\nabla_s b\|_\infty$ instead of (27) and then one has

$$\|\nabla V^*\|_\infty \leq \frac{\|\nabla_s r\|_\infty}{\rho - \|\nabla_s b\|_\infty}.$$

In the following lemma, we bound the difference between the solutions of two HJB equations in terms of the differences in their coefficients. Specifically, let $\hat{V}$ denote the solution to (5) with the coefficients r, b, and Σ replaced by $\hat{r}$, $\hat{b}$, and $\hat{\Sigma}$, respectively.

Lemma 3.7. *Let Assumption 2/(a) hold, and assume $\hat{r}, \hat{b}, \hat{\Sigma}, \nabla_s \hat{r}, \nabla_s \hat{b}, \nabla_s \hat{\Sigma}$ are uniformly bounded. If there exist $\varepsilon_r, \varepsilon_b, \varepsilon_\Sigma \geq 0$, such that,*

$$\sup_{s,a} |r(s,a) - \hat{r}(s,a)| \leq \varepsilon_r, \quad \sup_{s,a} |b(s,a) - \hat{b}(s,a)| \leq \varepsilon_b \quad \text{and} \quad \sup_{s,a} |\sigma(s,a) - \hat{\sigma}(s,a)| \leq \varepsilon_\Sigma,$$

then

$$\sup_s |V^*(s) - \hat{V}(s)| \leq \frac{2}{\rho}(\tilde{C}\varepsilon_\Sigma + 2L\varepsilon_b + \varepsilon_r), \tag{28}$$

where

$$\tilde{C} := 2\sqrt{3d\,(L\|\nabla_s r\|_\infty + 2L^2\|\nabla_s b\|_\infty + 3dL^2\|\nabla_s \sigma\|^2_\infty)}.$$

and L is the Lipschitz constant of $V(s)$.

By Lemma 3.5, one can easily verify the drift and diffusion term of $\hat{V}^*_i$ satisfy the assumption of Lemma 3.7 with $\varepsilon_r = 0, \varepsilon_b = \hat{C}_i \left\|\mathcal{L}^i_{b,\Sigma} b\right\|_\infty \Delta t^i$ and $\varepsilon_\Sigma = \hat{C}_i \left(\left\|\mathcal{L}^i_{b,\Sigma}\Sigma\right\|_\infty + \|h_i\|_\infty + 4\|b\|_\infty\right) \Delta t^i$. From Lemma 3.6, one has $L = \frac{2\|\nabla_s r\|_\infty}{\rho - \ln 2/\tau}$ for non-constant σ, or $\frac{\|\nabla_s r\|_\infty}{\rho - \|\nabla_s b\|_\infty}$ for constant σ. Inserting those quantities into (28) completes the proof of the theorem. □

3.3.2. *Perspective 2: Approximation of the optimal policy.* The following theorem quantifies the difference between the optimal policy π^* and the optimal policy $\hat{\pi}_i^*$ from Optimal-PhiBE, $\hat{\pi}_i^*$, where π^* is defined in (2) and $\hat{\pi}_i^*$ is defined in (22).

The difference between the two feedback policies is measured by the value function under the true dynamics. That is the true optimal value function $V^*(s)$ and the value function $V^{\hat{\pi}_i^*}$ under $\hat{\pi}_i^*$, which is also the solution to the following PDE,

$$\beta V(s) = r(s, \hat{\pi}_i^*(s)) + b(s, \hat{\pi}_i^*(s)) \cdot \nabla V(s) + \frac{1}{2}\Sigma(s, \hat{\pi}_i^*(s)) : \nabla^2 V(s). \tag{29}$$

Theorem 3.8. *Let Assumption 2 hold, and further assume that the optimal policy $\hat{\pi}^*(s)$ is measurable, then*

$$\begin{aligned}\sup_{s\in\mathbb{R}^d}\left\|V^*(s) - V^{\hat{\pi}_i^*}(s)\right\|_\infty \leq & \frac{2C_i\left\|\nabla_s r\right\|_\infty}{\beta L_\beta}\left[\left\|\mathcal{L}_{b,\Sigma}^i b\right\|_\infty\right. \\ & \left. +6\sqrt{dL_\beta + d\|\nabla_s b\|_\infty + d^2\|\nabla_s\sigma\|_\infty^2}\left(\left\|\mathcal{L}_{b,\Sigma}^i\Sigma\right\|_\infty + \left\|h_i\right\|_\infty + 3\left\|b\right\|_\infty\right)\right]\Delta t^i\end{aligned}$$

Specially, when $\Sigma \equiv 0$ (deterministic dynamics) with Assumption 2/(a), (b), and $\left\|\nabla_s b\right\|_\infty < \beta$, one has

$$\left\|\hat{V}_i^* - V^*\right\|_\infty \leq \frac{2C_i\left\|\nabla_s r\right\|_\infty\left\|\mathcal{L}_b^i b\right\|_\infty}{\beta(\beta - \left\|\nabla_s b\right\|_\infty)}\Delta t^i$$

with $C_i, L_\beta, h_i, \mathcal{L}_b$ being the same as Theorem 3.4.

The proof of the above theorem is provided in Section 7.4. The error bound for the optimal policy is similar to that of the value function. We believe this bound can be sharpened in specific cases.

In the following section, we derive a sharper error estimate specific to the LQR system. Compared to the above estimate for general dynamics, Section 4 provides improved precision with respect to the discount coefficient β and the oscillation in the action space.

4. LINEAR QUADRATIC REGULATOR (LQR)

4.1. **The problem setting.** Consider the following LQR problem,

$$\begin{aligned}V^*(s) = & \max_{a_t=\pi(s_t)} \mathbb{E}\left[\int_0^\infty e^{-\beta t} s_t^\top Q s_t + a_t^\top R a_t dt | s_0 = s\right] \\ s.t. \quad & ds_t = (As_t + Ba_t)dt + \sigma dB_t,\end{aligned} \tag{30}$$

Here, $s_t \in \mathbb{R}^d$ and $a_t \in \mathbb{R}^m$ denote the state and action vectors, respectively, while $A, Q \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d\times m}$, and $R \in \mathbb{R}^{m\times m}$ are matrices. The constant $\sigma \in \mathbb{R}$ represents the noise level. Although the deterministic LQR problem corresponds to the case $\sigma \equiv 0$, where there is no stochasticity, we retain the expectation operator $\mathbb{E}$ in the value function definition to maintain a consistent notation throughout the paper. Moreover, unlike the rest of the paper where we assume $\beta > 0$, we also consider the case $\beta = 0$ for the deterministic LQR setting. In fact, the standard LQR problem corresponds to the case where both $\beta = 0$ and $\sigma = 0$.

In this section, we focus on the following problem: given that the continuous-time dynamics are unknown, and only the discrete-time transition distribution $p(s', \Delta t \mid s, a)$, the reward function $r(s, a)$, and the discount coefficient β are available, what is the discretization error associated with the Optimal-BE and Optimal-PhiBE? Here, $p(s', \Delta t \mid s, a)$ denotes the distribution of the state at time Δt given that the initial state is s and a constant action a is applied over the interval $[0, \Delta t)$. Importantly, in this section, we not only lack access to the exact continuous-time dynamics, but we also make no assumption that the underlying system is linear.

By solving the linear SDE for the dynamics, the discrete-time transition distribution $p_{\Delta t}(s'|s, a)$ driven by the continuous-time dynamics (30) is given by

$$p_{\Delta t}(s'|s,a) \sim \mathcal{N}\left((\hat{A}_1\Delta t + I)s + \hat{B}_1 a\Delta t, \sigma^2 C_A \Delta t\right), \quad \text{with} \quad C_A = \frac{1}{\Delta t}\int_0^t e^{A(t-s)}e^{A^\top(t-s)}\,ds, \tag{31}$$

which represents the distribution of the state at time $t + \Delta t$, given that action a is applied constantly over the interval $[t, t + \Delta t)$ from the initial state s. Here

$$\hat{A}_1 = \frac{1}{\Delta t}(e^{Aj\Delta t} - I), \quad \hat{B}_1 = A^{-1}\hat{A}_1 B, \tag{32}$$

Note that C_A can be simplified to $C_A = \frac{1}{2\Delta t}A^{-1}(e^{2A\Delta t} - I)$ when A is symmetric and invertible. Note that for the deterministic LQR (when $\sigma = 0$), $p_{\Delta t}(s'|s, a)$ becomes a deterministic mapping $p_{\Delta t}(s, a)$, which can be written as

$$s_{\Delta t} = p_{\Delta t}(s, a) = (\hat{A}_1 \Delta t + I)s + \hat{B}_1 a \Delta t. \tag{33}$$

We will first review the optimal value function $V^*(s)$ and optimal policy $\pi^*(s)$ for the above LQR problem in Section 4.2. In Section 4.3, we will discuss, given the discrete-time transition density (31), what the optimal policies induced from Optimal-BE and Optimal-PhiBE look like. Finally, in Section 4.4, we will analyze the discretization error and discuss why, and under what conditions, Optimal-PhiBE provides a better approximation to the LQR problem.

4.2. **LQR with known dynamics.** When the dynamics are known, by the classical LQR theorem [1], under the following assumptions, the LQR problem (30) admits a unique optimal control.

Assumption 3 (Wellposedness for LQR)**.** *For $Q, R, A, B \in \mathbb{R}^{d\times d}, \beta \geq 0$, we assume that*

a) *$(A - \beta/2, B)$ are stabilizable, $(A - \beta/2, Q)$ are detectable*[1]*.*
b) *Both Q, R are negative definite.*

Under the above assumptions, the solution to (30) can be equivalently written as the unique solution to the following HJB equation under Assumption 3 as follows,

$$\beta V^*(s) = \max_a [s^\top Q s + a^\top R a + (As + Ba) \cdot \nabla_s V^*(s)] + \frac{\sigma^2}{2}\Delta V^*(s), \tag{34}$$

and the optimal policy is given by,

$$\pi^*(s) = \text{argmax}_a [s^\top Q s + a^\top R a + (As + Ba) \cdot \nabla_s V^*(s)].$$

It is well known that the solution to the standard LQR problem (i.e., $\beta = 0$) is given by the algebraic Riccati equation [1]. For completeness, we provide the Riccati equation for the case $\beta > 0$, along with the explicit solution in the one-dimensional setting in Proposition 4.1.

Proposition 4.1. *Under Assumption 3, the optimal value function $V^*(s) = s^\top P s + \frac{\sigma^2}{\beta} diag(P)$, and the optimal policy $\pi^*(s) = Ks$ with $K = -R^{-1}B^\top P$, where P is the unique negative definite matrix that satisfies*

$$\beta P = Q - PBR^{-1}B^\top P + A^\top P + PA \tag{35}$$

When $d = 1$, the optimal policy are $\pi^*(s) = Ks$ with

$$K = \frac{(\beta/2 - A) - \sqrt{(\beta/2 - A)^2 + \frac{QB^2}{R}}}{B}. \tag{36}$$

The proof of the Proposition is given in Section 7.5.

Remark 4.2. *Note that for the case when $\sigma = \beta = 0$, the solution $V^*(s)$ is unique up to a constant, so we add an additional condition that $V^*(0) = 0$ for $\beta = 0$ to ensure the wellposedness of the HJB equation* (34) *when $\beta = 0$. However, even without this additional condition, the optimal linear policy $\pi^*(s)$ is still unique because it only depends on $\nabla_s V^*(s)$.*

Assumption 3 guarantees the existence and uniqueness of the negative definite matrix solution to the Ricatti equation..

[1]We do not reiterate the formal definitions of stabilizability and detectability for conciseness, as our theoretical development does not depend explicitly on them. However, we assume these conditions hold for all LQR systems considered in this paper.

4.3. **Approximations given discrete-time transition.** In order to unify the symbols, we define

$$\hat{A}_i = \frac{1}{\Delta t}\sum_{j=1}^{i} a_j^{(i)}(e^{Aj\Delta t} - I), \quad \hat{B}_i = A^{-1}\hat{A}_i B, \tag{37}$$

with $a^{(i)}$ defined in (21).

We begin by formulating the Optimal-PhiBE and Optimal-BE for the LQR problem. Directly inserting the transition distribution (31) into the Optimal-BE and Optimal-PhiBE results in the following two equations,

$$\text{Optimal-PhiBE:} \quad \beta\hat{V}_i^*(s) = \max_a [s^\top Q s + a^\top R a + \hat{b}_i(s,a)\cdot\nabla\hat{V}_i^*(s)]. \tag{38}$$

$$\text{Optimal-BE:} \quad \tilde{V}^*(s) = \max_a \left[(s^\top\tilde{Q}s + a^\top\tilde{R}a) + \gamma\int_{\mathbb{S}}\tilde{V}^*(s')p_{\Delta t}(s'|s,a)\,ds'\right], \tag{39}$$

For the Optimal-PhiBE, it is worth noting that we use the deterministic formulation to approximate both stochastic and deterministic LQR problems. This choice is justified by Proposition 4.1, which shows that the optimal policy $\pi^*(s)$ in the LQR setting is independent of the noise level σ. Therefore, we treat both stochastic and deterministic LQR as deterministic systems in our analysis. However, it is important to note that while the underlying dynamics are treated as deterministic, the discrete-time transition distribution available to us remains stochastic.

For higher-order Optimal-PhiBE ($i \geq 2$), we require the distribution of $s_{j\Delta t}$ given $s_0 = s$ and $a_\tau = a$ for all $\tau \in [0, j\Delta t)$. This distribution can be obtained from one step transition density $p_{\Delta t}(s'|s,a)$ iteratively using the following relation,

$$p_{j\Delta t}(s'|s,a) = \int p_{\Delta t}(s'|\hat{s},a)p_{(j-1)\Delta t}(\hat{s}|s,a)d\hat{s}, \quad \text{for } j \geq 2.$$

For the Optimal-BE, the default and natural choices of $\tilde{Q}, \tilde{R}, \gamma$ are

$$\tilde{Q} = Q\Delta t, \quad \tilde{R} = R\Delta t, \quad \gamma = e^{-\beta\Delta t}. \tag{40}$$

However they can be different depending on different approximation methods. For example, if one approximates $\int_0^{\Delta t} e^{-\beta t} r(s_t, a_t)dt$ using $r(s_0,a_0)\int_0^{\Delta t} e^{-\beta t}dt$, then

$$\tilde{Q} = \frac{1}{\beta}(1-\gamma-\beta\Delta t)Q, \quad \tilde{R} = \frac{1}{\beta}(1-\gamma-\beta\Delta t)R.$$

In the general setting, it is unclear which provides a better approximation. However, there is an optimal $\tilde{Q}, \tilde{R}, \gamma$ for the LQR problem, and we will discuss it right after Theorem 4.3 in this section. For now, we use $\tilde{Q}$, $\tilde{R}$, and γ to denote arbitrary choices.

Note that both formulations, Optimal-PhiBE and Optimal-BE, operate without assuming that the underlying control problem is an LQR. They rely only on discrete-time information, namely (31). As shown in Theorem 4.3, when the discrete-time transition dynamics $p_{\Delta t}(s'|s,a)$ are given by (31), Optimal-PhiBE preserves the structure of an LQR problem, corresponding to a modified LQR system with slightly different dynamics. In contrast, Optimal-BE corresponds to a different control problem that involves nontrivial noise.

Furthermore, the optimal policies can be explicitly derived from the Optimal-PhiBE (38) and the Optimal-BE (39). Based on these policies, we provide a precise estimate of the discretization error in the optimal policy. This error analysis is presented in Section 4.4.

The following theorem states that the optimal policy $\tilde{\pi}^*(s), \hat{\pi}_i^*(s)$ obtained from the Optimal-BE (39) and the Optimal-PhiBE (38) can be viewed as the optimal policy obtained from two different stochastic optimal control problems.

Theorem 4.3. *The optimal policy $\hat{\pi}_i^*(s)$ obtained from the Optimal-PhiBE* (38) *is the same as the one obtained from the following stochastic LQR problem,*

$$\begin{aligned} V^*(s) &= \max_{a_t=\pi(s_t)} \mathbb{E}\left[\int_0^\infty e^{\beta t}\left(s_t^\top Q s_t + a_t^\top R a_t\right)dt | s_0 = s\right] \\ s.t. \quad & ds_t = (\hat{A}_i s_t + \hat{B}_i a_t)dt + \sigma dB_t, \end{aligned} \tag{41}$$

with $\hat{A}_i, \hat{B}_i$ defined in (37).

The optimal policy $\tilde{\pi}^*(s)$ obtained from the Optimal-BE (39) is the same as the one obtained from the following stochastic control problem,

$$
\begin{aligned}
V^*(s) = \max_{a_t=\pi(s_t)} \mathbb{E}\left[\int_0^\infty e^{-\hat{\beta}t}\left(s_t^\top \hat{Q} s_t + a_t^\top \hat{R} a_t\right) dt | s_0 = s\right] \\
s.t. \quad ds_t = (\hat{A}_1 s_t + \hat{B}_1 a_t)dt + \left(\hat{A}_1 s_t + \hat{B}_1 a_t\right)\sqrt{\Delta t}\, dB_t,
\end{aligned} \tag{42}
$$

with

$$
\hat{\beta} = \frac{1}{\Delta t \gamma} - \frac{1}{\Delta t}, \quad \hat{Q} = \frac{1}{\gamma \Delta t}\tilde{Q}, \quad \hat{R} = \frac{1}{\gamma \Delta t}\tilde{R}.
$$

Here B_t is a scalar Wiener process.

Proof. The proof of the above theorem is provided in Section 7.6. □

One first notes even when the reward function is known, due to the discretized nature of the MDP framework, the equivalent optimal control problem still has an $O(\Delta t)$ error for the discount coefficient and reward function. Second, for a specific choice of γ, $\tilde{Q}$, and $\tilde{R}$, one can eliminate the bias in the objective function. Specifically, setting

$$
\gamma = \frac{1}{\beta \Delta t + 1}, \quad \tilde{Q} = \Delta t \gamma Q, \quad \tilde{R} = \Delta t \gamma R, \tag{43}
$$

ensures that $\hat{\beta} = \beta$, $\hat{Q} = Q$, and $\hat{R} = R$.

For the remainder of this section, we focus on the discretization error under the optimal choice of $\tilde{Q}$, $\tilde{R}$, and γ. It is important to note that the optimal discretization in (43) is specific to the LQR setting and may not be optimal for general control problems. Moreover, the purpose of this optimal choice is to ensure that the equivalent continuous-time optimal control problem does not introduce bias into the objective function; however, this does not necessarily imply that the optimal choice always leads to a smaller discretization error compared to the default choice. In particular, since the error introduced by the default parameter choice (40) is of order $O(\beta^2 \Delta t)$, adjusting γ, $\tilde{Q}$, and $\tilde{R}$ does not affect the leading-order error when β is small. The dominant source of error remains the transition dynamics.

The MDP framework only utilizes discretized transition dynamics and does not leverage the underlying continuous-time structure. As a result, it effectively treats a deterministic LQR problem as a stochastic LQR problem with $O(\Delta t)$ noise. Furthermore, this artificial noise depends on both the state s and action a, making the discretization error sensitive to both the reward function and the dynamics. We will elaborate on these issues in Section 4.4. Fundamentally, MDP is designed for discrete-time decision-making processes. Even if the continuous-time dynamics are known, there is no mechanism within the Optimal-BE framework to incorporate this structure. In contrast, the Optimal-PhiBE framework not only provides an accurate objective function but also preserves the correct structure of the LQR problem, with only a slight modification to the dynamics matrices A and B.

4.4. **Error analysis.** Since the optimal policies derived from both Optimal-PhiBE and Optimal-BE are linear, we can directly compare their linear coefficients. The following theorem characterizes the difference between the approximated optimal policy and the true optimal policy in the one-dimensional case for both Optimal-PhiBE and Optimal-BE.

Theorem 4.4. *When $d = 1$, for the Optimal-BE, set*

$$
\gamma = \frac{1}{\beta \Delta t + 1}, \quad \tilde{Q} = \Delta t \gamma Q, \quad \tilde{R} = \Delta t \gamma R,
$$

then the difference between the optimal control $\tilde{\pi}^*(s) = \tilde{K}s$ obtained from (39) *and the true optimal control $\pi^*(s) = Ks$ can be bounded by*

$$
\left|\tilde{K} - K\right| \lesssim \left[|B|\left(\sqrt{\frac{Q}{R}} + \frac{|A|}{|B|}\right)^2 + |A - \beta/2|\frac{|A|}{|B|}\right]\Delta t + O(\Delta t^2).
$$

The difference between the optimal control $\hat{\pi}_i^*(s) = \hat{K}_i s$ from the Optimal-PhiBE (38) *and the true optimal control $\pi^*(s) = Ks$ can be bounded by*

$$\left|\hat{K}_i - K\right| \leq \beta \frac{|A|^i}{|B|} \Delta t^i + O(\Delta t^{2i}),$$

Especially, when $\beta = 0$, the Optimal-PhiBE approximation gives the exact optimal control, i.e.,

$$\hat{K}_i = K; \tag{44}$$

Proof. The proof of the above theorem is given in Section 7.7. □

Several remarks are in order. First, when $\beta = 0$, the first-order Optimal-PhiBE exactly recovers the optimal policy, whereas Optimal-BE provides only a first-order approximation. It is counterintuitive that Optimal-PhiBE can recover the exact continuous-time optimal policy using only discrete-time transition information. Notably, this phenomenon occurs when $\beta = 0$, a setting typically considered more challenging than cases with larger β.

