REGRET OF EXPLORATORY POLICY IMPROVEMENT AND q-LEARNING

WENPIN TANG AND XUN YU ZHOU

Abstract. We study the convergence of q-learning and related algorithms introduced by Jia and Zhou (J. Mach. Learn. Res., 24 (2023), 161) for controlled diffusion processes. Under suitable conditions on the growth and regularity of the model parameters, we provide a quantitative error and regret analysis of both the exploratory policy improvement algorithm and the q-learning algorithm.

Key words: Continuous-time reinforcement learning, q-learning, regret analysis, entropy regularization, policy iteration/improvement, backward stochastic differential equation (BSDE), stochastic approximation.

1. Introduction

Reinforcement learning (RL) is an active subarea in machine learning, which has successfully been applied to solve complex decision-making problems such as playing board games [31, 32] and video games [22], driving autonomously [18, 21], and more recently, aligning large language models and text-to-image generative models with human preference [3, 7, 24]. RL research has predominantly focused on Markov decision processes (MDPs) in discrete time and space; see [34] for a detailed account of theory and applications for MDPs.

Wang, Zariphopoulou, and Zhou [40] are the first to formulate and develop an entropy-regularized, exploratory control framework for RL with controlled diffusion processes, which is inherently in continuous time with continuous state spaces and possibly continuous action (control) spaces. In this framework, stochastic relaxed control is utilized to represent exploration, capturing the notion of "trial and error" that is the core to RL. Subsequent work aims at laying theoretical foundation for model-free RL in continuous-time by a martingale approach [14, 15, 16], and by policy optimization [44]. Here, by "model-free" we mean that the underlying dynamics are diffusion processes but their coefficients along with the reward functions are unknown. The key insight of [14, 15, 16] is that one can derive learning objectives from the martingale structure underpinning the continuous-time RL. The theoretical results in those papers naturally lead to various "model-free" algorithms for general RL tasks, in the sense that they learn optimal policies directly without attempting to learn/estimate the model parameters. Many of these algorithms recover existing RL algorithms for MDPs that were often proposed in a heuristic manner. However, convergence and regret analysis of the algorithms, which has occupied a central stage in the RL study for MDPs, is still lacking for the diffusion counterpart. To our best knowledge, the only works that carry out a *model-free* convergence analysis and derive sublinear regrets are [12] for a class of stochastic linear–quadratic (LQ) control problems and [11] for continuous-time mean–variance portfolio selection, both of which apply/apapt the policy gradient algorithms developed in [15] and exploit heavily the special structures of the problems.

The purpose of the paper is to fill this gap by providing quantitative analysis of (little) *q-learning* introduced in [16] and related algorithms for generally nonlinear RL problems. While (big) Q-learning is a key method for discrete-time MDP RL, the Q-function collapses in continuous time as it no longer depends on actions when the time step is infinitesimally small. [16] proposes the notion of the q-function, which is the first-order derivative of the Q-function with respect to time discretization.

Date: November 2, 2024.

The q-function is entirely a continuous-time notion, which is closely related to the (generalized) Hamiltonian. The significance of this function lies in several aspects: 1) it is the function that needs to be learned in order to learn the optimal policies, rather than the individual functions appearing in the model; 2) it is the Gibbs exponent that can be used to improve the current policy; and 3) it is *learnable* (through an argument using Itô's formula) by observable/coumputable data including the temporal differences and the reward signals. This leads to q-learning, which alternates between stochastic approximation (to learn the value functions and the q-functions) and policy iteration/improvement (to update and improve the policies). The general q-learning theory and algorithms have already been applied to various specific problems, including in particular diffusion models for generative AI [9, 43].

The q-learning has two components:

$$\text{Learning values and policies} \;+\; \text{Policy iteration (PI)},$$

where "values" include those of the value functions and q-functions and "policies" are encoded by the q-functions. In the classical control setting, PI (or Howard's algorithm [1, 10]) improves the policies gradually in order to approximate the optimal one. So it is also called policy improvement; see [17, 16, 35] for recent studies. In the RL setting, an entropy regularizer is added to encourage exploration, which gives rise to the exploratory control problem [36]. When the model parameters are known (and the value function can be accessed), one can simply use a version of PI for the exploratory control problem. We call it *exploratory policy improvement*, which can be viewed as a model-based RL approach; see [19, 30, 38] for recent development. The model-free RL approach, on the other hand, consists of learning directly the value function and the optimal policy, without estimating the model parameters.

Our contribution is to provide a quantitative convergence analysis for both the model-based exploratory policy improvement and the model-free q-learning, while the former serves as a technical prerequisite for the latter. Specially, we prove:

- the exponential convergence of exploratory policy improvement (Theorem 3.2);
- an explicit error bound of q-learning, depending on the regularity of the model parameters and the learning rate (Theorems 4.7 and 4.8).

As an intermediate step to study the q-learning algorithm, we introduce *semi-q-learning* in which the value functions and the reward functions are assumed to be known, so we only need to learn the q-functions. We establish a bound on the value approximation in terms of the q-function approximation (Theorem 4.2), which is crucial for the convergence analysis of (semi-)q-learning.

Our proofs rely on both probabilistic and analytical arguments, including backward stochastic differential equations (BSDEs), partial differential equations (PDEs), perturbation analysis, and stochastic approximation. We hope that this work will trigger and motivate further research in continuous-time RL, and especially in its emerging application to diffusion model alignment [9, 43] and [37, Section 7.3].

The remainder of the paper is organized as follows. In Section 2, we provide background on the exploratory control problem and q-learning. In Section 3, we study exploratory policy improvement, and the q-learning algorithm is considered in Section 4. We conclude with Section 5.

2. Background

In this section, we provide background on continuous-time RL, and in particular q-learning, for controlled diffusion processes. We first introduce notations that will be used throughout.

- $\mathbb{R}$ is the set of real numbers. For $x, y \in \mathbb{R}$, $x \wedge y$ (resp. $x \vee y$) is the smaller (resp. larger) number of x and y.
- For a vector x, $|x|$ is the Euclidean norm of x; for $\mathcal{A} \subset \mathbb{R}^d$, $|\mathcal{A}|$ is the volume of $\mathcal{A}$.

- For a matrix A, A^T is the transpose of A, $\text{Tr}(A)$ is the trace of A, and $|A|$ is the spectral norm of A.
- For a function f on X, $|f|_\infty := \sup_X |f(x)|$ denotes its sup-norm, and $|f|_{\mathbb{L}^1(X)} := \int_X |f(x)|dx$ is the $\mathbb{L}^1$-norm of f.
- For $f : [0,\infty) \times \mathbb{R}^d \ni (t,x) \to \mathbb{R}$, $\partial_t f$ is its (partial) derivative in t, and $\nabla f = \left(\frac{\partial f}{\partial x_i}\right)_i$ and $\nabla^2 f = \left(\frac{\partial f}{\partial x_i \partial x_j}\right)_{i,j}$ are its gradient and Hessian in x respectively.
- For two probability density functions $p(\cdot)$ and $q(\cdot)$, $d_{TV}(p(\cdot), q(\cdot)) := \sup_A |\int_A p(a)da - \int_A q(a)da|$ is the total variation distance between $p(\cdot)$ and $q(\cdot)$.
- For a σ-field $\mathcal{F}$, $\mathbb{L}^2(\mathcal{F})$ is the set of $\mathcal{F}$-measurable square-integrable random variables.
- For any $\theta \in \mathbb{R}$ and $0 \le t \le T$, $\mathbb{H}_t^\theta := \{$predictable processes $(x_s,\, t \le s \le T)$ satisfying $\mathbb{E}\int_t^T e^{\theta s}|x_s|^2 ds < \infty\}$ with the $\mathbb{H}_t^\theta$-norm $|x|_{\mathbb{H}_t^\theta} := (\mathbb{E}\int_t^T e^{\theta s}|x_s|^2 ds)^{\frac{1}{2}}$.
- $a = \mathcal{O}(b)$ or $a \lesssim b$ means that a/b is bounded as some problem parameter tends to 0 or ∞.
- We use C for a generic constant whose values may change from line to line.