Second, when $\beta > 0$, i-th order Optimal-PhiBE provides i-th order accuracy, as shown in the right plot of Figure 4. We mainly discuss the error difference between first-order Optimal-PhiBE and Optimal-BE. Both discretization errors are $O(\Delta t)$. However, the first-order Optimal-PhiBE error is independent of the reward function, whereas the Optimal-BE error is influenced by it. In particular, if the reward function oscillates rapidly in the state space but changes slowly in the action space, the resulting Optimal-BE error is large. This observation is consistent with Theorem 3.8 and aligns with the conclusions in [50]. In many reinforcement learning applications, where the goal is sparse or defined by reaching specific states, the reward function must be designed to vary sharply in the state space, which leads to significant discretization error. This helps explain why sparse rewards often result in slow convergence and instability during training. When the system is state-dominant, that is, with large $|A|$ and small $|B|$, both error increases. However, the Optimal-PhiBE benefits from an additional discount coefficient β, which results in smaller errors when β is small, which is a typical setting in RL applications. This, too, is counterintuitive, as smaller β (i.e., less discounting) usually corresponds to harder problems. In contrast, when the system is control-dominant, that is, with small $|A|$ and large $|B|$, Optimal-PhiBE yields small error, while Optimal-BE remains inaccurate since its error increases with $|B|$. As shown in the second plot of Figure 4, the Optimal-BE error remains small only when $|A|$ and Q/R are small, and $|B|$ is neither too small nor too large. If any of these conditions are violated, the error becomes significant.

Finally, we discuss the effect of β on the discretization error. It is important to note that the error bound for Optimal-PhiBE is not sharp when β is large. As shown in the third plot of Figure 4, the error of Optimal-PhiBE increases linearly with respect to β when β is small, but begins to decrease as β continues to grow. In contrast, the error of Optimal-BE is relatively insensitive to changes in β. Numerically, we observe that the Optimal-BE error appears to achieve a minimum at a positive value of β. Moreover, there exits a small regime where the Optimal-BE error is smaller than that of Optimal-PhiBE.

To summarize, the first-order Optimal-PhiBE outperforms Optimal-BE in several practically relevant scenarios. These include cases where the discount coefficient β is small, the reward function varies rapidly in the state space but slowly in the action space, or the natural system evolves slowly without control. All of these conditions are common in real world applications.

Next, we consider the multi-dimensional case where $d > 1$. Extending the results from the one-dimensional case is not straightforward. First, in contrast to the one-dimensional setting where well-posedness always holds, the multi-dimensional case requires Assumption 3 to ensure well-posedness. To rigorously establish the results, we need to verify this property within the Optimal-PhiBE framework, which is presented in Lemma 4.5. Second, extending the discretization error analysis from one to multiple dimensions is also nontrivial. In the one-dimensional case, the explicit form of the optimal policy allows for direct error computation. However, in higher dimensions, such an explicit form is generally unavailable. For the Optimal-PhiBE formulation, since it remains a continuous LQR problem, the discretization error can be analyzed using perturbation theory for the continuous algebraic Riccati equation (ARE), as developed in the existing literature [1]. On the other hand, the solution under the Optimal-BE formulation satisfies a discrete ARE, which differs in structure from the continuous case. While there are perturbation results available for the discrete ARE [4], to our knowledge, no existing work directly addresses the error incurred when interpreting a continuous LQR

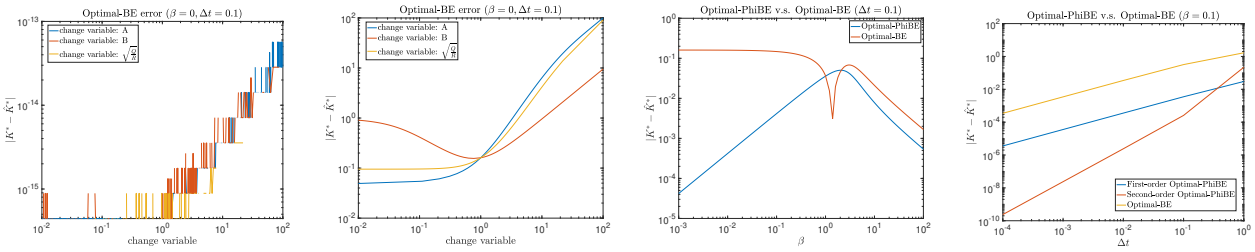


FIGURE 4. Comparison of the optimal policy error from Optimal-PhiBE and Optimal-BE. The left plot shows that Optimal-PhiBE exactly recovers the optimal policy when $\beta = 0$. The second plot illustrates how the reward and dynamics influence the error of Optimal-BE. The third plot shows how the discount coefficient β affects the errors. The right plot demonstrates that when $\beta > 0$, both Optimal-PhiBE and Optimal-BE achieve first-order approximation with respect to Δt, while the second-order PhiBE achieves second-order approximation.

system through a discrete lens. Since this is not the main focus of our work, we leave a rigorous treatment of the discretization error for Optimal-BE to future study. Nonetheless, as demonstrated in the numerical experiments in Section 6, the discretization error is influenced in a similar manner as in the one-dimensional case.

We first prove the wellposedness of the PhiBE approximation.

Lemma 4.5. *For Δt sufficiently small, if $(A - \beta/2, B)$ is stablizable and $(A - \beta/2, Q)$ is detectable, then $(\hat{A}_i - \beta/2, \hat{B}_i)$ is also stablizable and $(\hat{A}_i - \beta/2, Q)$ is also detectable where $\hat{A}_i, \hat{B}_i$ are defined in* (37)*.*

The proof of the above lemma is given in Section 7.8. Next, we present the error analysis for the Optimal-PhiBE in multi-dimensional case.

Theorem 4.6. *When $d > 1$, for Δt sufficiently small, such that $\hat{A}_i, \hat{B}_i$ still satisfies the Assumption 3/a), then when $\beta = 0$*

$$\hat{K}_i = K;$$

when $\beta > 0$,

$$\left\| \hat{K}_i - K \right\| \leq \beta p \hat{C}_i \kappa(A) \kappa(B) \left\| A \right\|^i \left\| B^{-1} \right\| \Delta t^i + O(\Delta t^{i+1})$$

where p defined in (106) *is a constant depending on $B^{-1}A, B, R, Q, \beta$, and $\hat{C}_i$ defined in* (53) *is a constant only depending on the order.*

Proof. The proof of the above Theorem is given in Section 7.9. □

Several remarks are in order. First, the discretization error associated with Optimal-PhiBE is mainly influenced by the system dynamics, represented by the matrices A and B, and the discount coefficient β. When A and B are well-conditioned, the behavior of the discretization error is similar to that of the one-dimensional case. In particular, Optimal-PhiBE recovers the exact optimal policy when $\beta = 0$. For positive β, the error remains small if the natural dynamics evolve slowly in the state space and respond strongly to control inputs. This scenario is typically considered easier to control when the system dynamics are known. In contrast, if the matrices A and B are poorly conditioned or the system is dominated by state dynamics, then the underlying system is inherently more difficult to control, and the discretization error from Optimal-PhiBE becomes larger.

5. Model-free Policy Iteration algorithm

In Section 2.1 - 4, we assumed access to the discrete transition dynamics (7) and introduced Optimal-PhiBE as a theoretical framework. We established that the solution to Optimal-PhiBE (18) provides a close approximation to the optimal value function of the original problem (4) and induces a policy that closely approximates the optimal policy. Consequently, finding the solution to Optimal-PhiBE (18) is of significant importance.

However, directly solving (18) is challenging, even when the discrete transition dynamics (7) are accessible. In Section 5.1, we outline the Policy Iteration (PI) algorithm for solving (18) in a linear space, assuming access to the discrete transition dynamics. In Section 5.2, we propose a model-free algorithm based on the framework introduced in Section 5.1, which only uses collected discrete data points.

5.1. **Policy Iteration (PI) for solve HJB equation in linear function space.** In this section, we review the policy iteration algorithm to solve the HJB equation (5) or the Optimal-PhiBE equation (18) when b, Σ or $(\hat{bi}, \hat{\Sigma}_i)$ are known [5, 12, 13, 32, 34, 43]. In a policy iteration framework, the optimal value function and policy are updated iteratively. Starting with the policy π^k from the previous iteration, the value function V^k under policy π^k defined as follows

$$\beta V^k(s) = r^{\pi_k}(s) + b^{\pi_k}(s) \cdot \nabla V^k(s) + \frac{1}{2}\Sigma^{\pi_k}(s) : \nabla^2 V^k(s). \tag{45}$$

is evaluated. Then the *state-action value function* $q^k(s,a)$, as introduced in [16, 42], is estimated under this policy, which is defined as follows,

$$q^k(s,a) = r(s,a) + b(s,a) \cdot \nabla V^k(s) + \frac{1}{2}\Sigma(s,a) : \nabla^2 V^k(s). \tag{46}$$

Intuitively, the state-action value function evaluates how advantageous it is to take action a in state s based on the expected cumulative reward, which is derived as the limit of the rescaled advantage function. The policy is then updated by setting $\pi^{k+1}(s) = \arg\max_{a\in\mathcal{A}} q^k(s,a)$, based on the current estimate $q^k(s,a)$. It is known that the value functions V^k converge exponentially fast as $k \to \infty$.

Numerically, one can approximate V^k, q^k in a linear function space $\{\theta^\top \Phi(s) \mid \theta \in \mathbb{R}^n\}$ and $\{w^\top \Psi(s,a) \mid w \in \mathbb{R}^m\}$, where $\Phi(s) = (\phi_i(s))_{i=1}^n$ and $\Psi(s,a) = (\psi_i(s,a))_{i=1}^m$ represent the corresponding column vector of basis functions. Applying the Galerkin method to (45) and (46), one has an estimate of the coefficient θ, w.

$$\begin{aligned}
&\left[\int \Phi(s)\left(\beta\Phi(s) - \left(b^{\pi_k}(s)\cdot\nabla\Phi(s) + \frac{1}{2}\Sigma^{\pi_k}(s) : \nabla^2\Phi(s)\right)\right)^\top \mu(s)\,ds\right]\theta = \left[\int r^{\pi_k}(s)\Phi(s)\,ds\right], \\
&\left[\int \Psi(s,a)\Psi(s,a)^\top\,ds\,da\right] w = \left[\int \left(r(s,a) + b(s,a)\cdot\nabla V^k(s) + \frac{1}{2}\Sigma(s,a) : \nabla^2 V^k(s)\right)\Psi(s,a)\nu(s,a)\,ds\,da\right],
\end{aligned} \tag{47}$$

where the operator $b \cdot \nabla + \frac{1}{2}\Sigma : \nabla^2$ is applied entry-wise to the basis vector $\Phi(s), \Psi(s,a)$. The Galerkin method approximates the solution using an ansatz and projects the PDE onto the chosen linear space by multiplying both sides with $\Phi(s)$ and integrating with respect to a measure. This yields two independent linear systems for θ and ω. Note that $V^k(s)$ in the second equation is represented by $V^k(s) = \theta^\top\Phi(s)$, where $\theta \in \mathbb{R}^n$ is obtained by solving the first linear system.

After obtaining $q(s,a) = \omega^\top \Psi(s,a)$, the updated policy is defined as,

$$\pi_{k+1} = \mathrm{argmax}_a q(s,a).$$

5.2. **Model-free algorithm.** In the previous section, we presented a policy iteration framework for solving general HJB equations within a linear function space, assuming access to the drift and diffusion terms (b, Σ) or to the discrete-time transition dynamics. When only trajectory data (6) are accessible, integrals are approximated by empirical sums over the collected data points, while expectations are estimated using unbiased samples drawn from the data.

We now present the details of the model-free PI algorithm based on Optimal-PhiBE. We follow the same policy iteration pipeline: for a given policy π_k, we first approximate the corresponding value function V^k within a linear function space. Using this approximation, we then estimate the state-action value function q^k in another linear space, and finally update the policy based on q^k.

To approximate V^k, [50] proposed a data-dependent method for solving the linear system (47) derived from the PDE formulation of PhiBE (16). This method is summarized in Algorithm 1, and we omit the technical details of its derivation here.

Once the approximation of V^k is obtained, we proceed to estimate q^k from data. Specifically, we aim to find an approximation of the form $q_w^k(s,a) = w^\top \Psi(s,a)$ by solving the following linear system,

$$\left[\int \Psi(s,a)\Psi(s,a)^\top \, ds\, da\right] w = \left[\int \left(r(s,a) + \hat{b}_i(s,a)\cdot \nabla V^k(s) + \frac{1}{2}\hat{\Sigma}_i(s,a) : \nabla^2 V^k(s)\right)\Psi(s,a)\, ds\, da\right]. \tag{48}$$

where $\hat{b}_i$ and $\hat{\Sigma}_i$ are defined in (19) and (20) respectively. Since the integrals and expectations are not directly computable, we approximate them using collected empirical data $\{s^l_{i\Delta t}, a^l_{i\Delta t}, r^l_{i\Delta t}\}_{i=0,l=1}^{i=I,l=L}$. For example, an unbiased estimate of

$$\hat{b}_1(s^l_{j\Delta t}, a^l_{j\Delta t}) = \mathbb{E}\left[\frac{1}{\Delta t}(s_{\Delta t} - s_0)\middle| s_0 = s^l_{j\Delta t}, a_\tau = a^l_{j\Delta t} \text{ for } \tau \in [0,\Delta t)\right]$$

is given by

$$\hat{b}_1(s^l_{j\Delta t}, a^l_{j\Delta t}) \approx \frac{1}{\Delta t}(s^l_{(j+1)\Delta t} - s^l_{j\Delta t}).$$

Similarly,

$$\hat{\Sigma}_1(s^l_{j\Delta t}, a^l_{j\Delta t}) \approx \frac{1}{\Delta t}(s^l_{(j+1)\Delta t} - s^l_{j\Delta t})(s^l_{(j+1)\Delta t} - s^l_{j\Delta t})^\top.$$

Thus, when $i = 1$, the integrand on the right-hand side of (48) evaluated at a single data point $(s^l_{j\Delta t}, a^l_{j\Delta t})$ admits the following unbiased estimator,

$$\begin{aligned}&\left(r(s^l_{j\Delta t}, a^l_{j\Delta t}) + \left(\frac{1}{\Delta t}(s^l_{(j+1)\Delta t} - s^l_{j\Delta t})\right)\cdot \nabla V^k(s^l_{j\Delta t})\right.\\ &\left. + \frac{1}{2}\left(\frac{1}{\Delta t}(s^l_{(j+1)\Delta t} - s^l_{j\Delta t})(s^l_{(j+1)\Delta t} - s^l_{j\Delta t})^\top\right) : \nabla^2 V^k(s^l_{j\Delta t})\right)\Psi(s^l_{j\Delta t}, a^l_{j\Delta t})\end{aligned}$$

which is denoted by $f_{j,l}(w)$. For the integrand on the left-hand side of (48), its evaluation at the data point $(s^l_{j\Delta t}, a^l_{j\Delta t})$ is given by $\Psi(s^l_{j\Delta t}, a^l_{j\Delta t})\Psi(s^l_{i\Delta t}, a^l_{i\Delta t})^\top$, which is denoted by $g_{j,l}$. Then, to obtain w, it suffices to solve the following linear system,

$$\left[\sum_{l=1}^{I}\sum_{j=0}^{m-1} g_{j,l}\right] w = \left[\sum_{l=1}^{I}\sum_{j=0}^{m-1} f_{j,l}\right].$$

An alternative approach to finding w is gradient descent (Algorithm 2), where the function approximation is achieved by minimizing a loss function. This method can be generalized to nonlinear function approximation. While we do not provide a detailed explanation of this method here, it serves as another effective tool for solving such problems.

The practical framework of our algorithm is summarized as follows, while the complete pseudocode is provided in Algorithm 4.

Preparation: Collect trajectory data $\{s^l_{i\Delta t}, a^l_{i\Delta t}, r_{i\Delta t}\}_{i=0,l=1}^{i=I,l=L}$ with $a_j \sim$Unif$(\mathcal{A})$ and s^l_0 randomly sampled from the state space. This data consists of state-action pairs along with the corresponding rewards, where I denotes the number of samples per trajectory and L represents the number of trajectories. This batch of collected data will be used to evaluate the state-action value function q^k in each iteration.

At the k-th iteration:

Step 1: For the policy $\pi^k(s)$ from the previous iteration, collect data $\{\tilde{s}^l_{i\Delta t}, \tilde{a}^l_{i\Delta t} = \pi^k(s^l_{i\Delta t}), \tilde{r}_{i\Delta t}\}_{i=0,l=1}^{i=I,l=L}$ according to policy π^k. Once the data is collected, compute the approximated value function $V^k(s) = \theta^\top \Phi(s)$, using the Galerkin method outlined in Algorithm 1, which is proposed in [50].

Step 2: Use the value function V^k and trajectory data $\{s^l_{i\Delta t}, a^l_{i\Delta t}, r^l_{i\Delta t}\}_{i=0,l=1}^{i=I,l=L}$ to approximate the state-action value function q^k, either by applying the Galerkin method (Algorithm 3) or Gradient Descent (Algorithm 2).

Step 3: Update the policy using $\pi^{k+1}(s) = \arg\max_a \hat{q}^{\pi^k}(s,a)$.

Algorithm 1 PhiBE_Policy_Evaluation($\Delta t, \beta, B, \Phi, i$) - i-th order Policy evaluation

Input: discrete time step Δt, discount coefficient β, discrete-time trajectory data $B = \{(s^l_{j\Delta t}, r^l_{j\Delta t})_{j=0}^m\}_{l=1}^I$ generated by applying corresponding policy π, finite bases $\Phi(s) = (\phi_1(s), ..., \phi_n(s))^\top$ and the order of the method i.

Output: Value function $\hat{V}^\pi$.

Compute

$$A_i = \sum_{l=1}^{I} \sum_{j=0}^{m-i} \Phi(s^l_{j\Delta t}) \Big[\beta \Phi(s^l_{j\Delta t}) - \hat{\mu}_r(s^l_{j\Delta t}) \cdot \nabla \Phi(s^l_{j\Delta t}) - \frac{1}{2} \hat{\Sigma}_r(s^l_{j\Delta t}) : \nabla^2 \Phi(s^l_{j\Delta t}) \Big]^\top,$$

where

$$\hat{\mu}_i(s^l_{j\Delta t}) = \sum_{k=1}^{i} a^i_k (s^l_{(j+k)\Delta t} - s_{j\Delta t}) \text{ and } \hat{\Sigma}_i(s^l_{j\Delta t}) = \frac{1}{\Delta t} \sum_{k=1}^{i} a^i_k (s^l_{(j+k)\Delta t} - s_{j\Delta t})(s^l_{(j+k)\Delta t} - s_{j\Delta t})^\top$$

with a^i_k defined in (15).

Compute

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

Compute

$$\theta = A_i^{-1} b_i.$$

return $\hat{V}^\pi(s) = \theta^\top \Phi(s)$.

Algorithm 2 Q_Gradient_descent_PhiBE($\Delta t, \beta, B, \Psi, \hat{V}^\pi, w_0, \alpha, i$) - i-th order Gradient descent for $\hat{q}^\pi$

Input: discrete time step Δt, discount coefficient β, discrete-time trajectory data $B = \{(s^l_{j\Delta t}, a^l_{j\Delta t}, r^l_{j\Delta t})_{j=0}^m\}_{l=1}^I$, finite bases $\Psi(s,a) = (\psi_1(s,a), ..., \psi_n(s,a))^\top$, approximated value function $\hat{V}^\pi$, initial coefficient with respect to the finite bases w_0, step size of the gradient descent α, and the order of the method i.

Output: Continuous q-function for policy π.

Initialize $w = w_0$.

while not *Stopping Criterion Satisfied* **do**

Compute $\nabla \hat{V}^\pi(s)$, $\nabla^2 \hat{V}^\pi(s)$, and $q^\pi_w(s,a) = w^\top \Psi(s,a)$ as functions using their representation under the bases.

Compute approximate gradient

$$\hat{F}(w) = \frac{1}{|B|} \sum_{l=1}^{I} \sum_{j=0}^{m-i} \Big[r^l_{j\Delta t} + \hat{\mu}_i(s^l_{j\Delta t}, a^l_{j\Delta t}) \cdot \nabla \hat{V}^\pi(s^l_{j\Delta t}) + \tfrac{1}{2} \hat{\Sigma}_i(s^l_{j\Delta t}, a^l_{j\Delta t}) : \nabla^2 \hat{V}^\pi(s^l_{j\Delta t}) - q^\pi_w(s^l_{j\Delta t}, a^l_{j\Delta t}) \Big] \left(-\Psi(s^l_{j\Delta t}, a^l_{j\Delta t}) \right),$$

where

$$\hat{\mu}_i(s^l_{j\Delta t}, a^l_{j\Delta t}) = \sum_{k=1}^{i} a^i_k (s^l_{(j+k)\Delta t} - s_{j\Delta t}) \text{ and } \hat{\Sigma}_r(s^l_{j\Delta t}, a^l_{j\Delta t}) = \frac{1}{\Delta t} \sum_{k=1}^{i} a^i_k (s^l_{(j+k)\Delta t} - s_{j\Delta t})(s^l_{(j+k)\Delta t} - s_{j\Delta t})^\top$$

with a^i_k defined in (15).

Update the coefficient vector:

$$w \leftarrow w - \alpha \hat{F}(w).$$

end while

return $\hat{q}^\pi(s,a) = w^\top \Psi(s,a)$.

6. Numerical Experiments

6.1. **Linear Quadratic Regulator (LQR) Problem.** In this subsection, we consider the infinite-horizon linear quadratic regulator (LQR) problem. The state evolves according to the (stochastic) differential equation:

$$ds_t = (As_t + Ba_t)\,dt + \sigma\,dB_t, \quad s_0 = s,$$

Algorithm 3 Q_Galerkin_phibe($\Delta t, \beta, B, \Psi, \hat{V}^\pi, i$) - i-th order Galerkin method for $\hat{q}^\pi$

Input: discrete time step Δt, discount coefficient β, discrete-time trajectory data $B = \{(s^l_{j\Delta t}, a^l_{j\Delta t}, r^l_{j\Delta t})_{j=0}^m\}_{l=1}^I$, finite bases $\Psi(s,a) = (\psi_1(s,a), ..., \psi_n(s,a))^\top$, approximated value function $\hat{V}^\pi$, and the order of the method i.

Output: Continuous q-function for policy π.

Compute

$$A_i^q = \sum_{l=1}^{I} \sum_{j=0}^{m-i} \Psi(s^l_{j\Delta t}, a^l_{j\Delta t}) \Psi(s^l_{j\Delta t}, a^l_{j\Delta t})^\top$$

Compute

$$b_i^q = \sum_{l=1}^{I} \sum_{j=0}^{m-i} \left[r^l_{j\Delta t} + \hat{\mu}_r(s^l_{j\Delta t}, a^l_{j\Delta t}) \cdot \nabla \hat{V}^\pi(s^l_{j\Delta t}) + \frac{1}{2} \hat{\Sigma}_i(s^l_{j\Delta t}, a^l_{j\Delta t}) : \nabla^2 \hat{V}^\pi(s^l_{j\Delta t}) \right] \Psi(s^l_{j\Delta t}, a^l_{j\Delta t}),$$

where

$$\hat{\mu}_i(s^l_{j\Delta t}, a^l_{j\Delta t}) = \sum_{k=1}^{i} a_k^i (s^l_{(j+k)\Delta t} - s_{j\Delta t}) \text{ and } \hat{\Sigma}_i(s^l_{j\Delta t}, a^l_{j\Delta t}) = \frac{1}{\Delta t} \sum_{k=1}^{i} a_k^i (s^l_{(j+k)\Delta t} - s_{j\Delta t})(s^l_{(j+k)\Delta t} - s_{j\Delta t})^\top$$

with a_k^i defined in (15).

Compute

$$w = (A_i^q)^{-1} b_i^Q.$$

return $\hat{q}^\pi(s,a) = w^\top \Psi(s,a)$.

Algorithm 4 Optimal_Phibe($\Delta t, \beta, \Phi, \Psi, \pi_0, \alpha, i$) - i-th order Optimal-PhiBE algorithm for finding the optimal policy

Input: discrete time step Δt, discount coefficient β, finite bases for policy evaluation $\Phi(s) = (\phi_1(s), ..., \phi_{n'}(s))^\top$, finite bases for q-approximation $\Psi(s,a) = (\psi_1(s,a), ..., \psi_n(s,a))^\top$, initial policy π_0, step size of the gradient descent α, and the order of the method i.

Output: Optimal policy $\pi^*(s)$, optimal value function $V^{\pi^*}(s)$.

Initialize $\pi(s) = \pi_0(s)$.

Generate $B^q = \{(s^l_{j\Delta t}, a^l_{j\Delta t}, r^l_{j\Delta t})_{j=0}^m\}_{l=1}^I$ for continuous q-function approximation.

while not *Stopping Criterion Satisfied* **do**

  Generate data $B^\pi = \{(s^l_{j\Delta t}, r^l_{j\Delta t})_{j=0}^{m'}\}_{l=1}^{I'}$ by applying policy π.

  Call Algorithm 1 to obtain

$$\hat{V}^\pi(s) = \text{Phibe\_Policy\_Evaluation}(\Delta t, \beta, B^\pi, \Phi, i).$$

  Call Algorithm 3 to obtain

$$\hat{q}^\pi(s,a) = \text{q\_Galerkin\_phibe}(\Delta t, \beta, B^q, \Psi, \hat{V}^\pi, i),$$

  or inherit w_0 from the last iteration and call Algorithm 2 to obtain

$$\hat{q}^\pi(s,a) = \text{q\_Gradient\_Descent\_phibe}(\Delta t, \beta, B^q, \Psi, \hat{V}^\pi, w_0, \alpha, i).$$

  Update the optimal policy:

$$\pi(s) \leftarrow \text{argmax}_a \, \hat{q}^\pi(s,a).$$

end while

return $\pi^*(s) = \pi(s)$, $V^{\pi^*}(s) = \hat{V}^\pi(s)$.

where $A, B \in \mathbb{R}^{d\times d}$ and $\sigma \geq 0$. The objective is to maximize the value function

$$\pi^*(s) = \text{argmax}_\pi V^\pi(s), \quad \text{where} \quad V^\pi(s) = \mathbb{E}\left[\int_0^\infty e^{-\beta t} (s_t^\top Q s_t + a_t^\top R a_t)\, dt, \middle| s_0 = s \right] \quad \text{with} \quad a_t = \pi(s_t),$$

where Q and R are negative definite matrices in $\mathbb{R}^{d\times d}$. In all the experiments, we assume that observations can only be made at discrete intervals of Δt, and actions can only be updated at discrete intervals of Δt.

6.1.1. *Comparison of Policy Iteration Algorithms in Deterministic and Stochastic LQR Settings.* In this section, we compare our proposed PI based on Optimal-PhiBE (Algorithm 4) (first-order and second-order) with PI based on Optimal-BE (Algorithm 7) under both the deterministic and stochastic settings, considering one-dimensional ($d = 1$) and two-dimensional ($d = 2$) systems.

For the one-dimensional deterministic case (Figure 5), we consider the following four examples:

- (Case 1) $A = 1$, $B = 1$, $R = -1$, $Q = -1$, $\sigma = 0$, $\beta = 0$, and $\Delta t = 2$.
- (Case 2) $A = 1$, $B = 0.1$, $R = -1$, $Q = -1$, $\sigma = 0$, $\beta = 0$, and $\Delta t = 1$.
- (Case 3) $A = 1$, $B = 1$, $R = -0.01$, $Q = -100$, $\sigma = 0$, $\beta = 0$, and $\Delta t = 0.1$.
- (Case 4) $A = 100$, $B = 1$, $R = -1$, $Q = -1$, $\sigma = 0$, $\beta = 0$, and $\Delta t = 0.01$.

For the one-dimensional stochastic case (Figure 6), we consider the same $A, B, R, Q, \Delta t$ as the one-dimensional deterministic case, but with different $\beta = 0.01, \sigma = 1$.

And in two-dimensional deterministic case (Figure 7), we consider the following four examples:

- (Case 1) $A = \begin{pmatrix} -9.375 & -3.125 \\ -3.125 & -9.375 \end{pmatrix}$, $B = \begin{pmatrix} 10 & 1 \\ 1 & 10 \end{pmatrix}$, $R = \begin{pmatrix} -12 & -3 \\ -3 & -8 \end{pmatrix}$, $Q = \begin{pmatrix} -10 & -2 \\ -2 & -10 \end{pmatrix}$, $\sigma = 0$, $\beta = 0$, and $\Delta t = 2$.
- (Case 2) $A = \begin{pmatrix} -1.875 & -0.625 \\ -0.625 & -1.875 \end{pmatrix}$, $B = \begin{pmatrix} 0.6 & 0.2 \\ 0.2 & 0.6 \end{pmatrix}$, $R = \begin{pmatrix} -12 & -3 \\ -3 & -8 \end{pmatrix}$, $Q = \begin{pmatrix} -10 & -2 \\ -2 & -10 \end{pmatrix}$, $\sigma = 0$, $\beta = 0$, and $\Delta t = 1$.
- (Case 3) $A = \begin{pmatrix} -0.9375 & -0.3125 \\ -0.3125 & -0.9375 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1.01 \end{pmatrix}$, $R = \begin{pmatrix} -9 & -8.5 \\ -8.5 & -9 \end{pmatrix}$, $Q = \begin{pmatrix} -106.6667 & 93.3333 \\ 93.3333 & -106.6667 \end{pmatrix}$, $\sigma = 0$, $\beta = 0$, and $\Delta t = 0.1$.
- (Case 4) $A = \begin{pmatrix} 10.6667 & -9.3333 \\ -9.3333 & 10.6667 \end{pmatrix}$, $B = \begin{pmatrix} 0.9 & 0.85 \\ 0.85 & 0.88 \end{pmatrix}$, $R = \begin{pmatrix} -12 & -3 \\ -3 & -8 \end{pmatrix}$, $Q = \begin{pmatrix} -10 & -2 \\ -2 & -10 \end{pmatrix}$, $\sigma = 0$, $\beta = 0$, and $\Delta t = 0.01$.

Finally in two-dimensional stochastic case (Figure 8), we consider the same $A, B, R, Q, \Delta t$ as the two-dimensional deterministic case, but with different $\beta = 0.01, \sigma = 1$.

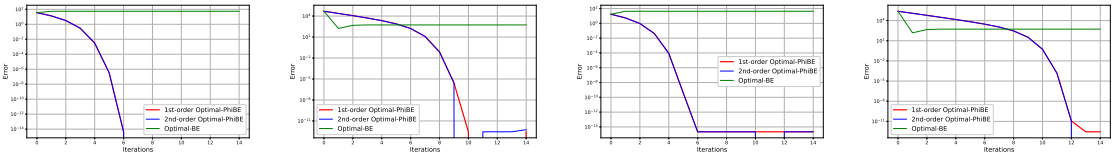
For both algorithms, we use identical configurations in each experiment. For conciseness, details on the ground-truth optimal policies and value functions, as well as the complete experimental setup, are provided in Appendix 9. Each algorithm is run for 15 iterations per example, and we report the L^2 distance between the value function of the iterated policy and the ground-truth optimal value function.

The one-dimensional results for the deterministic and stochastic settings are shown in Figure 5 and Figure 6, respectively, where the x-axis indicates the number of iterations and the y-axis reports the $L^2([-3, 3])$ distance. The corresponding two-dimensional results are presented in Figure 7 and Figure 8, where the y-axis shows the $L^2([-3, 3]^2)$ distance.

It can be observed that, across all settings, the policy iteration (PI) method based on Optimal-PhiBE consistently outperforms the PI algorithm based on Optimal-BE. In the deterministic settings for both 1D and 2D systems, Optimal-PhiBE is able to closely recover the exact solution, even in moderately ill-conditioned cases. In the stochastic settings, Optimal-PhiBE again demonstrates superior performance, consistently providing much lower L^2 errors compared to Optimal-BE. Notably, in ill-conditioned scenarios—such as Case 4, Optimal-PhiBE still maintains a relatively small error, highlighting its robustness to problem conditioning.

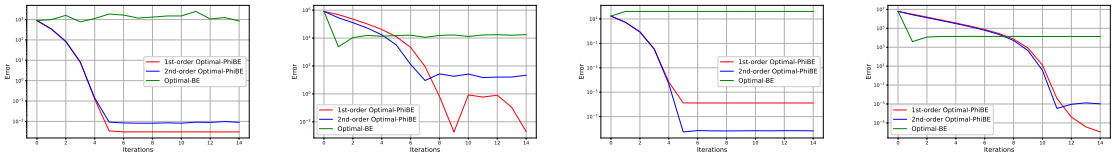
6.1.2. *Role of Δt in deterministic LQR problem.* In this section, we validate the error order with respect to Δt, as established in Theorem 4.4 and Theorem 4.6.

We vary $\Delta t \in \{5\times10^{-3}, 10^{-2}, 5\times10^{-2}, 10^{-1}, 1, 2.5\}$ and evaluate the performance of PI based on Optimal-PhiBE (first- and second-order) alongside the PI based on Optimal-BE. To minimize noise induced by random data sampling, we use batch sizes significantly larger than the algorithms actually require, for instance, 6×10^6 data points per iteration in the 1-D case and 6×10^4 data points in the 2-D case. All three algorithms are configured identically, sharing the same initializations.



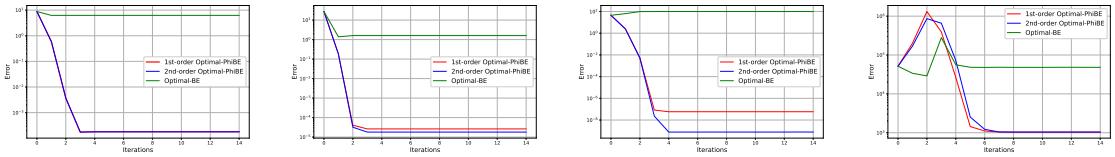
(A) (B) (C) (D)

FIGURE 5. Comparison in the one-dimensional deterministic case. (A) Case 1, where Δt is large. (B) Case 2, where $|A/B|$ is large. (C) Case 3, where $|Q/R|$ is large. (D) Case 4, where $|A|$ is large.



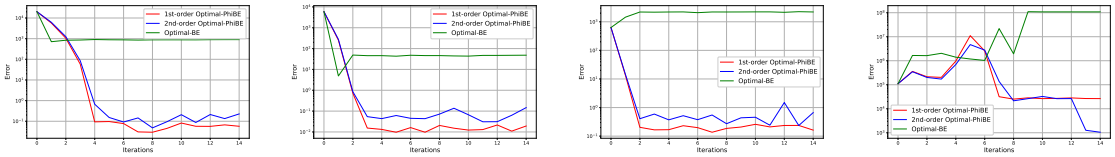
(A) (B) (C) (D)

FIGURE 6. Comparison in the one-dimensional stochastic case. (A) Case 1, where Δt is large. (B) Case 2, where $|A/B|$ is large. (C) Case 3, where $|Q/R|$ is large. (D) Case 4, where $|A|$ is large.



(A) (B) (C) (D)

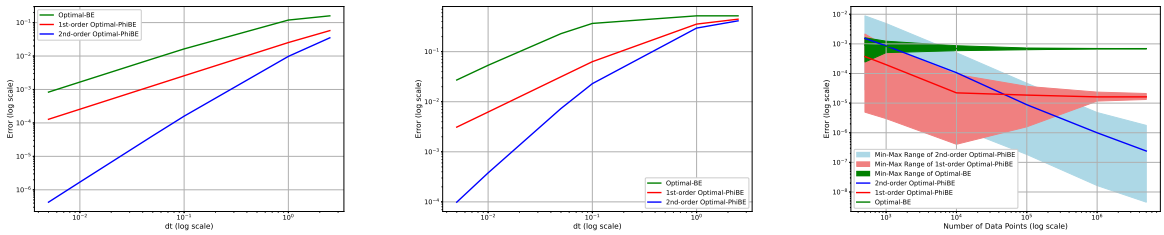
FIGURE 7. Comparison in the two-dimensional deterministic case. (A) Case 1, where Δt is large. (B) Case 2, where $\|A\|/\|B\|$ is large. (C) Case 3, where $\|Q\|/\|R\|$ is large. (D) Case 4, where $\|A\|$ is large.



(A) (B) (C) (D)

FIGURE 8. Comparison in the two-dimensional stochastic case. (A) Case 1, where Δt is large. (B) Case 2, where $\|A\|/\|B\|$ is large. (C) Case 3, where $\|Q\|/\|R\|$ is large. (D) Case 4, where $\|A\|$ is large.

For each value of Δt, we conduct 30 experiments and compute the mean of $\|K - K^*\|$ over these runs, where K^* is the coefficient from the ground truth optimal policy $\pi^*(s) = K^*s$, and K is the corresponding coefficient estimated by the algorithms.



(A) Time step scaling in 1D case

(B) Time step scaling in 2D case

(C) Effect of Batch Size

FIGURE 9. Time step scaling and effect of batch size

The corresponding LQR problem setups are as follows:

- (1-dimension) $A = -1.0, B = 0.5, \sigma = 0, R = -1, Q = -1, \beta = 1$.
- (2-dimension) $A = \begin{pmatrix} -9.375 & -3.125 \\ -3.125 & -9.375 \end{pmatrix}$, $B = \begin{pmatrix} 10 & 1 \\ 1 & 10 \end{pmatrix}$, $\sigma = 0$, $R = \begin{pmatrix} -12 & -3 \\ -3 & -8 \end{pmatrix}$, $Q = \begin{pmatrix} -10 & -2 \\ -2 & -10 \end{pmatrix}$, $\beta = 10$.

The results (log-log graph) are presented in Figure 9a and 9b.

It can be observed that both PI based on Optimal- PhiBE (first order) and the PI based on Optimal-BE algorithms exhibit first-order error, while the PI based on Optimal-PhiBE (second order) algorithm achieves second-order error. This result validates the theoretical guarantees established in Theorem 4.4 and Theorem 4.6, confirming the expected error rates.

6.1.3. *Role of the number of data points.* In this section, we investigate how the number of data points affects the error.

Let B denote the number of data points used per iteration. We vary $B \in \{5\times10^2, 10^3, 10^4, 10^5, 10^6, 5\times10^6\}$ and evaluate the performance of PI based on Optimal-PhiBE (first- and second-order) algorithm alongside the PI based on Optimal-BE algorithm. The time step is fixed at $\Delta t = 0.1$, and all three algorithms share the same initializations and configurations.

The LQR setup is $A = -1.0, B = 0.5, \sigma = 0.1, R = -1, Q = -1, \beta = 3$.

For each value of B, data is collected from $\lfloor B/4 \rfloor$ trajectories at time steps $0, \Delta t, 2\Delta t$, and $3\Delta t$. We conduct 30 experiments for each B, compute the mean, and record the maximum and minimum of $\|V - V^*\|$ across these runs, measured in $L^2[-3, 3]$, where V^* is the value function corresponding to the ground truth optimal policy π^*, and V is the estimated value function obtained from the algorithms. The results (log-log graph) are presented in Figure 9c.

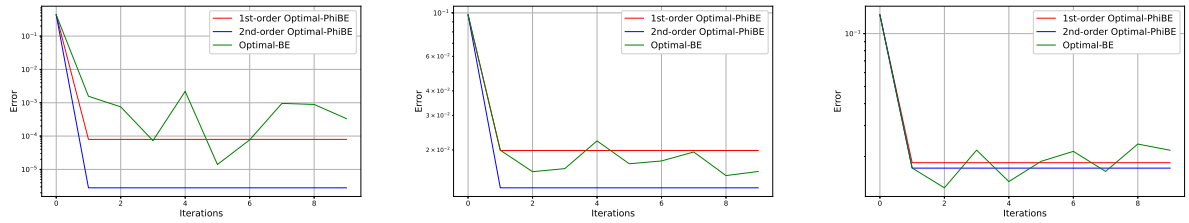
PI based on Optimal-PhiBE (both first- and second-order) demonstrates clear improvement as the number of data points increases, with the second-order variant benefiting even more significantly. In contrast, PI based on BE does not show the same level of improvement. While the PI based on second-order Optimal-PhiBE generally has higher variance, given a sufficient amount of data, it consistently outperforms the PI based on Optimal-BE approach and PI based on first-order Optimal-PhiBE.

6.2. **Merton's Portfolio Optimization Problem.** In this section, we compare the performance of PI based on Optimal-PhiBE and PI based on Optimal-BE on Merton's Portfolio Optimization Problem. Merton's portfolio problem [22] is one of the fundamental models in continuous-time finance, providing a framework for optimal asset allocation using a stochastic differential equation.

An investor allocates wealth between a risk-free asset E_t with interest rate r and a risky asset F_t with return μ and volatility σ. Their dynamics are,

$$dE_t = rE_t\,dt, \quad dF_t = \mu F_t\,dt + \sigma F_t\,dB_t,$$

where B_t is standard Brownian motion. The investor chooses a fraction π_t of wealth to invest in the risky asset. Borrowing is allowed at rate $r_b > r$, but to ensure well-posed dynamics, π_t is constrained to either



(A) Case 1 (B) Case 2 (C) Case 3

FIGURE 10. Comparison in Merton's Portfolio Optimization Problem

$[0, 1]$ or $(1, \infty)$ for all t. The wealth process follows,

$$dW_t = \begin{cases} (\pi_t \mu + (1 - \pi_t) r) W_t \, dt + \sigma \pi_t W_t \, dB_t, & \text{for } \pi_t \in [0, 1], \\ (\pi_t \mu - (\pi_t - 1) r_b) \, W_t \, dt + \sigma \pi_t W_t \, dB_t, & \text{for } \pi_t > 1. \end{cases}$$

The goal is to choose π_t to maximize the expected discounted utility,

$$\arg\max_{\pi_t} \mathbb{E}\left[\int_0^\infty e^{-\beta t} U(W_t) \, dt\right],$$

where

$$U(W) = \frac{W^{1-\gamma}}{1-\gamma}$$

is a power utility function with $\gamma > 0$ representing the risk aversion of the investor.

It turns out that the optimal allocation is a constant depending on the problem

$$\pi_t^* = \begin{cases} \frac{\mu - r}{\gamma \sigma^2}, & \text{if } \frac{\mu - r}{\gamma \sigma^2} \leq 1, \\ \max\left\{1, \frac{\mu - r_b}{\gamma \sigma^2}\right\}, & \text{if } \frac{\mu - r}{\gamma \sigma^2} > 1. \end{cases}$$

We evaluate the performance of the algorithms across three different settings. In each case, we fix $\gamma = 0.5$, which indicates a risk-tolerant investor. Additionally, we set $\Delta t = 1/12$ to simulate monthly data collection. The specific configurations for each setting are as follows:

- (Case 1: high risk premium) $r = 0.02, r_b = 0.05, \mu = 0.08, \sigma = 0.2, \beta = 0.2$.
- (Case 2: moderate returns) $r = 0.02, r_b = 0.05, \mu = 0.06, \sigma = 0.3, \beta = 0.15$.
- (Case 3: borderline leverage) $r = 0.02, r_b = 0.05, \mu = 0.07, \sigma = 0.3, \beta = 0.2$.

The ground truth optimal policies and corresponding value functions for each case are as follows:

- (Case 1) Optimal policy: $\pi_t^* = 1.5$; optimal value function: $V^*(s) = 12.2137\sqrt{s}$.
- (Case 2) Optimal policy: $\pi_t^* = 0.8888$; optimal value function: $V^*(s) = 15.2542\sqrt{s}$.
- (Case 3) Optimal policy: $\pi_t^* = 1.0$; optimal value function: $V^*(s) = 11.3475\sqrt{s}$.

For both algorithms, we maintain identical configurations throughout the experiments. In each iteration, we collect 10^7 data points from $\lfloor 10^7/6 \rfloor$ trajectories. Data is recorded at the times $0, \Delta t, 1\Delta t, 2\Delta t, 3\Delta t, 4\Delta t$, and $5\Delta t$, effectively gathering data over the span of half a year. The basis for the value function V is $\{\sqrt{s}\}$, while the basis for Q is $\{\sqrt{s}, \sqrt{s}a, \sqrt{s}a^2\}$. Each algorithm is run for 10 iterations, and the $L^2([-3, 3])$ distance between the estimated value function of the iterated policy and the ground-truth optimal value function is recorded. The results are presented in Figure 10, where the x-axis denotes the number of iterations, and the y-axis shows the corresponding $L^2([-3, 3])$ distance.

We observe that both PI based on Optimal-PhiBE and PI based on Optimal-BE achieve small L^2 errors, confirming that both methods can approximate the optimal value function effectively. However, PI based on Optimal-PhiBE consistently achieves lower errors across all cases and exhibits greater numerical stability throughout the iterations.

7. Proofs

7.1. **Proof of Lemma 3.5.** The proof of Lemma 3.5 is based on the following two lemmas.

Lemma 7.1. *Define operator* $\Pi_{i,\Delta t}f(s) = \frac{1}{\Delta t}\mathbb{E}[\sum_{j=1}^{i} a_j^{(i)} f(s_{j\Delta t} - s_0)|s_0 = s, a_\tau = a \text{ for } \tau \in [0, i\Delta t)]$ *with* $a_j^{(i)}$ *defined in* (21) *and* $f(s)|_{s=0} = 0$*, then*

$$\Pi_{i,\Delta t}f(s) = \mathcal{L}_{b(s',a),\Sigma(s',a)}f(s'-s)|_{s'=s} + \frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j^{(i)} \int_0^{j\Delta t} \mathbb{E}[\mathcal{L}_{b,\Sigma}^{i+1} f(s_t - s_0)|s_0 = s, a_\tau = a \text{ for } \tau \in [0, i\Delta t)]t^i dt.$$

Proof. First note that

$$\Pi_{i,\Delta t}f(s) = \frac{1}{\Delta t}\sum_{j=1}^{i} a_j^{(i)} \int_{\mathbb{S}} f(s'-s)\rho(s', j\Delta t|s)ds', \tag{49}$$

where $\rho(s', t|s), 0 \le t \le i\Delta t$ is the solution to the following PDE

$$\begin{cases} \partial_t \rho(s', t|s) = \nabla_{s'} \cdot \left[-b(s', a)\rho(s', t|s) + \frac{1}{2}\nabla_{s'} \cdot [\Sigma(s', a)\rho(s', t|s)]\right] \\ \rho(s', 0|s) = \delta_s(s') \end{cases} \tag{50}$$

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 (49) 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 $a_j^{(i)}$, one has $\sum_{j=0}^{i} a_j^{(i)} j^k = \sum_{j=1}^{i} a_j^{(i)} j^k = \begin{cases} 1, k = 1 \\ 0, k \ge 2. \end{cases}$. The first part can be simplified to

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

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

$$II = \frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j^{(i)} \int_0^{j\Delta t} \mathbb{E}[\mathcal{L}_{b,\Sigma}^{i+1} f(s_t - s_0)|s_0 = s, a_\tau = a \text{ for } \tau \in [0, i\Delta t)]t^i dt,$$

which completes the proof. □

Lemma 7.2. *For* $p(t, s, a) = \mathbb{E}\left[f(s_t, a)|s_0 = s, a_\tau = a, \tau \in [0, T)\right]$ *with* $0 \le t < T$ *and* s_t *driven by the SDE* (1)*, one has*

$$\|\nabla_s p(t, s, a)\|_\infty \le e^{ct}\|\nabla_s f(s, a)\|_\infty, \quad \text{with } c = \|\nabla_s b\|_\infty + \frac{1}{2}\|\nabla_s \sigma\|_\infty^2$$

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

$$\|p(s,t)\|_\infty \le \|b\|_\infty \sqrt{e^t-1}, \quad \|\nabla p(s,t)\|_\infty \le \|\nabla b(s)\|_\infty \sqrt{e^t-1}.$$

Proof. Note that $p(t,s,a)$ satisfies the following backward Kolmogorov equation [30],

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

let $q_l=\partial_{s_l} p$, one has

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

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

$$\begin{aligned}
\partial_t\left(\frac{1}{2}\|q\|_2^2\right) =& \mathcal{L}_{b,\Sigma}\left(\frac{1}{2}\|q\|_2^2\right) \underbrace{-\frac{1}{2}\sum_l (\nabla q_l)^\top \Sigma(\nabla q_l)}_{I} + q^\top \nabla b\cdot q + \underbrace{\sum_l \frac{1}{2}(\partial_{s_l}\Sigma : \nabla q) q_l}_{II},\\
\le& \mathcal{L}_{b,\Sigma}\left(\frac{1}{2}\|q\|_2^2\right) + \left(\|\nabla b\|_\infty + \frac{1}{2}\|\nabla\sigma\|_\infty^2\right)\|q\|_2^2,\\
\partial_t\left(\frac{1}{2}e^{-ct}\|q\|_2^2\right) \le& \mathcal{L}_{b,\Sigma}\left(\frac{1}{2}e^{-ct}\|q\|_2^2\right), \quad \text{with } c=\|\nabla b\|_\infty + \frac{1}{2}\|\nabla\sigma\|_\infty^2
\end{aligned}$$

where the first inequality is due to the following,

$$\begin{aligned}
2I =& -\sum_l (\sigma^\top \nabla q_l)^\top (\sigma^\top \nabla q_l) = -\sum_{k,l} (\sigma_{\cdot k}\cdot \nabla q_l)^2,\\
2II =& \sum_l (\partial_{s_l}\Sigma : \nabla q) q_l = \sum_l \left[(\sigma\partial_{s_l}\sigma^\top) : \nabla^2 p + (\partial_{s_l}\sigma\sigma^\top) : \nabla^2 p\right] q_l = \sum_l \left[(\sigma\partial_{s_l}\sigma^\top) : \nabla^2 p + (\sigma\partial_{s_l}\sigma^\top)^\top : (\nabla^2 p)^\top\right] q_l\\
=& 2\sum_l \left[(\sigma\partial_{s_l}\sigma^\top) : \nabla^2 p\right] q_l = 2\left[\sum_{l,i,j,k} \sigma_{ik}\partial_{s_l}\sigma_{jk}\partial_{s_i s_j} p\right] q_l = 2\sum_{j,k}(\sigma_{\cdot k}\cdot\nabla q_j)\left(\sum_l \partial_{s_l}\sigma_{jk} q_l\right)\\
\le& \sum_{j,k}(\sigma_{\cdot k}\cdot\nabla q_j)^2 + \sum_{j,k}\left(\sum_l \partial_{s_l}\sigma_{jk} q_l\right)^2 \le \sum_{j,k}(\sigma_{\cdot k}\cdot\nabla q_j)^2 + \left(\sum_{j,k}\|\nabla\sigma_{jk}\|^2\right)\|q\|^2,
\end{aligned}$$

which leads to

$$2(I+II) \le \|\nabla\sigma\|_2^2 \|q\|^2, \quad \text{where } \|\nabla\sigma\|_\infty^2 = \sup_{s\in\mathbb{S}}\|\nabla\sigma\|_2^2, \quad \|\nabla\sigma\|_2^2 = \sum_{j,k,l}(\partial_{s_l}\sigma_{jk})^2.$$

In addition,

$$q^\top\nabla b q = \sum_{k,l}(\partial_{s_l} b_k) q_k q_l \le \sqrt{\sum_{k,l}(\partial_{s_l} b_k)^2}\sqrt{\sum_{k,l}(q_k q_l)^2} \le \|\nabla b\|_\infty \|q\|^2, \quad \text{where } \|\nabla b\|_\infty = \sup_{s\in\mathbb{S}}\|\nabla b\|_2.$$

Let $g(t,s,a)=\frac{1}{2}e^{-ct}\|q(t,s,a)\|_2^2$, then $\partial_t g \le \mathcal{L}_{b,\Sigma} g$. Let $g_1(t,s,a)=\mathbb{E}[\|\nabla_s f(s_t,a)\|_2^2/2|s_0=s,a_\tau=a,\tau\in[0,T)]$, then $g_1(t,s,a)$ satisfies $\partial_t g_1=\mathcal{L}_{b,\Sigma} g_1$, with $g_1(0,s,a)=g(0,s,a)$. Since $\|g_1(t,s,a)\|_\infty \le \frac{1}{2}\|\nabla_s f\|_\infty$, by comparison theorem, one has

$$g(t,s,a), g_1(t,s,a) \le \frac{1}{2}\|\nabla_s f\|_\infty^2.$$

which yields,

$$\|q(t,s,a)\|_2^2 \le e^{ct}\|\nabla_s f(s,a)\|_\infty^2, \quad \text{where } \|\nabla_s f(s,a)\|_\infty^2 = \sup_{s,a}\nabla_s f(s,a).$$

This completes the proof for the first inequality.

For the second $p(t,s,a)=\mathbb{E}[f(s_t)(s_t-s_0)|s_0=s,a_\tau=a,\tau\in[0,T)]$, first note that it satisfies the following PDE,

$$\partial_t p = \mathcal{L}_{b,\Sigma} p + b(s), \quad \text{with } p(0,s,a)=0. \tag{51}$$

Multiplying it with $p^\top$ gives,

$$\partial_t\left(\frac{1}{2}\|p\|_2^2\right) \le \mathcal{L}_{b,\Sigma}\left(\frac{1}{2}\|p\|_2^2\right) + \frac{1}{2}\|b\|_\infty^2 + \frac{1}{2}\|p\|_2^2, \quad \text{for } \forall a > 0. \tag{52}$$

Let $g(t,s,a) = \frac{1}{2}\|p\|_2^2 e^{-t} + \frac{1}{2}\|b\|_\infty^2 e^{-t}$, then one has $\partial_t g \le \mathcal{L}_{b,\Sigma} g$, with $g(0,s) = \frac{\|b\|_\infty^2}{2}$. Similarly, by comparison theorem, one has $\|g(t,s,a)\|_\infty \le \frac{\|b\|_\infty^2}{2}$, which implies,

$$\|p(t,\cdot)\|_\infty \le \|b\|_\infty \sqrt{e^t - 1}.$$

On the other hand, taking ∇ to (51) and multiply $q^\top$ to it gives,

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

Let $g(t,s,a) = \frac{1}{2}\|q\|_2^2 e^{-ct} + \frac{1}{2}\|\nabla b\|_\infty^2 e^{-ct}$ with $c = 2\|\nabla b\|_\infty + \|\nabla\sigma\|_\infty^2 + 1$, then

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

By the comparison theorem, one has $g(t,s,a) \le \frac{1}{2}\|\nabla_s b\|_\infty^2$, which implies

$$\|q\|_2 \le \|\nabla b\|_\infty \sqrt{e^t - 1}.$$

□

Now we are ready to prove Lemma 3.5

Proof. First note that $\hat{b}_i(s) = \Pi_{i,\Delta t} f(s)$ with $f(s) = s$. Therefore, By Lemma 7.1, one has,

$$\begin{aligned}
\hat{b}_i(s,a) =& b(s,a) + \frac{1}{\Delta t i!}\sum_{j=1}^i a_j^{(i)} \int_0^{j\Delta t} \mathbb{E}[\mathcal{L}_{b,\Sigma}^{i+1} f(s_t - s_0) | s_0 = s, a_\tau = a, \tau \in [0, i\Delta t)] t^i dt. \\
\le& b(s,a) + \frac{1}{\Delta t i!}\sum_{j=1}^i \left|a_j^{(i)}\right| \left\|\mathcal{L}_{b,\Sigma}^{i+1} f(s)\right\|_\infty \int_0^{j\Delta t} t^i dt \\
=& b(s,a) + \hat{C}_i \left\|\mathcal{L}_{b,\Sigma}^i b\right\|_\infty \Delta t^i,
\end{aligned}$$

where

$$\hat{C}_i = \frac{\sum_{j=1}^i j^{i+1}\left|a_j^{(i)}\right|}{(i+1)!}, \tag{53}$$

which gives (25). The uniform boundedness of $\hat{b}(s)$ is followed by the above inequality and the uniform boundedness of $b(s)$.

Next, we prove the uniform boundedness of $\nabla_s \hat{b}$. Note that

$$\nabla_s \hat{b}_i(s,a) = \nabla_s b(s,a) + \frac{1}{\Delta t i!}\sum_{j=1}^i a_j^{(i)} \int_0^{j\Delta t} \nabla_s p(t,s,a) t^i dt.$$

where

$$p(t,s,a) = \mathbb{E}[\mathcal{L}_{b,\Sigma}^{i+1} f(s_t - s_0) | s_0 = s, a_\tau = a, \tau \in [0, i\Delta t)] = \mathbb{E}[\mathcal{L}_{b,\Sigma}^i b(s_t, a) | s_0 = s, a_\tau = a, \tau \in [0, i\Delta t)].$$

By the first inequality of Lemma 7.2, one has $\|\nabla_s p(t,s,a)\| \le e^{cT}\left\|\nabla_s \mathcal{L}_{b,\Sigma}^i b(s,a)\right\|_\infty$ for $T = i\Delta t$, which is uniformly bounded with Assumption 2/(a), (b). Therefore, the uniform boundedness of $\left\|\nabla_s \hat{b}_i(s,a)\right\|_\infty$ is followed by the above inequality and the uniform boundedness of $\|\nabla_s b(s,a)\|_\infty$.

Next, one notes that $\hat{\Sigma}_i = \Pi_{i,\Delta t} f(s)$ with $f(s) = s^\top s$. Therefore, one has

$$\begin{aligned}\hat{\Sigma}_i(s) =& \Sigma(s) \\ &+ \frac{1}{\Delta t i!}\sum_{j=1}^{i} a_j^{(i)} \int_0^{j\Delta t} \mathbb{E}[\mathcal{L}_{b,\Sigma}^i \left(b(s_t)(s_t-s_0)^\top + (s_t-s_0)b^\top(s_t) + \Sigma(s_t)\right)|s_0 = s] t^i dt.\end{aligned} \tag{54}$$

Note that

$$\begin{aligned}h_i(s_t) :=& \mathcal{L}_{b,\Sigma}^i \left(b(s_t)(s_t-s_0)^\top + (s_t-s_0)b^\top(s_t)\right) \\ &- \left[\mathcal{L}_{b,\Sigma}^i b(s_t)(s_t-s_0)^\top + (s_t-s_0)(\mathcal{L}_{b,\Sigma}^i b(s_t))^\top\right]\end{aligned} \tag{55}$$

is a function that only depends on the derivative $\nabla^j\Sigma, \nabla^j b$ up to $2i$-th order, which can be bounded under Assumption 2/(c). Thus applying the second inequality of Lemma 7.2 yields

$$\begin{aligned}&\left|\mathbb{E}[\mathcal{L}_{b,\Sigma}^i \left(b(s_t)(s_t-s_0)^\top + (s_t-s_0)b^\top(s_t) + \Sigma(s_t)\right)|s_0 = s, a_\tau = a, \tau \in [0, i\Delta t)]\right| \\ \le& \left|\mathcal{L}_{b,\Sigma}^i \Sigma\right| + |h(s)| + \left|\mathbb{E}\left[\mathcal{L}_{b,\Sigma}^i b(s_t)(s_t-s_0)^\top + (s_t-s_0)(\mathcal{L}_{b,\Sigma}^i b(s_t))^\top | s_0 = s, a_\tau = a, \tau \in [0, i\Delta t)\right]\right| \\ \le& \left\|\mathcal{L}_{b,\Sigma}^i \Sigma\right\|_\infty + \|h(s)\|_\infty + 2\|b(s)\|_\infty \sqrt{e^t - 1}.\end{aligned}$$

Hence, one has,

$$\begin{aligned}&\int_0^{j\Delta t} \mathbb{E}[\mathcal{L}_{b,\Sigma}^i \left(b(s_t,a)(s_t-s_0)^\top + (s_t-s_0)b^\top(s_t,a) + \Sigma(s_t)\right)|s_0 = s, a_\tau = a, \tau \in [0, i\Delta t)] t^i dt \\ \le& \frac{1}{i+1}\left(\left\|\mathcal{L}_{b,\Sigma}^i \Sigma\right\|_\infty + \|h\|_\infty\right)(j\Delta t)^{i+1} + 3\|b(s)\|_\infty \int_0^{j\Delta t} t^{i+1/2} dt, \quad \text{for } j\Delta t \le 1 \\ \le& \frac{1}{i+1}\left(\|h_i\|_\infty + \left\|\mathcal{L}_{b,\Sigma}^i \Sigma\right\|_\infty + 3\|b\|_\infty\right)(j\Delta t)^{i+1}\end{aligned}$$

where $\sqrt{e^t - 1} \le \frac{3}{2}\sqrt{t}$ for $t \le 1$ are used in the first inequality, and $\frac{(j\Delta t)^{i+3/2}}{i+3/2} \le \frac{(j\Delta t)^{i+1}}{i+1}$ for $j\Delta t \le 1$ is used in the second inequality. Plugging the above inequality back to (54) implies

$$\left\|\hat{\Sigma}_i(s,a) - \Sigma(s,a)\right\|_\infty \le \hat{C}_i(\left\|\mathcal{L}_{b,\Sigma}^i \Sigma\right\|_\infty + \|h_i(s)\|_\infty + 3\|b\|_\infty)\Delta t^i,$$

where $h_i(s)$ is defined in (55) that only depends on the derivative $\nabla^j\Sigma, \nabla^j b$ up to $2i$-th order. The uniform boundedness of $\hat{\Sigma}$ is followed by the above inequality and the uniform boundedness of $\Sigma(s,a)$.

To prove the the uniform boundedness of $\nabla_s\hat{\Sigma}_i(s,a)$, one needs to bound

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

where

$$p(s,t) = \mathbb{E}[\mathcal{L}_{b,\Sigma}^i \Sigma(s_t,a) + h(s_t,a) + \mathcal{L}_{b,\Sigma}^i b(s_t,a)(s_t-s_0)^\top + (s_t-s_0)(\mathcal{L}_{b,\Sigma}^i b(s_t,a))^\top | s_0 = s, a_\tau = a, \tau \in [0, i\Delta t)]$$

with $h(s,a)$ defined in (55). Therefore, by the first and third inequalities in Lemma 7.2, one has

$$\|\nabla p(t,s,a)\|_\infty \le e^{cT}\left(\left\|\nabla_s \mathcal{L}_{b,\Sigma}^i \Sigma\right\|_\infty + \|\nabla_s h_i\|_\infty\right) + \|\nabla b\|_\infty \sqrt{e^T - 1}, \quad \text{with } T = i\Delta t,$$

which is uniformly bounded by Assumption 2. Therefore, the uniformly boundedness of $\nabla\hat{\Sigma}_i(s,a)$ is followed by the uniformly boundedness of $\nabla\Sigma(s,a)$ and the above inequality.

□

7.2. **Proof of Lemma 3.6.**

Proof. Since $\pm\|r\|_\infty/\beta$ is a sub- and a super- solution to (5), respectively, the first claim follows from the comparison principle. Below, we will often write ∇h for $\nabla_s h$ where h can be r, b and σ.

Let $s^1, s^2 \in \mathbb{R}^d$, and suppose that π^1 is such that

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

with s_t^1 satisfying

$$ds_t^1 = b(s_t^1, \pi^1(s_t^1))dt + \sigma(s_t^1, \pi^1(s_t^1))dB_t, \quad s_0^1 = s^1.$$

Let s_t^2 solve the following SDE

$$ds_t^2 = b(s_t^2, \pi^1(s_t^1))dt + \sigma(s_t^2, \pi^1(s_t^1))dB_t, \quad s_0^2 = s^2.$$

Since π^1 might not be the optimal policy for $V^*(s^2)$, we have

$$V^*(s^1) - V^*(s^2) \geq \mathbb{E}\left[\int_0^\infty e^{-\beta t}(r(s_t^1, \pi^1(s_t^1)) - r(s_t^2, \pi^1(s_t^1)))dt\right]. \tag{56}$$

Below, we will estimate $\mathbb{E}[|s_t^1 - s_t^2|^2]$. It follows from the SDEs that

$$\begin{aligned} |s_t^1 - s_t^2| &\leq |s_0^1 - s_0^2| + \int_0^t |b(s_\tau^1, \pi^1(s_\tau^1)) - b(s_\tau^2, \pi^1(s_\tau^1))|d\tau + \left|\int_0^t (\sigma(s_\tau^1, \pi^1(s_\tau^1)) - \sigma(s_\tau^2, \pi^1(s_\tau^1)))dB_\tau\right| \\ &\leq |s_0^1 - s_0^2| + \|\nabla b\|_\infty \int_0^t |s_\tau^1 - s_\tau^2|d\tau + \left|\int_0^t (\sigma(s_\tau^1, \pi^1(s_\tau^1)) - \sigma(s_\tau^2, \pi^1(s_\tau^1)))dB_\tau\right| \end{aligned} \tag{57}$$

where we used the initial data and that b is uniformly Lipschitz continuous in s.

Let $t_0 > 0$, and let $t \in [0, t_0]$. By the Burkholder-Davis-Gundy inequality (see [33, Chapter IV]), there exists a dimensional constant C_d such that

$$\begin{aligned} &\mathbb{E}\left[\sup_{t\in[0,t_0]}\left|\int_0^t (\sigma(s_\tau^1, \pi^1(s_\tau^1)) - \sigma(s_\tau^2, \pi^1(s_\tau^1)))dB_\tau\right|\right] \\ &\quad \leq C_d\,\mathbb{E}\left[\left(\int_0^{t_0} |\sigma(s_\tau^1, \pi^1(s_\tau^1)) - \sigma(s_\tau^2, \pi^1(s_\tau^1))|^2 d\tau\right)^{1/2}\right] \leq C_d\|\nabla\sigma\|_\infty t_0^{\frac{1}{2}}\mathbb{E}\left[\sup_{t\in[0,t_0]}|s_t^1 - s_t^2|\right]. \end{aligned}$$

We obtain from (57) that

$$\mathbb{E}\left[\sup_{t\in[0,t_0]}|s_t^1 - s_t^2|\right] \leq |s_0^1 - s_0^2| + t_0\|\nabla b\|_\infty\mathbb{E}\left[\sup_{t\in[0,t_0]}|s_t^1 - s_t^2|\right] + t_0^{\frac{1}{2}}C_d\|\nabla\sigma\|_\infty\mathbb{E}\left[\sup_{t\in[0,t_0]}|s_t^1 - s_t^2|\right].$$

Thus, after taking t_0 to be small such that

$$t_0\|\nabla b\|_\infty + t_0^{1/2}C_d\|\nabla\sigma\|_\infty \leq \frac{1}{2},$$

we get

$$\mathbb{E}\left[\sup_{t\in[0,t_0]}|s_t^1 - s_t^2|\right] \leq 2|s_0^1 - s_0^2|.$$

After iteration, this yields for all $t > 0$,

$$\mathbb{E}\left[|s_t^1 - s_t^2|\right] \leq 2^{t/t_0+1}|s_0^1 - s_0^2|. \tag{58}$$

It follows from (56), (58) and the uniform Lipchitz continuity of $r(s, a)$ in s that

$$\begin{aligned} V^*(s^1) - V^*(s^2) &\geq -\|\nabla r\|_\infty \int_0^\infty e^{-\beta t}\mathbb{E}|s_t^1 - s_t^2|dt \\ &\geq -2\|\nabla r\|_\infty|s_0^1 - s_0^2|\int_0^\infty e^{-\beta t}2^{t/t_0}dt \geq -\frac{2\|\nabla r\|_\infty}{\beta - (\ln 2)/t_0}|s_0^1 - s_0^2|, \end{aligned}$$

which yields the second claim.

For the last claim, let s_t^1, s_t^2 be as before. Since σ is independent of s_τ^1 in (57), we can replace (57) by

$$|s_t^1 - s_t^2| \leq |s_0^1 - s_0^2| + \|\nabla b\|_\infty \int_0^t |s_\tau^1 - s_\tau^2|d\tau.$$

By Grönwall's inequality, we get

$$|s_t^1 - s_t^2| \leq |s_0^1 - s_0^2|e^{\|\nabla b\|_\infty t}.$$

Since $\beta - \|\nabla b\|_\infty > 0$, it follows that

$$V^*(s) - V^*(y) \geq -\int_0^\infty e^{-\beta t}\|\nabla r\|_\infty |s_0^1 - s_0^2| e^{\|\nabla b\|_\infty t} dt \geq -\frac{\|\nabla r\|_\infty}{\beta - \|\nabla b\|_\infty}|s_0^1 - s_0^2|,$$

which implies the last claim. □

7.3. **Proof of Lemma 3.7.**

Proof. Let us only prove the upper bound for $V^*(s) - \hat{V}(s)$; the proof for the other direction is almost identical. By Lemma 3.6, we have

$$|V^*(\cdot)|, |\hat{V}(\cdot)| \leq C_1 := \max\{\|r\|_\infty, \|\hat{r}\|_\infty\}/\beta. \tag{59}$$

Let

$$2\delta := \sup_{s\in\mathbb{R}^d}(V^*(s) - \hat{V}(s)),$$

and apparently $\delta \leq C_1$. Let R_1 be large enough such that

$$\sup_{s\in B_{R_1}}(V^*(s) - \hat{V}(s)) \geq \delta. \tag{60}$$

Let $R_2 \geq 2R_1$ to be determined. We take a smooth, radially symmetric, and radially non-decreasing function $\phi : \mathbb{R}^d \to [0,\infty)$ such that

$$\phi(\cdot) \equiv 0 \text{ in } B_{R_1}, \quad \phi(\cdot) = C_1 \text{ outside } B_{R_2}, \tag{61}$$

and

$$|\nabla\phi(s)| \leq 4C_1/R_2, \qquad |\nabla^2\phi(s)| \leq C_1/R_2 \quad \text{for all } s. \tag{62}$$

Such ϕ clearly exists when R_2 is sufficiently large.

It follows from (59), (60) and (61) that there exists $s^0 \in B_{R_2}$ such that

$$V^*(s^0) - \hat{V}(s^0) - 2\phi(s^0) = \sup_{s\in\mathbb{R}^d}\Big(V^*(s) - \hat{V}(s) - 2\phi(s)\Big) =: \delta' \geq \delta. \tag{63}$$

Next, for some $\rho \gg 1$ to be determined, there are $s^1, s^2 \in B_{R_2}$ such that

$$\begin{aligned} &V^*(s^1) - \hat{V}(s^2) - \phi(s^1) - \phi(s^2) - \rho|s^1 - s^2|^2 \\ &\qquad = \sup_{s,s'\in\mathbb{R}^d}\Big(V^*(s) - \hat{V}(s') - \phi(s) - \phi(s') - \rho|s - s'|^2\Big) \\ &\qquad \geq V^*(s^0) - \hat{V}(s^0) - 2\phi(s^0) = \delta'. \end{aligned} \tag{64}$$

Since $\phi \geq 0$, this implies

$$V^*(s^1) - \hat{V}(s^2) \geq \delta'. \tag{65}$$

In view of (62), we have

$$|\phi(s^1) - \phi(s^2)| \leq 4C_1|s^1 - s^2|/R_2.$$

Using this, Lipschitz continuity of V^*, and (64), we obtain

$$\begin{aligned} \delta' &\leq V^*(s^1) - \hat{V}(s^2) - 2\phi(s^2) - \rho|s^1 - s^2|^2 + 4C_1|s^1 - s^2|/R_2 \\ &\leq V^*(s^2) - \hat{V}(s^2) - 2\phi(s^2) + L|s^1 - s^2| - \rho|s^1 - s^2|^2 + 4C_1|s^1 - s^2|/R_2 \\ &\leq \delta' + L|s^1 - s^2| - \rho|s^1 - s^2|^2 + 4C_1|s^1 - s^2|/R_2, \end{aligned}$$

where in the last inequality, we also used (63). This then simplifies to

$$|s^1 - s^2| \leq (L + 4C_1/R_2)/\rho. \tag{66}$$

If only $\hat{V}$ is known to be Lipschitz, similarly, we have

$$\begin{aligned} \delta' &\leq V^*(s^1) - \hat{V}(s^1) - 2\phi(s^1) + L|s^1 - s^2| - \rho|s^1 - s^2|^2 + 4C_1|s^1 - s^2|/R_2 \\ &\leq \delta' + L|s^1 - s^2| - \rho|s^1 - s^2|^2 + 4C_1|s^1 - s^2|/R_2. \end{aligned}$$

We comment that in the proof, we only need one of V^* and $\hat{V}$ to be Lipschitz continuous.

Now we proceed by making use of (64). Since V^* and $\hat{V}$ are, respectively, solutions to HJB equation (5) and (18), the Crandall-Ishii lemma [7, Theorem 3.2] yields that there are matrices X_1, $X_2 \in \mathcal{S}^d$ satisfying the following:

$$-(2\rho + |J|)I \leq \begin{pmatrix} X_1 & 0 \\ 0 & -X_2 \end{pmatrix} \leq J + \frac{1}{2\rho}J^2, \quad \text{with } J := 2\rho \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}, \tag{67}$$

and

$$\begin{aligned} \beta V^*(s^1) - \max_a \left[r(s^1,a) + b(s^1,a)\cdot p_1 + \frac{1}{2}\operatorname{Tr}(\Sigma(s^1,a)(X_1 + \nabla^2\phi(s^1))) \right] \leq 0 \\ \leq \beta \hat{V}(s^2) - \max_a \left[\hat{r}(s^2,a) + \hat{b}(s^2,a)\cdot p_2 + \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,a)(X_2 - \nabla^2\phi(s^2))) \right], \end{aligned} \tag{68}$$

where

$$p_1 := 2\rho(s^1 - s^2) + \nabla\phi(s^1), \quad p_2 := 2\rho(s^1 - s^2) - \nabla\phi(s^2). \tag{69}$$

For $i = 1, 2$, let a_i be one argmax of

$$\left[r(s^i,a) + b(s^i,a)\cdot p_i + \frac{1}{2}\operatorname{Tr}(\Sigma(s^i,a)X_i) \right].$$

and then by the assumptions,

$$\begin{aligned} -\max_a \left[\hat{r}(s^2,a) + \hat{b}(s^2,a)\cdot p_2 + \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,a)X_2) \right] &\leq -\left[\hat{r}(s^2,a_1) + \hat{b}(s^2,a_1)\cdot p_2 + \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,a_1)X_2) \right] \\ &\leq -\left[r(s^2,a_1) + b(s^2,a_1)\cdot p_2 + \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,a_1)X_2) \right] + \varepsilon_r + \varepsilon_b|p_2|. \end{aligned}$$

Thus, also using the regularity of r and b,

$$\begin{aligned} &\max_a \left[r(s^1,a) + b(s^1,a)\cdot p_1 + \frac{1}{2}\operatorname{Tr}(\Sigma(s^1,a)(X_1 + \nabla^2\phi(s^1))) \right] \\ &\qquad - \max_a \left[\hat{r}(s^2,a) + \hat{b}(s^2,a)\cdot p_2 + \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,a)(X_2 - \nabla^2\phi(s^2))) \right] \\ &\leq \left[r(s^1,a_1) + b(s^1,a_1)\cdot p_1 + \frac{1}{2}\operatorname{Tr}(\Sigma(s^1,a_1)(X_1 + \nabla^2\phi(s^1))) \right] \\ &\qquad - \left[r(s^2,a_1) + b(s^2,a_1)\cdot p_2 + \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,a_1)(X_2 - \nabla^2\phi(s^2))) \right] + \varepsilon_r + \varepsilon_b|p_2| \\ &\leq \|\nabla r\|_\infty |s^1 - s^2| + \|\nabla b\|_\infty |p_2||s^1 - s^2| + \|b\|_\infty |p_1 - p_2| + \frac{1}{2}\operatorname{Tr}(\Sigma(s^1,a_1)X_1) \\ &\qquad - \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,a_1)X_2) + \frac{1}{2}\operatorname{Tr}(\Sigma(s^1,a_1)\nabla^2\phi(s^1)) + \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,a_2)\nabla^2\phi(s^2)) + \varepsilon_r + \varepsilon_b|p_2|. \end{aligned} \tag{70}$$

Note that $J + \frac{1}{2\rho}J^2 = 6\rho J$ and (67) yields $X_1 \leq X_2$. Similarly as done in [7, Example 3.6], we multiply (67) by the nonnegative symmetric matrix

$$\begin{pmatrix} \sigma(s^1,a_1)\sigma(s^1,a_1)^T & \sigma(s^2,a_1)\sigma(s^1,a_1)^T \\ \sigma(s^1,a_1)\sigma(s^2,a_1)^T & \sigma(s^2,a_1)\sigma(s^2,a_1)^T \end{pmatrix}$$

on the left-hand side, and take traces to obtain

$$\begin{aligned} \frac{1}{2}\operatorname{Tr}(\Sigma(s^1,a_1)X_1) - \frac{1}{2}\operatorname{Tr}(\Sigma(s^2,a_1)X_2) &\leq 3\rho \operatorname{Tr}\left[(\sigma(s^1,a_1) - \sigma(s^2,a_1))(\sigma(s^1,a_1) - \sigma(s^2,a_1))^T \right] \\ &\leq 3\rho \|\sigma(s^1,a_1) - \sigma(s^2,a_1)\|_F^2 \leq 3\rho d \|\nabla\sigma\|_\infty^2 |s^1 - s^2|^2 \end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, and we used that

$$\sqrt{\operatorname{Tr}(AA^T)} = \|A\|_F \leq \sqrt{d}\|A\|_2 \quad \text{for any } m \times n \text{ matrix } A \text{ with } m, n \leq d.$$

Since $|J| = 4\rho$, (67) yields $|X_2| \le 6\rho$. Thus, by the assumption that $\|\Sigma - \hat{\Sigma}\|_\infty \le \varepsilon_\Sigma$, we get

$$\begin{aligned}\frac{1}{2}\operatorname{Tr}(\Sigma(s^1,a_1)X_1) - \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,a_1)X_2) &\le \frac{1}{2}\operatorname{Tr}(\Sigma(s^1,a_1)X_1) - \frac{1}{2}\operatorname{Tr}(\Sigma(s^2,a_1)X_2) + \frac{d}{2}\|\Sigma - \hat{\Sigma}\|_\infty |X_2| \\ &\le 3\rho d\|\nabla\sigma\|_\infty^2 |s^1 - s^2|^2 + 3\rho d\varepsilon_\Sigma\end{aligned}$$

Next, by (62) and the assumptions, there exists $C > 0$ such that

$$\frac{1}{2}\operatorname{Tr}(\Sigma(s^1,a_1)\nabla^2\phi(s^1)) + \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,a_2)\nabla^2\phi(s^2)) \le C/R_2. \tag{71}$$

Also using (66) and (69), we obtain

$$\begin{aligned}&|p_1|,\ |p_2| \le 2(L + 4C_1/R_2) + 4C_1/R_2 \le 2L + 12C_1/R_2, \\ &|p_1 - p_2| = |\nabla\phi(s^1) + \nabla\phi(s^2)| \le 8C_1/R_2.\end{aligned} \tag{72}$$

Now, we plugging the above estimates (62), (66), (70)–(72) into (68) to get

$$\begin{aligned}\beta(V^*(s^1) - \hat{V}(s^2)) &\le \|\nabla r\|_\infty |s^1 - s^2| + \|\nabla b\|_\infty |s^1 - s^2||p_2| + \|b\|_\infty |p_1 - p_2| + \varepsilon_r \\ &\quad + \varepsilon_b|p_2| + 3\rho d\varepsilon_\Sigma + 3\rho d\|\nabla\sigma\|_\infty^2 |s^1 - s^2|^2 + C/R_2 \\ &\le C_2/\rho + (C_3/\rho + \varepsilon_b)(2L + 12C_1/R_2) + \varepsilon_r + 3\rho d\varepsilon_\Sigma + C_4/\rho + (C + 8C_1\|b\|_\infty)/R_2\end{aligned}$$

where

$$C_2 := (L + 4C_1/R_2)\|\nabla r\|_\infty, \quad C_3 := (L + 4C_1/R_2)\|\nabla b\|_\infty, \quad C_4 := 3d(L + 4C_1/R_2)^2\|\nabla\sigma\|_\infty^2.$$

Using (65) and then passing R_2 to infinity yield

$$\beta\delta' \le C_5/\rho + 2L\varepsilon_b + \varepsilon_r + 3\rho d\varepsilon_\Sigma.$$

with

$$C_5 := L\|\nabla r\|_\infty + 2L^2\|\nabla b\|_\infty + 3dL^2\|\nabla\sigma\|_\infty^2.$$

We then take ρ to be $\sqrt{C_5/(3d\varepsilon_\Sigma)}$, so that

$$\sup_{s\in\mathbb{R}^d}(V^*(s) - \hat{V}(s)) = 2\delta \le 2\delta' \le \frac{2}{\beta}\left(2\sqrt{3dC_5\varepsilon_\Sigma} + 2L\varepsilon_b + \varepsilon_r\right).$$

□

7.4. **Proof of Theorem 3.8.**

Proof. The proof is very similar to the one of Lemma 3.7, except a few estimates due to the different equations. Let us denote $U := V^{\hat{\pi}_i^*}$ and $\hat{V} := \hat{V}_i^*$ for simplicity. In view of Lemma 3.7, it suffices to estimate the difference between U and $\hat{V}$.

As before, we consider a slightly more general case by letting $\hat{V}$ solve (5) with the coefficients r, b, and Σ replaced by $\hat{r}$, $\hat{b}$, and $\hat{\Sigma}$, respectively, and letting U solve

$$\beta U(s) = r(s,\hat{a}) + b(s,\hat{a})\cdot\nabla U(s) + \frac{1}{2}\operatorname{Tr}(\Sigma(s,\hat{a})\nabla^2 U(s)), \tag{73}$$

where $\hat{a}(s)$ is one argmax of

$$\max_a\left[\hat{r}(s,a) + \hat{b}(s,a)\cdot\nabla\hat{V}(s) + \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s,a)\nabla^2\hat{V}(s))\right].$$

We assume that $\varepsilon_r, \varepsilon_b$ and ε_Σ are such that

$$\sup_{s,a}|r(s,a) - \hat{r}(s,a)| \le \varepsilon_r, \quad \sup_{s,a}|b(s,a) - \hat{b}(s,a)| \le \varepsilon_b \quad \text{and} \quad \sup_{s,a}|\Sigma(s,a) - \hat{\Sigma}(s,a)| \le \varepsilon_\Sigma.$$

We also assume that at least one of U and $\hat{V}$ are uniformly Lipschitz continuous with Lipschitz constant L. It suffices to estimate $|U - \hat{V}|$.

Again we only show the upper bound for $\hat{V}(s) - U(s)$. It follows from Lemma 3.6 (the control set for U is a singleton), we have

$$|\hat{V}(\cdot)|,\ |U(\cdot)| \le C_1 := \max\{\|r\|_\infty, \|\hat{r}\|_\infty\}/\beta. \tag{74}$$

Set

$$2\delta := \sup_{s\in\mathbb{R}^d}(\hat{V}(s)-U(s)) \le 2C_1.$$

Let R_1 be large enough such that

$$\sup_{s\in B_{R_1}}(\hat{V}(s)-U(s)) \ge \delta. \tag{75}$$

For $R_2 \ge 2R_1$, take ϕ the same as in the proof of Lemma 3.7 so that (61) and (62) hold.

Due to (74), (75) and (61), there exists $s^0 \in B_{R_2}$ such that

$$\hat{V}(s^0)-U(s^0)-2\phi(s^0) = \sup_{s\in\mathbb{R}^d}\left(\hat{V}(s)-U(s)-2\phi(s)\right) =: \delta' \ge \delta. \tag{76}$$

For any fixed $\rho \gg 1$, there are $s^1, s^2 \in B_{R_2}$ such that

$$\begin{aligned}
&\hat{V}(s^1)-U(s^2)-\phi(s^1)-\phi(s^2)-\rho|s^1-s^2|^2 \\
&\qquad = \sup_{s,s'\in\mathbb{R}^d}\left(\hat{V}(s)-U(s')-\phi(s)-\phi(s')-\rho|s-s'|^2\right) \\
&\qquad \ge \hat{V}(s^0)-U(s^0)-2\phi(s^0) = \delta'.
\end{aligned} \tag{77}$$

By (62), we have

$$|\phi(s^1)-\phi(s^2)| \le 4C_1|s^1-s^2|/R_2. \tag{78}$$

If $\hat{V}$ is Lipschitz continuous, (76), (77) and (78) yield

$$\begin{aligned}
\delta' &\le \hat{V}(s^1)-U(s^2)-2\phi(s^2)-\rho|s^1-s^2|^2+4C_1|s^1-s^2|/R_2 \\
&\le \hat{V}(s^2)-U(s^2)-2\phi(s^2)+L|s^1-s^2|-\rho|s^1-s^2|^2+4C_1|s^1-s^2|/R_2 \\
&\le \delta'+L|s^1-s^2|-\rho|s^1-s^2|^2+4C_1|s^1-s^2|/R_2,
\end{aligned}$$

while if U is Lipschitz continuous, we get

$$\begin{aligned}
\delta' &\le \hat{V}(s^1)-U(s^1)-2\phi(s^1)+L|s^1-s^2|-\rho|s^1-s^2|^2+4C_1|s^1-s^2|/R_2 \\
&\le \delta'+L|s^1-s^2|-\rho|s^1-s^2|^2+4C_1|s^1-s^2|/R_2.
\end{aligned}$$

In both cases, we find

$$|s^1-s^2| \le (L+4C_1/R_2)/\rho. \tag{79}$$

Since $\hat{V}$ and U are, respectively, solutions to (5) (with the coefficients r, b and Σ replaced by $\hat{r}$, $\hat{b}$ and $\hat{\Sigma}$) and (73), the Crandall-Ishii lemma [7, Theorem 3.2] yields that there are matrices X_1, $X_2 \in \mathcal{S}^d$ satisfying the following:

$$-(2\rho+|J|)I \le \begin{pmatrix} X_1 & 0 \\ 0 & -X_2 \end{pmatrix} \le J+\frac{1}{2\rho}J^2, \quad \text{with } J := 2\rho\begin{pmatrix} I & -I \\ -I & I \end{pmatrix}, \tag{80}$$

and

$$\begin{aligned}
&\beta\hat{V}(s^1)-\left[\hat{r}(s^1,\hat{a})+\hat{b}(s^1,\hat{a})\cdot p_1+\frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^1,\hat{a})(X_1+\nabla^2\phi(s^1)))\right] \le 0 \\
&\qquad \le \beta U(s^2)-\left[r(s^2,\hat{a})+b(s^2,\hat{a})\cdot p_2+\frac{1}{2}\operatorname{Tr}(\Sigma(s^2,\hat{a})(X_2-\nabla^2\phi(s^2)))\right],
\end{aligned} \tag{81}$$

where

$$p_1 := 2\rho(s^1-s^2)+\nabla\phi(s^1), \quad p_2 := 2\rho(s^1-s^2)-\nabla\phi(s^2). \tag{82}$$

Then (72) holds the same. By the assumptions on $r, b, \hat{r}$ and $\hat{b}$, and (72), (79) and (62),

$$\begin{aligned}
&\left[\hat{r}(s^1,\hat{a}) + \hat{b}(s^1,\hat{a})\cdot p_2\right] - \left[r(s^2,\hat{a}) + b(s^2,\hat{a})\cdot p_1\right] \\
&\le \left[r(s^1,\hat{a}) + b(s^1,\hat{a})\cdot p_1\right] - \left[r(s^2,\hat{a}) + b(s^2,\hat{a})\cdot p_2\right] + \varepsilon_r + \varepsilon_b|p_1| \\
&\le \varepsilon_r + \varepsilon_b|p_1| + \|\nabla r\|_\infty |s^1 - s^2| + \|\nabla b\|_\infty |p_1||s^1 - s^2| + \|b\|_\infty |p_1 - p_2| \\
&\le \varepsilon_r + \varepsilon_b(2L + 12C_1/R_2) + \left(\|\nabla r\|_\infty + \|\nabla b\|_\infty(2L + 12C_1/R_2)\right)(L + 4C_1/R_2)/\rho + \|b\|_\infty 8C_1/R_2.
\end{aligned}$$

For the second order terms, as before, we multiply (80) by the nonnegative symmetric matrix

$$\begin{pmatrix} \sigma(s^1,\hat{a})\sigma(s^1,\hat{a})^T & \sigma(s^2,\hat{a})\sigma(s^1,\hat{a})^T \\ \sigma(s^1,\hat{a})\sigma(s^2,\hat{a})^T & \sigma(s^2,\hat{a})\sigma(s^2,\hat{a})^T \end{pmatrix}$$

and take traces to obtain

$$\begin{aligned}
\frac{1}{2}\operatorname{Tr}(\Sigma(s^1,\hat{a})X_1) - \frac{1}{2}\operatorname{Tr}(\Sigma(s^2,\hat{a})X_2) &\le 3\rho \operatorname{Tr}\left[(\sigma(s^1,\hat{a}) - \sigma(s^2,\hat{a}))(\sigma(s^1,\hat{a}) - \sigma(s^2,\hat{a}))^T\right] \\
&\le 3\rho\|\sigma(s^1,\hat{a}) - \sigma(s^2,\hat{a})\|_F^2 \le 3\rho d\|\nabla\sigma\|_\infty^2 |s^1 - s^2|^2 \\
&\le 3\rho d\|\nabla\sigma\|_\infty^2 (L + 4C_1/R_2)^2/\rho^2.
\end{aligned}$$

Since $|J| = 4\rho$, (80) yields $|X_2| \le 6\rho$. By the assumption that $\|\Sigma - \hat{\Sigma}\|_\infty \le \varepsilon_\Sigma$, we obtain

$$\begin{aligned}
\frac{1}{2}\operatorname{Tr}(\Sigma(s^1,\hat{a})X_1) - \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,\hat{a})X_2) &\le \frac{1}{2}\operatorname{Tr}(\Sigma(s^1,\hat{a})X_1) - \frac{1}{2}\operatorname{Tr}(\Sigma(s^2,\hat{a})X_2) + \frac{d}{2}\|\Sigma - \hat{\Sigma}\|_\infty |X_2| \\
&\le 3\rho d\|\nabla\sigma\|_\infty^2 |s^1 - s^2|^2 + 3\rho d\varepsilon_\Sigma \le 3d\|\nabla\sigma\|_\infty^2 (L + 4C_1/R_2)^2/\rho + 3\rho d\varepsilon_\Sigma
\end{aligned}$$

As before, by (62), there exists $C > 0$ such that

$$\frac{1}{2}\operatorname{Tr}(\Sigma(s^1,\hat{a})\nabla^2\phi(s^1)) + \frac{1}{2}\operatorname{Tr}(\hat{\Sigma}(s^2,a_2)\nabla^2\phi(s^2)) \le C/R_2.$$

Plugging these estimates into (81), we obtain

$$\begin{aligned}
\beta(\hat{V}(s^1) - U(s^2)) \le{}& \varepsilon_r + \varepsilon_b(2L + 12C_1/R_2) + \left(\|\nabla r\|_\infty + \|\nabla b\|_\infty(2L + 12C_1/R_2)\right)(L + 4C_1/R_2)/\rho \\
&+ \|b\|_\infty 8C_1/R_2 + 3d\|\nabla\sigma\|_\infty^2 (L + 4C_1/R_2)^2/\rho + 3\rho d\varepsilon_\Sigma + C/R_2.
\end{aligned}$$

Use (75) and take $R_2 \to \infty$ to get from the above that

$$\beta\delta' \le C_5/\rho + 2L\varepsilon_b + \varepsilon_r + 3\rho d\varepsilon_\Sigma.$$

with

$$C_5 := L\|\nabla r\|_\infty + 2L^2\|\nabla b\|_\infty + 3dL^2\|\nabla\sigma\|_\infty^2.$$

Taking $\rho := \sqrt{C_5/(3d\varepsilon_\Sigma)}$ yields

$$\sup_{s\in\mathbb{R}^d}(\hat{V}(s) - U(s)) = 2\delta \le 2\delta' \le \frac{2}{\beta}\left(2\sqrt{3dC_5\varepsilon_\Sigma} + 2L\varepsilon_b + \varepsilon_r\right).$$

Similarly, we obtain the same estimate for $\sup_{s\in\mathbb{R}^d} U(s) - \hat{V}(s)$. Combining these bounds and applying Lemma 3.7, we conclude that

$$\sup_{s\in\mathbb{R}^d}|V^*(s) - U(s)| \le \frac{4}{\beta}\left(2\sqrt{3dC_5\varepsilon_\Sigma} + 2L\varepsilon_b + \varepsilon_r\right).$$

Finally, the conclusion follows from the argument in the last paragraph of the proof of Theorem 3.4. □

7.5. **Proof of Proposition 4.1.**

Proof. The standard LQR problem usually sets $\beta = 0$. In our case, we can use the same method to derive the optimal value function and optimal policy for $\beta \neq 0$. First we know that the optimal value function is quadratic in s, i.e. $V(s) = s^\top Ps + c$ therefore, and it satisfies the following HJB

$$\beta s^\top Ps + \beta c = \max_a \left\{ s^\top Qs + a^\top Ra + 2(As + Ba)^\top Ps \right\} + \sigma^2 \mathrm{diag}(P)$$

When $R \prec 0$, one has

$$\pi^*(s) = \mathrm{argmax} \left\{ s^\top Qs + a^\top Ra + 2(As + Ba)^\top Ps \right\} = Ks, \quad \text{where} \quad K = -R^{-1}B^\top P.$$

Inserting it into the HJB gives

$$V^*(s) = s^\top Ps + \frac{\sigma^2}{\beta}\mathrm{diag}(P), \quad \text{where} \quad \beta P = Q - PBR^{-1}B^\top P + A^\top P + PA.$$

Under assumption 3, there exists a unique negative definite solution to the above Riccati equation.

The one-dimensional solution (36) is obtained by solving (35) analytically. □

7.6. **Proof of Theorem 4.3.** The proof is based on the following Lemma.

Lemma 7.3. *When Q, R are negative definition, $(A - \beta/2, B, C, D)$ is mean-square stabilizable, and $(A - \beta/2, Q, C)$ is detectable, the optimal policy to the following stochastic LQR problem*

$$\begin{aligned} \tilde{V}^*(s) = \max_{a_t = \pi(s_t)} \mathbb{E}\left[\int_0^\infty e^{-\beta t} \left(s_t^\top Q s_t + a_t^\top R a_t \right) dt | s_0 = s \right] \\ s.t. \quad ds_t = (As_t + Ba_t)dt + (Cs_t + Da_t)\, dB_t. \quad \textit{with } B_t \textit{ a scalar Wiener process} \end{aligned} \tag{83}$$

is $\pi^*(s) = Ks$, where

$$K = -(R + D^\top PD)^{-1}(B^\top P + D^\top PC)s. \tag{84}$$

and P is the unique negative definite matrix that satisfies

$$(A - \beta/2)^\top P + P(A - \beta/2) - (PB + C^\top PD)(R + D^\top PD)^{-1}(B^\top P + D^\top PC) + Q + C^\top PC = 0. \tag{85}$$

The optimal policy to the following stochastic LQR problem

$$\begin{aligned} \tilde{V}^*(s) = \max_{a_t = \pi(s_t)} \mathbb{E}\left[\int_0^\infty e^{-\beta t} \left(s_t^\top Q s_t + a_t^\top R a_t \right) dt | s_0 = s \right] \\ s.t. \quad ds_t = (As_t + Ba_t)dt + \sigma dB_t. \quad \textit{with } B_t \textit{ a scalar Wiener process} \end{aligned} \tag{86}$$

is also $\pi^*(s) = Ks$ with the same K defined as (84) for $C = D = 0$.

Remark 7.4. *When $\sigma = 0$, then the optimal value function is $V^*(s) = s^\top Ps$ with P satisfies (85) that exists for $\beta \geq 0$. When $\sigma \neq 0$, then the optimal value function is $V^*(s) = s^\top Ps + \frac{\sigma^2}{\beta}\,\mathrm{Tr}(P)$, which means that the optimal value function is well-defined only when $\beta > 0$.*

Proof. The HJB equation for the optimal control problem (83) is the following,

$$\beta V^*(s) = \max_a \{ s^\top Qs + a^\top Ra + \frac{1}{2}(Cs + Da)^\top \nabla^2 V^*(s)(Cs + Da) \}.$$

Assume that the optimal solution is in terms of

$$V^*(s) = s^\top Ps.$$

Inserting it into the HJB gives,

$$\beta s^\top Ps = \max_a \left\{ s^\top Qs + a^\top Ra + 2(As + Ba)^\top Ps + (Cs + Da)^\top P(Cs + Da) \right\}$$

Taking gradient w.r.t. a for the RHS gives,

$$2Ra + 2B^\top Ps + 2D^\top P(Cs + Da) = 0, \quad \pi^*(a) = -(R + D^\top PD)^{-1}(B^\top P + D^\top PC)s$$

Inserting the above policy to the RHS of the equation gives the Riccati equation (85) for P.

The HJB equation for the optimal contorl problem (86) is

$$\beta V^*(s) = \max_a \{s^\top Q s + a^\top R a + \frac{1}{2}\sigma^2 \Delta V^*(s)\}.$$

Assume that the optimal solution is in terms of

$$V^*(s) = s^\top P s + e.$$

Inserting it into the HJB gives,

$$\beta(s^\top P s + e) = \max_a \left\{ s^\top Q s + a^\top R a + 2(As + Ba)^\top P s + \sigma^2 \operatorname{Tr}(P) \right\}$$

Taking gradient w.r.t. a for the RHS gives,

$$2Ra + 2B^\top P s = 0, \quad \pi^*(a) = -R^{-1} B^\top P s$$

Inserting the above policy to the RHS of the equation gives the Riccati equation (85) for P and $e = \frac{\sigma^2 \operatorname{Tr}(P)}{\beta}$ □

Now we are ready to proof Theorem 4.3.

Proof. **RL approximation.** For the deterministic case, where $\sigma = 0$, the Optimal-BE can be viewed as a discrete-time infinite horizon LQR problem.

$$\tilde{V}^*(s) = \max_a \left[(s^\top \tilde{Q} s + a^\top \tilde{R} a) + \gamma \tilde{V}^*(p_{\Delta t}(s,a)) \right], \quad p_{\Delta t}(s,a) = \tilde{A} s + \tilde{B} a \tag{87}$$

where

$$\tilde{A} = \hat{A}_1 \Delta t + I, \quad \tilde{B} = \hat{B}_1 \Delta t.$$

Therefore, the optimal policy should be a linear policy $\pi^*(s) = \tilde{K} s$ and optimal value function should be quadratic $\tilde{V}^*(s) = s^\top \tilde{P} s$ [4]. Inserting $\tilde{V}^*(s) = s^\top \tilde{P} s$ to (87) gives

$$s^\top \tilde{P} s = \max_a \left(s^\top \tilde{Q} s + a^\top \tilde{R} a + \gamma (\tilde{A} s + \tilde{B} a)^\top \tilde{P} (\tilde{A} s + \tilde{B} a) \right) \tag{88}$$

Multiplying $\frac{1}{\gamma \Delta t}$ on both sides and then substract $\frac{1}{\Delta t} s^\top \tilde{P} s$ on both sides gives,

$$\frac{1}{\Delta t \gamma} s^\top \tilde{P} s - \frac{1}{\Delta t} s^\top \tilde{P} s = \max_a \left(s^\top \left(\frac{1}{\Delta t \gamma} \tilde{Q} \right) s + a^\top \left(\frac{1}{\Delta t \gamma} \tilde{R} \right) a + \frac{1}{\Delta t} (\tilde{A} s + \tilde{B} a)^\top \tilde{P} (\tilde{A} s + \tilde{B} a) - \frac{1}{\Delta t} s^\top \tilde{P} s \right),$$

Furthermore,

$$\begin{aligned} &\frac{1}{\Delta t} (\tilde{A} s + \tilde{B} a)^\top \tilde{P} (\tilde{A} s + \tilde{B} a) - \frac{1}{\Delta t} s^\top \tilde{P} s = \frac{1}{\Delta t} (\tilde{A} s + \tilde{B} a - s)^\top \tilde{P} (\tilde{A} s + \tilde{B} a - s) + \frac{2}{\Delta t} (\tilde{A} s + \tilde{B} a - s)^\top \tilde{P} s \\ =&(\hat{A}_1 s + \hat{B}_1 a)^\top \tilde{P} (\hat{A}_1 s + \hat{B}_1 a) \Delta t + 2 (\hat{A}_1 s + \hat{B}_1 a)^\top \tilde{P} s, \end{aligned} \tag{89}$$

where the following equality is used to obtain the above second equality,

$$\frac{1}{\Delta t} (\tilde{A} s + \tilde{B} a - s) = \hat{A}_1 s + \hat{B}_1 a.$$

Then one can equivalently write (88) in the following form,

$$\hat{\beta} s^\top \tilde{P} s = \max_a \left(s^\top \hat{Q} s + a^\top \hat{R} a + (\hat{A}_1 s + \hat{B}_1 a) \cdot (2 \tilde{P} s) + (\hat{A}_1 s + \hat{B}_1 a)^\top \tilde{P} (\hat{A}_1 s + \hat{B}_1 a) \Delta t \right), \tag{90}$$

with

$$\hat{\beta} = \frac{1}{\Delta t \gamma} - \frac{1}{\Delta t}, \quad \hat{Q} = \frac{1}{\Delta t \gamma} \tilde{Q}, \quad \hat{P} = \frac{1}{\Delta t \gamma} \tilde{P}.$$

When $\hat{R}$ is negative definite, then the optimal policy induced by the RHS of (90) is

$$\hat{R} a^* + \hat{B}_1^\top \tilde{P} s + \hat{B}_1^\top \tilde{P} (\hat{A}_1 s + \hat{B}_1 a^*) \Delta t = 0, \quad \tilde{\pi}^*(s) = -(\hat{R} + \hat{B}_1^\top \tilde{P} \hat{B}_1 \Delta t)^{-1} \left(\hat{B}_1^\top \tilde{P} + \hat{B}_1^\top \tilde{P} \hat{A}_1 \Delta t \right) s.$$

Inserting it back to (90) gives the Riccati equation for $\tilde{P}$,

$$\hat{\beta} \tilde{P} = -\left(\hat{B}_1^\top \tilde{P} + \hat{B}_1^\top \tilde{P} \hat{A}_1 \Delta t \right)^\top (\hat{R} + \hat{B}_1^\top \tilde{P} \hat{B}_1 \Delta t)^{-1} \left(\hat{B}_1^\top \tilde{P} + \hat{B}_1^\top \tilde{P} \hat{A}_1 \Delta t \right) + \hat{Q} + \hat{A}_1^\top \tilde{P} + \tilde{P} \hat{A}_1 + \hat{A}_1^\top \tilde{P} \hat{A}_1 \Delta t.$$

Therefore, according to Lemma 7.3, by replacing β, A, B, C, D, Q, R with $\hat{\beta}, \hat{A}_1, \hat{B}_1, \hat{A}_1\sqrt{\Delta t}, \hat{B}_1\sqrt{\Delta t}, \hat{Q}, \hat{R}$, one comes to the conclusion that the optimal policy derived from (87) is the same as the solution to (41).

For the stochastic case, where $\sigma \neq 0$, one assumes that the optimal solution to the Optimal-BE is in the form of $\hat{V}^*(s) = s^\top \tilde{P} s + b$ with constant b. Inserting it into (39) gives,

$$(s^\top \tilde{P} s + b) = \max_a \left\{ s^\top \tilde{Q} s + a^\top \tilde{R} a + \gamma \int (s')^\top \tilde{P} s' \rho_{\Delta t}(s'|s,a) ds' + \gamma b \right\}$$
$$(s^\top \tilde{P} s + b) = \max_a \left\{ s^\top \tilde{Q} s + a^\top \tilde{R} a + \gamma(\tilde{A}s + \tilde{B}a)^\top \tilde{P}(\tilde{A}s + \tilde{B}a) \right\} + \gamma\sigma^2 \operatorname{Tr}(\tilde{P}C_A)\Delta t + \gamma b$$

Multiplying $\frac{1}{\gamma \Delta t}$ on both sides and then substract $\frac{1}{\Delta t} s^\top \tilde{P} s$ on both sides gives,

$$\frac{1}{\gamma\Delta t}(s^\top \tilde{P} s + b) = \max_a \left\{ s^\top \left(\frac{1}{\gamma\Delta t}\tilde{Q}\right) s + a^\top \left(\frac{1}{\gamma\Delta t}\tilde{R}\right) a + \frac{1}{\Delta t}(\tilde{A}s + \tilde{B}a)^\top \tilde{P}(\tilde{A}s + \tilde{B}a) \right\} + \sigma^2 \operatorname{Tr}(\tilde{P}C_A) + \frac{b}{\Delta t}$$
$$\hat{\beta}(s^\top \tilde{P} s + b) = \max_a \left\{ (s^\top \hat{Q} s + a^\top \hat{R} a) + \frac{1}{\Delta t}(\tilde{A}s + \tilde{B}a)^\top \tilde{P}(\tilde{A}s + \tilde{B}a) - \frac{1}{\Delta t} s^\top \tilde{P} s \right\} + \sigma^2 \operatorname{Tr}(\tilde{P}C_A)$$

Based on the same equality as (89), one has

$$\hat{\beta} s^\top \tilde{P} s = \max_a \left(s^\top \hat{Q} s + a^\top \hat{R} a + (\hat{A}_1 s + \hat{B}_1 a) \cdot (2\tilde{P} s) + (\hat{A}_1 s + \hat{B}_1 a)^\top \tilde{P}(\hat{A}_1 s + \hat{B}_1 a)\Delta t \right),$$
$$\hat{\beta} b = \sigma^2 \operatorname{Tr}(\tilde{P}C_A),$$

Note the optimal policy $\tilde{\pi}^*(s)$ is driven by the RHS of equation for $\tilde{P}$, which is the same as (90). This implies that the Optimal-BE solution (39) is the same for different σ. Therefore, based on Lemma 7.3, one completes the proof for Optimal-BE part of the Lemma.

PhiBE approximation. First note that given $s_0 = s$ and $a_\tau = a$ for $\tau \in [0, i\Delta t)$, one has

$$\mathbb{E}[s_{j\Delta t}] = e^{Aj\Delta t} s + A^{-1}(e^{Aj\Delta t} - I)Ba.$$

By the definition of $\hat{b}_i$ in (19), one has

$$\hat{b}_i(s,a) = \frac{1}{\Delta t}\sum_{j=1}^{i} a_j^{(i)}(e^{Aj\Delta t} - I)s + \frac{1}{\Delta t}\sum_{j=1}^{i} a_j^{(i)} A^{-1}(e^{Aj\Delta t} - I)Ba = \hat{A}_i s + \hat{B}_i a$$

where $\hat{A}_i, \hat{B}_i$ are defined in (37). This implies that the Optimal-PhiBE solves the following deterministic LQR problem,

$$V_i^*(s) = \max_{a_t = \pi(s_t)} \mathbb{E}\left[\int_0^\infty e^{-\beta t} \left(s_t^\top Q s_t + a_t^\top R a_t \right) dt | s_0 = s \right]$$
$$s.t. \quad ds_t = (\hat{A}_i s_t + \hat{B}_i a_t) dt$$

According to Lemma 7.3, the optimal control for the stochastic LQR with $C = D = 0$ and $\sigma \neq 0$ is the same as the deterministic LQR problem. Therefore, one completes the proof for the second part of the Lemma on the Optimal-PhiBE. □

7.7. **Proof of Theorem 4.4.** Before the proof of Theorem 4.4, we need a Lemma to characterize the difference between $(\hat{A}_i, \hat{B}_i)$ and (A, B). Define

$$\epsilon_i^A = \hat{A}_i - A, \quad \epsilon_i^B = \hat{B}_i - B. \tag{91}$$

we give an upper bound for $\epsilon_i^A, \epsilon_i^B$ in the following lemma.

Lemma 7.5. *For Δt sufficiently small, s.t. $\exp(\|A\| i\Delta t) \leq C(\|A\| i\Delta t + 1)$, one can bound the spectrum norm of ϵ_i^A and ϵ_i^B by*

$$\left\|\epsilon_i^A\right\| \leq \hat{C}_i \|A\|^{i+1} \Delta t^i + C\hat{C}_i \|A\|^{2+1} \Delta t^{i+1}, \quad \left\|\epsilon_i^B\right\| \leq \hat{C}_i \left\|A^{-1}\right\| \|A\|^{i+1} \|B\| \Delta t^i + C\hat{C}_i \left\|A^{-1}\right\| \|A\|^{i+2} \|B\| \Delta t^{i+1}.$$

Proof. 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

$$\|R_{ij}\| \leq \frac{\|A\|^{i+1}(j\Delta t)^{i+1}}{(i+1)!} e^{\|A\| j\Delta t} = \frac{\|A\|^{i+1}(j\Delta t)^{i+1}}{(i+1)!} C_A, \quad \text{with} \quad C_A = e^{\|A\| i\Delta t}.$$

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

$$\left\| \hat{A}_i - A \right\| \leq \frac{1}{\Delta t} \sum_{j=1}^{i} \left| a_j^{(i)} \right| \|R_{ij}\| \leq \frac{\sum_{j=1}^{i} \left| a_j^{(i)} \right| j^{i+1}}{(i+1)!} \|A\|^{i+1} \Delta t^i C_A = C_A \hat{C}_i \|A\|^{i+1} \Delta t^i. \tag{92}$$

where $\hat{C}_i$ is defined in (53). Similarly, one has,

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

where the last equality is due to the definition of $a_j^{(i)}$ in (21). Therefore, one has

$$\left\| \hat{B}_i - B \right\| \leq \frac{1}{\Delta t} \sum_{j=1}^{i} \left| a_j^{(i)} \right| \left\| A^{-1} \right\| \|R_{ij}\| \|B\| C_A \leq C_A \hat{C}_i \left\| A^{-1} \right\| \|A\|^{i+1} \|B\| \Delta t^i. \tag{93}$$

When Δt is sufficiently small, s.t. $C_A \leq C(\|A\| i\Delta t + 1)$, one ends up the inequality in the Lemma.

□

Now we are ready to prove Theorem 4.4.

Proof. **RL approximation.** According to Lemma 7.3 and Theorem 4.3, the optimal policy from the Optimal-BE is $\tilde{\pi}^*(s) = \tilde{K}s$, with $\tilde{K}$ defined as

$$\tilde{K} = -(R + \hat{B}_1^\top \tilde{P} \hat{B}_1 \Delta t)^{-1} (\hat{B}_1^\top \tilde{P} + \hat{B}_1^\top \tilde{P} \hat{A}_1 \Delta t).$$

and $\tilde{P}$ satisfying

$$(\hat{A}_1 - \beta/2)^\top \tilde{P} + \tilde{P}(\hat{A}_1 - \beta/2) - (\tilde{P}\hat{B}_1 + \hat{A}_1^\top \tilde{P} \hat{B}_1 \Delta t)(R + \hat{B}_1^\top \tilde{P} \hat{B}_1 \Delta t)^{-1} (\hat{B}_1^\top \tilde{P} + \hat{B}_1^\top \tilde{P} \hat{A}_1 \Delta t) + Q + \hat{A}_1^\top \tilde{P} \hat{A}_1 \Delta t = 0. \tag{94}$$

In one-dimensional case, one can equivelently write the above equation as

$$-(1 + \beta\Delta t)\hat{B}_1^2 \tilde{P}^2 + ((2\hat{A}_1 - \beta)R + Q\hat{B}_1^2 \Delta t + \hat{A}_1^2 R \Delta t)\tilde{P} + QR = 0.$$

Since the coefficient for $\tilde{P}^2$ is negative and the constant is positive, therefore there always exists two solutions and only one of them is negative. By replacing $\tilde{P}$ by

$$\tilde{P} = -(\hat{B}_1 + \hat{B}_1^2 \tilde{K} \Delta t + \hat{A}_1 \hat{B}_1 \Delta t)^{-1} R \tilde{K}$$

in (94), one can rewrite the equation in terms of $\tilde{K}$,

$$-(1 + \hat{A}_1 \Delta t)\tilde{K}^2 + \left[-\frac{2\hat{A}_1}{\hat{B}_1} + \frac{\beta}{\hat{B}_1} + \frac{Q}{R} \hat{B}_1 \Delta t - \frac{\hat{A}_1^2}{\hat{B}_1} \Delta t \right] \tilde{K} + \left[\frac{Q}{R} + \frac{Q}{R} \hat{A}_1 \Delta t \right] = 0$$

Since $\frac{\hat{A}_1}{\hat{B}_1} = \frac{A}{B}$, one can equivalently write the above equation as

$$-(B+\epsilon_2)\tilde{K}^2 + [-2A+\beta+\epsilon_1]\,\tilde{K} + \left[\frac{QB}{R}+\epsilon_0\right] = 0$$

with

$$\epsilon_2 = B\hat{A}_1\Delta t, \quad \epsilon_1 = \beta\left(\frac{B}{\hat{B}_1}-1\right) + \frac{Q}{R}\hat{B}_1 B\Delta t - \frac{\hat{A}_1^2 B}{\hat{B}_1}\Delta t, \quad \epsilon_0 = \frac{QB}{R}\hat{A}_1\Delta t.$$

Comparing it to the optimal policy from the true dynamics

$$-BK^2 + (-2A+\beta)\,K + \frac{QB}{R} = 0,$$

one can write the difference $|K-\tilde{K}|$ in terms of $a=-B, b=-2A+\beta, c=QB/R$ and $\epsilon_i, i=1,2,3$,

$$\begin{aligned}
|K-\tilde{K}| =& \left|\frac{-b+\sqrt{b^2-4ac}}{2a} - \frac{-(b+\epsilon_1)+\sqrt{(b+\epsilon_1)^2-4(a+\epsilon_2)(c+\epsilon_0)}}{2(a+\epsilon_2)}\right| \\
=& \frac{1}{2|a|}\left|\left(-1+\frac{b}{\sqrt{b^2-4ac}}\right)\epsilon_1 - \frac{2a}{\sqrt{b^2-4ac}}\epsilon_0 + \left(-\frac{2c}{\sqrt{b^2-4ac}} + \frac{-b+\sqrt{b^2-4ac}}{2a}\right)\epsilon_2\right| + O\left(\sum_j \epsilon_j^2\right) \\
\lesssim& \frac{1}{|a|}\left||\epsilon_1| + \frac{|a|}{\sqrt{|ac|}}|\epsilon_0| + \left(\frac{|c|}{\sqrt{|ac|}} + \frac{|b|+\sqrt{|ac|}}{|a|}\right)|\epsilon_2|\right| + O\left(\sum_j \epsilon_j^2\right) \\
\lesssim& \frac{1}{|B|}\left(|Q/RB^2 - A^2| + |AB|\sqrt{Q/R} + |A-\beta/2||A|\right)\Delta t + O(\Delta t^2) \\
\lesssim& \left[|B|\left(\sqrt{\frac{Q}{R}} + \frac{|A|}{|B|}\right)^2 + |A-\beta/2|\frac{|A|}{|B|}\right]\Delta t + O(\Delta t^2)
\end{aligned}$$

where the first equality is obtained by taking Taylor expansion of $\tilde{K}$ around $\epsilon_j = 0$ and the first inequality uses $|b|, 2\sqrt{ac} \le \sqrt{b^2-4ac} \lesssim |b| + \sqrt{ac}$. The second inequality is obtained by the fact that $\hat{A}_1 = A + O(\Delta t), \hat{B}_1 = B + O(\Delta t)$.

PhiBE Approximation. Since PhiBE can also be viewed as an LQR with approxiamted dynamics $\hat{A}_i, \hat{B}_i$. Therefore, according to Proposition 4.1, the optimal control is $\hat{\pi}_i^*(s) = \hat{K}_i s$ with

$$\hat{K}_i = \frac{\beta/2}{\hat{B}_i} - \frac{\hat{A}_i}{\hat{B}_i} + \sqrt{\left(\frac{\beta/2}{\hat{B}_i} - \frac{\hat{A}_i}{\hat{B}_i}\right)^2 + \frac{Q}{R}}. \tag{95}$$

By the definition of $\hat{A}_i, \hat{B}_i$ in (37), one has

$$\frac{\hat{A}_i}{\hat{B}_i} = \frac{A}{B}.$$

Inserting the above equality into (95) and setting $\beta = 0$, one has,

$$\hat{K}_i = -\frac{A}{B} + \sqrt{\left(-\frac{A}{B}\right)^2 + \frac{Q}{R}} = K,$$

which gives the equality in (44).

For the case where $\beta > 0$, by letting

$$\hat{B}_i = B + \epsilon_i^B, \quad D = \frac{A}{B},$$

one has

$$\hat{K}_i = \frac{\beta/2}{B+\epsilon_i^B} - D + \sqrt{\left(\frac{\beta/2}{B+\epsilon_i^B} - D\right)^2 + \frac{Q}{R}}. \tag{96}$$

Since

$$\frac{\beta/2}{B+\epsilon_i^B} = \frac{\beta/2}{B} + \frac{\beta/2}{B^2}\epsilon_i^B + O(\Delta t^{2i}), \quad \sqrt{(a+c)^2+d} = \sqrt{a^2+d} + \frac{a}{\sqrt{a^2+d}}c + O(c^2)$$

one can equivalently write (96) as

$$\hat{K}_i = \frac{\beta/2}{B} - D + \sqrt{\left(\frac{\beta/2}{B} - D\right)^2 + \frac{Q}{R}} + \frac{\beta/2}{B^2}\epsilon_i^B + \frac{|\frac{\beta/2}{B} - D|}{\sqrt{\left(\frac{\beta/2}{B} - D\right)^2 + \frac{Q}{R}}}\frac{\beta/2}{B^2}\epsilon_i^B + O(\Delta t^{2i}),$$

which implies

$$|\hat{K}_i - K| \leq \frac{\beta/2}{B^2}|\epsilon_i^B|\left(1 + \frac{|\frac{\beta/2}{B} - D|}{\sqrt{\left(\frac{\beta/2}{B} - D\right)^2 + \frac{Q}{R}}}\right) + O(\Delta t^{2i}) \leq \frac{\beta}{B^2}|\epsilon_i^B| + O(\Delta t^{2i}).$$

By the bound given in Lemma 3.5, one has

$$|\hat{K}_i - K| \leq \frac{\beta \hat{C}_i}{|B|}|A|^i \Delta t^i + O(\Delta t^{i+1}).$$

Note that the optimal solution $\hat{P}_i$ satisfies

$$-B^2P^2 + (2A - \beta)RP + QR = 0$$

with negative coefficient for the quadratic term and positive constant, so the well-posedness can also be guaranteed. □

7.8. **Proof of Lemma 4.5.**

Proof. We first prove the case where $\beta = 0$. Note that one can equivalently write $\hat{A}_i, \hat{B}_i$ in terms of

$$\hat{A}_i = A + \sum_{i=2}^{\infty} a_i A^i, \quad \hat{B}_i = B + \sum_{i=2}^{\infty} a_i A^{i-1}B.$$

If λ is the eigenvalue of A, then

$$\hat{\lambda} = \lambda + \sum_{i=2}^{\infty} a_i \lambda^i \tag{97}$$

is the eigenvalue of $\hat{A}_i$. In order to prove $(\hat{A}_i, \hat{B}_i)$ is stablizable, it is equivalent to prove that for $\mathrm{Re}(\hat{\lambda}) < 0$,

$$\mathrm{rank}[\hat{A}_i - \hat{\lambda}, \hat{B}_i] = d$$

This is equivalent to prove that the span of the column of $\hat{B}_i$ covers the null space of $\hat{A}_i - \hat{\lambda}$. Note that the null space of $A - \lambda I$ is the same as the null space of $\hat{A}_i - \hat{\lambda}$, so we only need to prove that the column of $\hat{B}_i$ covers the null space of $A - \lambda I$.

By the definition of $\hat{A}_i, \hat{B}_i$, one has,

$$\left\|\hat{\lambda} - \lambda\right\| \leq O(\Delta t^i) \tag{98}$$

which imples that if $\mathrm{Re}(\hat{\lambda}) < 0$, then $\mathrm{Re}(\lambda) < 0$ for sufficiently small Δt. Since (A, B) is stablizable, which implies that the span of the column of B covers the null space of $A - \lambda$ for $\mathrm{Re}(\lambda) < 0$, i.e., for $\forall v$ that satisfying $(A - \lambda)v = 0$, there exists a constant vector $c \in \mathbb{R}^d$, such that

$$v = Bc.$$

This leads to

$$\hat{B}_i c = Bc + \sum_{i=2}^{\infty} a_i A^{i-1}Bc = v + \sum_{i=2}^{\infty} a_i A^{i-1}v = \frac{v}{\lambda}\hat{\lambda},$$

which implies that any vector v in the null space of $A-\lambda$, there exists a constant vector $\frac{\lambda}{\hat{\lambda}}c$ s.t.

$$\hat{B}_i \begin{bmatrix} \lambda \\ \frac{\lambda}{\hat{\lambda}} c \end{bmatrix} = v$$

which completes the proof for the first part.

To prove $(\hat{A}_i, Q)$ is detectable, one needs to prove that if the eigenvector v of $\hat{A}_i$ such that $Qv = 0$, then the corresponding eigenvalue $\hat{\lambda}$ of $\hat{A}_i$ needs to have negative real part. Since A and $\hat{A}_i$ has the same eigenvector, and because (A, Q) is detectable, which implies that for this eigenvector v, the corresponding eigenvalue λ of A has negative eigenvalue. Since the difference between $\hat{\lambda}$ and λ are small according to (98) when Δt is sufficiently small, which implies that $\hat{\lambda}$ also have negative real part.

The above arguements all hold when $\beta \neq 0$, which completest the proof for this Lemma. □

7.9. **Proof of Theorem 4.6.**

Proof. Another equivalent condition that choosing the correct solution to the Riccati equaiton is the unique P such that all the real part of the eigenvalues of

$$A - BR^{-1}B^\top P - \beta/2 \tag{99}$$

are negative.

First one notes that by setting $M = PB$ and using the definition of P in Proposition 4.1, one can rewrite the definition of K in terms of M

$$K = -R^{-1}M, \quad \text{with} \quad \beta M B^{-1} = Q - MR^{-1}M^\top + D^\top M^\top + MD, \quad D = B^{-1}A. \tag{100}$$

and the condition (99) for P can be rewritten as

$$\text{the eigenvalue of } B(D - R^{-1}M^\top) - \frac{\beta}{2} \text{ are all negative.} \tag{101}$$

Since PhiBE can also be viewed as an LQR with approximated dynamics $\hat{A}_i, \hat{B}_i$ and based on the fact that

$$\hat{D} = \hat{B}_i^{-1}\hat{A}_i = B^{-1}A = D.$$

where $\hat{A}_i, \hat{B}_i$ are defined in (37), the optimal control under PhiBE approximation is $\hat{\pi}_i^*(s) = \hat{K}_i s$ with

$$\hat{K}_i = -R^{-1}\hat{M}, \quad \text{with} \quad \beta \hat{M} B^{-1} = Q - \hat{M}R^{-1}\hat{M}^\top + D^\top \hat{M}^\top + \hat{M}D, \quad D = \hat{B}_i^{-1}\hat{A}_i = B^{-1}A. \tag{102}$$

and the condition for $\hat{M}$ is

$$\text{the eigenvalue of } \hat{B}_i(D - R^{-1}\hat{M}^\top) - \frac{\beta}{2} \text{ are all negative.} \tag{103}$$

By the wellposedness assumption in Lemma 4.5, there exists a unique solution M and $\hat{M}$.

For $\beta = 0$, we only need to prove that the unique solution K to (100) - (101) also satisfies (102) - (103) for sufficiently small Δt. Note that by setting $\beta = 0$ for both (100) and (102), both M and $\hat{M}$ satisfy the same quadratic equation

$$MR^{-1}M^\top - D^\top M^\top - MD - Q = 0. \tag{104}$$

The only difference between K and $\hat{K}_i$ is the condition that one requires

$$\text{the eigenvalue of } C = B(D - R^{-1}M^\top), \quad C_i = \hat{B}_i(D - R^{-1}\hat{M}^\top) \text{ are all negative.} \tag{105}$$

By the definition of $\hat{B}_i$ and Lemma 7.5, one has

$$\|\lambda(C) - \lambda(C_i)\| \leq \kappa(C)\|C - C_i\| \leq \kappa(C)\left\|\epsilon_i^B(D - R^{-1}M^\top)\right\| \leq \kappa(C)\hat{C}_i\|A\|^i\|B\|\left\|(D - R^{-1}M^\top)\right\| O(\Delta t^i).$$

Since all the real part of the eigenvalues of C are negative, which implies that $\kappa(C) < \infty$, and combine with the fact that $\|M\| < \infty$, which implies that as long as Δt is sufficiently small, then all the eigenvalues of C_i is sufficiently close to C. Therefore, all the eigenvalues of C_i are all negative. This means that K satisfies (102) - (103), and it completes the proof for $\beta = 0$.

Now for $\beta \neq 0$, using the fact that

$$MB^{-1} = (MB^{-1})^\top = P, \quad \hat{M}\hat{B}_i^{-1} = (\hat{M}\hat{B}_i^{-1})^\top = \hat{P}_i$$

one can rewrite the two equations for $M, \hat{M}$ by

$$-MR^{-1}M^\top + (D + \frac{\beta}{2}B^{-1})^\top M^\top + M(D + \frac{1}{2}B^{-1}) + Q = 0,$$
$$-\hat{M}R^{-1}\hat{M}^\top + (D + \frac{\beta}{2}\hat{B}_i^{-1})^\top \hat{M}^\top + \hat{M}(D + \frac{1}{2}\hat{B}_i^{-1}) + Q = 0$$

There are many perturbation theorem that provides the error bound for $\left\|M - \hat{M}\right\|$, here we adopt the bound provided in [Theorem 3.1 of [40]], one can bound

$$\left\|M - \hat{M}\right\| \lesssim p\beta \left\|B^{-1} - \hat{B}_i^{-1}\right\| + O\left(\left\|B^{-1} - \hat{B}_i^{-1}\right\|^2\right)$$

where p is a constant depends on D, B, R, M and can be bounded by

$$p \leq \|M\| \left\|T^{-1}\right\| \tag{106}$$

with

$$T = I_d \otimes [(D + \frac{\beta}{2}B^{-1}) - R^{-1}M]^\top + [(D + \frac{\beta}{2}B^{-1}) - R^{-1}M]^\top \otimes I_d.$$

Now we bound $\left\|B^{-1} - \hat{B}_i^{-1}\right\|$. By Taylor expansion, one has

$$\hat{B}_i^{-1} = (B + \epsilon_i^B)^{-1} = B^{-1}\sum_{k=0}^{\infty}(-B^{-1}\epsilon_i^B)^k$$

which leads to

$$\left\|B^{-1} - \hat{B}_i^{-1}\right\| = \left\|B^{-1}\right\| \left\|I - \sum_{k=0}^{\infty}(-B^{-1}\epsilon_i^B)^k\right\| \leq \left\|B^{-1}\right\| \left[\left\|B^{-1}\right\| \left\|\epsilon_i^B\right\| + O(\left\|\epsilon_i^B\right\|^2)\right]$$
$$= \hat{C}_i \left\|A^{-1}\right\| \|A\|^{i+1} \|B\| \left\|B^{-1}\right\|^2 \Delta t^i + O(\Delta t^{i+1})$$

□

8. Algorithms

Algorithm 5 RL_Policy_Evaluation($\Delta t, \beta, B, \Phi$) - RL Policy evaluation method

1: **Input:** discrete time step Δt, discount coefficient β, discrete-time trajectory data $B = \{(s^l_{j\Delta t}, r^l_{j\Delta t})_{j=0}^m\}_{l=1}^I$ generated by applying corresponding policy π and finite bases $\Phi(s) = (\phi_1(s), ..., \phi_n(s))^\top$.

2: **Output:** Value function $\hat{V}^\pi$.

3: Compute

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

4: Compute

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

5: Compute

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

return $\hat{V}^\pi(s) = \theta^\top\Phi(s)$.

9. Experimental Details

9.1. Ground Truth Optimal Value Functions.

Algorithm 6 RL_Q_EVALUATION($\Delta t, \beta, B, \Psi$) - RL method for $\hat{Q}^\pi$

Input: discrete time step Δt, discount coefficient β, and discrete-time trajectory data $B = \{(s^l_{j\Delta t}, a^l_{j\Delta t}, r^l_{j\Delta t})_{j=0}^m\}_{l=1}^I$ generated by applying random actions at the first discrete time step then following policy π, finite bases $\Psi(s,a) = (\psi_1(s,a), ..., \psi_n(s,a))^\top$.

Output: Continuous Q-function for policy π.

Initialize $w = w_0$.

Compute

$$A = \sum_{l=1}^{I} \sum_{j=0}^{m-1} \Psi(s^l_{j\Delta t}, a^l_{j\Delta t}) \Big[\Psi(s^l_{j\Delta t}, a^l_{j\Delta t}) - e^{-\beta \Delta t} \Psi(s^l_{(j+1)\Delta t}, a^l_{j\Delta t}) \Big]^\top.$$

Compute

$$b = \sum_{l=1}^{I} \sum_{j=0}^{m-1} r^l_{j\Delta t} \Psi(s^l_{j\Delta t}, a^l_{j\Delta t}) \cdot \Delta t.$$

Compute

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

return $\hat{Q}^\pi(s,a) = \theta^\top \Psi(s,a)$.

Algorithm 7 RL_OPTIMAL($\Delta t, \beta, \Phi, \Psi, \pi_0$) - RL algorithm for finding the optimal policy

Input: discrete time step Δt, discount coefficient β, finite bases for policy evaluation $\Phi(s) = (\phi_1(s), ..., \phi_{n'}(s))^\top$, finite bases for Q-approximation $\Psi(s,a) = (\psi_1(s,a), ..., \psi_n(s,a))^\top$, initial policy π_0.

Output: Optimal policy $\pi^*(s)$, optimal value function $V^{\pi^*}(s)$.

Initialize $\pi(s) = \pi_0(s)$.

while not *Stopping Criterion Satisfied* **do**

  Generate data $B = \{(s^l_{j\Delta t}, a^l_{j\Delta t}, r^l_{j\Delta t})_{j=0}^m\}_{l=1}^I$ by applying random actions at the first discrete time step and then following policy π.

  Call Algorithm 6 to obtain

$$\hat{Q}^\pi(s,a) = \text{RL\_Q\_EVALUATION}(\Delta t, \beta, B, \Psi).$$

  Update the optimal policy:

$$\pi(s) \leftarrow \text{argmax}_a \, \hat{Q}^\pi(s,a).$$

end while

Generate data $B = \{(s^l_{j\Delta t}, r^l_{j\Delta t})_{j=0}^m\}_{l=1}^I$ by applying policy π.

Call Algorithm 5 to obtain

$$\hat{V}^\pi(s) = \text{RL\_POLICY\_EVALUATION}(\Delta t, \beta, B, \Phi).$$

return $\pi^*(s) = \pi(s)$, $V^{\pi^*}(s) = \hat{V}^\pi(s)$.

9.1.1. *One-dimensional Deterministic Case.*

- (Case 1) Optimal policy: $\pi(s) = -2.4142s$; optimal value function: $V^*(s) = -2.4142s^2$.
- (Case 2) Optimal policy: $\pi(s) = -20.050s$; optimal value function: $V^*(s) = -200.50s^2$.
- (Case 3) Optimal policy: $\pi(s) = -101.00s$; optimal value function: $V^*(s) = -1.0100s^2$.
- (Case 4) Optimal policy: $\pi(s) = -200.0050s$; optimal value function: $V^*(s) = -200.0050s^2$.

9.1.2. *One-dimensional Stochastic Case.*

- (Case 1) Optimal policy: $\pi(s) = -2.4057s$; optimal value function: $V^*(s) = -240.57 - 2.4057s^2$.
- (Case 2) Optimal policy: $\pi(s) = -19.95012s$; optimal value function: $V^*(s) = -19950.12 - 199.5012s^2$.
- (Case 3) Optimal policy: $\pi(s) = -101.00s$; optimal value function: $V^*(s) = -101.0000 - 1.0100s^2$.

- (Case 4) Optimal policy: $\pi(s) = -199.9950s$; optimal value function: $V^*(s) = -19999.50 - 199.9950s^2$.

9.1.3. *Two-dimensional Deterministic Case.*

- (Case 1) Optimal policy: $\pi(s) = \begin{pmatrix} -0.3994 & 0.1253 \\ 0.1163 & -0.5850 \end{pmatrix} s$; optimal value function: $V^*(s) = -0.4462s_1^2 - 0.4279s_2^2 + 0.0353s_1s_2$.
- (Case 2) Optimal policy: $\pi(s) = \begin{pmatrix} -0.1335 & 0.0258 \\ 0.0115 & -0.2110 \end{pmatrix} s$; optimal value function: $V^*(s) = -2.7454s_1^2 - 2.8184s_2^2 + 0.8011s_1s_2$.
- (Case 3) Optimal policy: $\pi(s) = \begin{pmatrix} -9.8068 & 9.5130 \\ 9.5107 & -9.8121 \end{pmatrix} s$; optimal value function: $V^*(s) = -7.7180s_1^2 - 7.6702s_2^2 + 5.9615s_1s_2$.
- (Case 4) Optimal policy: $\pi(s) = \begin{pmatrix} -494.1155 & 492.5599 \\ 509.9291 & -512.4116 \end{pmatrix} s$; optimal value function: $V^*(s) = -87470.0790s_1^2 - 87437.5963s_2^2 + 174878.7572s_1s_2$.

9.1.4. *Two-dimensional Stochastic Case.*

- (Case 1) Optimal policy: $\pi(s) = \begin{pmatrix} -0.3978 & 0.1247 \\ 0.1155 & -0.5830 \end{pmatrix} s$; optimal value function: $V^*(s) = -870.93 - 4.4447s_1^2 - 4.2647s_2^2 + 0.34674s_1s_2$.
- (Case 2) Optimal policy: $\pi(s) = \begin{pmatrix} -0.1330 & 0.0256 \\ 0.0112 & -0.2104 \end{pmatrix} s$; optimal value function: $V^*(s) = -554.7285 - 2.7371s_1^2 - 2.8102s_2^2 + 0.7937s_1s_2$.
- (Case 3) Optimal policy: $\pi(s) = \begin{pmatrix} -9.8036, & 9.5108 \\ 9.5085 & -9.8090 \end{pmatrix} s$; optimal value function: $V^*(s) = -1537.0394 - 7.7090s_1^2 - 7.6614s_2^2 + 5.9735s_1s_2$.
- (Case 4) Optimal policy: $\pi(s) = \begin{pmatrix} -493.9887 & 492.4366 \\ 509.8068 & -512.2836 \end{pmatrix} s$; optimal value function: $V^*(s) = -17486388.5994 - 87448.1489s_1^2 - 87415.7371s_2^2 + 174835.0342s_1s_2$.

9.2. **Experimental Setups.**

9.2.1. *One-dimensional Deterministic Case.* When collecting trajectory data, entries of actions and initial states are sampled uniformly from the same interval. In each iteration, we collect 96 data points, obtained from 16 trajectories, where data is recorded at times $0, \Delta t, 2\Delta t, 3\Delta t, 4\Delta t$, and $5\Delta t$ along each trajectory. Additionally, we employ the same basis functions, where the basis for V is $\{s^2\}$ and the basis for Q is $\{a^2, sa, s^2\}$.

9.2.2. *One-dimensional Stochastic Case.* When collecting trajectory data, entries of actions and initial states are sampled uniformly from the same interval. In each iteration, we collect 9996 data points, obtained from 1666 trajectories, where data is recorded at times $0, \Delta t, 2\Delta t$, $3\Delta t$, $4\Delta t$ and $5\Delta t$ along each trajectory. Additionally, we employ the same basis functions, where the basis for V is $\{1, s^2\}$ and the basis for Q is $\{1, a^2, sa, s^2\}$.

9.2.3. *Two-dimensional Deterministic Case.* When collecting trajectory data, entries of actions and initial states are sampled uniformly from the same interval. In each iteration, we collect 100 data points, obtained from 25 trajectories, where data is recorded at times $0, \Delta t, 2\Delta t$, and $3\Delta t$ along each trajectory. Additionally, we employ the same basis functions, where the basis for V is $\{s_1^2, s_2^2, s_1s_2\}$ and the basis for Q is the union of $\{s_1s_2.s_1^2, s_2^2\}$, $\{s_1a_1, s_1a_2, s_2a_1, s_2a_2\}$ and $\{a_1a_2, a_1^2, a_2^2\}$.

9.2.4. *Two-dimensional Stochastic Case.* When collecting trajectory data, entries of actions and initial states are sampled uniformly from the same interval. In each iteration, we collect 6×10^4 data points, obtained from 10^4 trajectories, where data is recorded at times $0, \Delta t, 2\Delta t, 3\Delta t, 4\Delta t$, and $5\Delta t$ along each trajectory. Additionally, we employ the same basis functions, where the basis for V is $\{1, s_1^2, s_2^2, s_1s_2\}$ and the basis for Q is the union of $\{1\}$, $\{s_1s_2.s_1^2, s_2^2\}$, $\{s_1a_1, s_1a_2, s_2a_1, s_2a_2\}$ and $\{a_1a_2, a_1^2, a_2^2\}$.

Acknowledgements

Y. Zhu is supported by the NSF grants No 2529107. Y. P. Zhang acknowledges support from NSF CAREER grant DMS-2440215, Simons Foundation Travel Support MPS-TSM-00007305, and a start-up grant at Auburn University.

References

[1] B. D. Anderson and J. B. Moore. Optimal control: linear quadratic methods. Courier Corporation, 2007.
[2] L. 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 vol.4, 1994.
[3] E. Bayraktar and A. D. Kara. Approximate q learning for controlled diffusion processes and its near optimality. SIAM Journal on Mathematics of Data Science, 5(3):615–638, 2023.
[4] D. Bertsekas. Dynamic programming and optimal control: Volume I, volume 4. Athena scientific, 2012.
[5] D. P. Bertsekas. Approximate policy iteration: A survey and some new methods. Journal of Control Theory and Applications, 9(3):310–335, 2011.
[6] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In International Conference on Machine Learning, pages 1283–1294. PMLR, 2020.
[7] M. G. Crandall, H. Ishii, and P.-L. Lions. User's guide to viscosity solutions of second order partial differential equations. Bull. Amer. Math. Soc. (N.S.), 27(1):1–67, 1992.
[8] K. De Asis and R. S. Sutton. An idiosyncrasy of time-discretization in reinforcement learning. arXiv preprint arXiv:2406.14951, 2024.
[9] K. Doya. Reinforcement learning in continuous time and space. Neural computation, 12(1):219–245, 2000.
[10] W. H. Fleming and R. W. Rishel. Deterministic and Stochastic Optimal Control. Springer-Verlag, 1975.
[11] W. H. Fleming and H. M. Soner. Controlled Markov Processes and Viscosity Solutions. Springer, 2nd edition, 2006.
[12] X. Guo, H. V. Tran, and Y. P. Zhang. Policy iteration for nonconvex viscous hamilton–jacobi equations. arXiv preprint arXiv:2503.02159, 2025.
[13] R. A. Howard. Dynamic programming and markov processes. 1960.
[14] W. Hua, H. Mei, S. Zohar, M. Giral, and Y. Xu. Personalized dynamic treatment regimes in continuous time: a bayesian approach for optimizing clinical decisions with timing. Bayesian Analysis, 17(3):849–878, 2022.
[15] J. Jia, W. E, and Z. Li. Policy optimization for markovian jump processes: A path integral control approach. Journal of Machine Learning Research (JMLR), 22:1–50, 2021.
[16] Y. Jia and X. Y. Zhou. q-learning in continuous time. Journal of Machine Learning Research, 24(161):1–61, 2023.
[17] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In Conference on learning theory, pages 2137–2143. PMLR, 2020.
[18] A. Karimi, J. Jin, J. Luo, A. R. Mahmood, M. Jagersand, and S. Tosatto. Dynamic decision frequency with continuous options. In 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 7545–7552. IEEE, 2023.
[19] J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. The International Journal of Robotics Research, 32(11):1238–1274, 2013.
[20] V. Konda and J. Tsitsiklis. Actor-critic algorithms. Advances in neural information processing systems, 12, 1999.
[21] J. Lee and R. S. Sutton. Policy iterations for reinforcement learning problems in continuous time and space—fundamental theory and methods. Automatica, 126:109421, 2021.
[22] R. C. Merton. Lifetime portfolio selection under uncertainty: The continuous-time case. The Review of Economics and Statistics, 51(3):247–257, 1969.
[23] R. C. Merton. Optimum consumption and portfolio rules in a continuous-time model. In Stochastic optimization models in finance, pages 621–661. Elsevier, 1975.
[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2015.
[25] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani. Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems. Automatica, 50(1):193–202, 2014.
[26] J. Moody and M. Saffell. Learning to trade via direct reinforcement. IEEE Transactions on Neural Networks, 12(4):875–889, 2001.
[27] W. Mou and Y. Zhu. On bellman equations for continuous-time policy evaluation i: discretization and approximation. arXiv preprint arXiv:2407.05966, 2024.
[28] R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. Journal of Machine Learning Research, 9:815–857, 2008.
[29] S. A. Murphy. Optimal dynamic treatment regimes. Journal of the Royal Statistical Society Series B: Statistical Methodology, 65(2):331–355, 2003.
[30] G. A. Pavliotis. Stochastic processes and applications. Springer, 2016.
[31] S. Pradhan and S. Yuksel. Discrete-time approximations of controlled diffusions with infinite horizon discounted and average cost. arXiv preprint arXiv:2502.05596, 2025.

[32] M. L. Puterman and S. L. Brumelle. On the convergence of policy iteration in stationary dynamic programming. Mathematics of Operations Research, 4(1):60–69, 1979.
[33] D. Revuz and M. Yor. Continuous martingales and Brownian motion, volume 293. Springer Science & Business Media, 2013.
[34] M. S. Santos and J. Rust. Convergence properties of policy iteration. SIAM Journal on Control and Optimization, 42(6):2094–2115, 2004.
[35] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In International conference on machine learning, pages 1889–1897. PMLR, 2015.
[36] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.
[37] A. Sciarretta, M. Back, and L. Guzzella. Optimal control of parallel hybrid electric vehicles. IEEE Transactions on control systems technology, 12(3):352–363, 2004.
[38] B. Siciliano and L. Villani. Robot force control. Springer Science & Business Media, 1999.
[39] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. V. D. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. Nature, 529(7587):484–489, 2016.
[40] J.-G. Sun. Perturbation theory for algebraic riccati equations. SIAM Journal on Matrix Analysis and Applications, 19(1):39–65, 1998.
[41] R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction. MIT press, 2018.
[42] C. Tallec, L. Blier, and Y. Ollivier. Making deep q-learning methods robust to time discretization. In International Conference on Machine Learning, pages 6096–6104. PMLR, 2019.
[43] W. Tang, H. V. Tran, and Y. P. Zhang. Policy iteration for the deterministic control problems—a viscosity approach. SIAM Journal on Control and Optimization, 63(1):375–401, 2025.
[44] H. V. Tran, Z. Wang, and Y. P. Zhang. Policy iteration for exploratory hamilton–jacobi–bellman equations. Applied Mathematics & Optimization, 91(2):50, 2025.
[45] C. J. Watkins and P. Dayan. Q-learning. Machine learning, 8:279–292, 1992.
[46] L. Yang and M. Wang. Sample-optimal parametric q-learning using linearly additive features. In International conference on machine learning, pages 6995–7004. PMLR, 2019.
[47] C. Yildiz, M. Heinonen, and H. Lähdesmäki. Continuous-time model-based reinforcement learning. In International Conference on Machine Learning, pages 12009–12018. PMLR, 2021.
[48] J. Yong and X. Y. Zhou. Stochastic controls: Hamiltonian systems and HJB equations. Number 43 in Applications of mathematics. Springer Science & Business Media, New York, 1999.
[49] J. Zhang, C. Ni, C. Szepesvari, M. 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.
[50] Y. Zhu. Phibe: A pde-based bellman equation for continuous time policy evaluation. arXiv preprint arXiv:2405.12535, 2024.
[51] D. M. Ziegler, N. Stiennon, J. Wu, T. Brown, A. Radford, D. Amodei, and P. F. Christiano. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593, 2019.