2.1. **Classical control problem.** Let $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\}_{t\ge 0})$ be a filtered probability space on which we define a d'-dimensional $\mathcal{F}_t$-adapted Brownian motion $(W_t,\, t \ge 0)$. Let $\mathcal{A}$ be a generic action space, and $u = (u_t,\, 0 \le t \le T)$ be a control which is an adapted process taking values in $\mathcal{A}$.

The stochastic control problem is to control the state variable $X_t \in \mathbb{R}^d$, whose dynamic is governed by the (controlled) stochastic differential equation (SDE):

$$dX_t^u = b(t, X_t^u, u_t)dt + \sigma(t, X_t^u, u_t)dW_t, \tag{2.1}$$

where $b : \mathbb{R}_+ \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^d$ is the drift coefficient, and $\sigma : \mathbb{R}_+ \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^{d\times d'}$ is the covariance matrix (or diffusion coefficient) of the state variable. Here the superscript 'u' in X_t^u emphasizes the dependence of the state variable on control u. The goal of the control problem is to maximize the total discounted reward (or objective functional), leading to the (optimal) value function:

$$\mathring{J}(t,x) := \sup_u \mathbb{E}\left[\int_t^T e^{-\beta(s-t)} r(s, X_s^u, u_s)ds + e^{-\beta(T-t)} h(X_T^u) \middle| X_t^u = x\right], \tag{2.2}$$

where $r : \mathbb{R}_+ \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}$ is the running reward, $h : \mathbb{R}^d \to \mathbb{R}$ is the terminal reward, and $\beta > 0$ is the discount factor.

In the classical setting, the functional forms of r, h, b, σ are known. The optimal control is generally represented as a deterministic mapping from the current time and state to the action space: $u_t^* = u^*(t, X_t^*)$. The mapping u^* is called an optimal feedback control *policy*, and the corresponding optimally controlled process $(X_t^*,\, t \ge 0)$ satisfies the following SDE:

$$dX_t^* = b(t, X_t^*, u^*(t, X_t^*))dt + \sigma(t, X_t^*, u^*(t, X_t^*))dW_t,$$

provided that it is well-posed. Optimal policies are derived by powerful approaches such as dynamic programming or maximum principle; see [8, 42] for detailed accounts of the classical stochastic control theory.

2.2. **Exploratory control problem for RL.** When the model parameters are unknown,[1] the RL approach explores the unknown environment and learns optimal controls through repeated trials and

[1] In many practical problems, the parameters b, σ are unknown while the rewards r, h may be specified in advance. The problem where r, h are unknown and need to be learned from the observed (optimal) actions is referred to as *inverse reinforcement learning* [23, 29].

errors. This calls for an essentially and drastically different school of thoughts from that of the classical control theory.

Inspired by this, [40] models exploration by a probability distribution of controls $\pi = (\pi_t(\cdot),\, 0 \le t \le T)$ over the action space $\mathcal{A}$ from which each trial is sampled. The exploratory state process is:

$$dX_t^\pi = \widetilde{b}(t, X_t^\pi, \pi_t)dt + \widetilde{\sigma}(t, X_t^\pi, \pi_t)d\widetilde{W}_t, \tag{2.3}$$

where $(\widetilde{W}_t,\, t \ge 0)$ is a d-dimensional $\mathcal{F}_t$-adapted Brownian motion, and the coefficients $\widetilde{b}(\cdot,\cdot,\cdot)$ and $\widetilde{\sigma}(\cdot,\cdot,\cdot)$ are defined by

$$\widetilde{b}(t,x,\pi) := \int_{\mathcal{A}} b(t,x,a)\pi(a)da \quad \text{and} \quad \widetilde{\sigma}(t,x,\pi) := \left(\int_{\mathcal{A}} \sigma(t,x,a)\sigma(t,x,a)^T \pi(a)da\right)^{\frac{1}{2}}, \tag{2.4}$$

for $(t,x,\pi) \in \mathbb{R}_+ \times \mathbb{R}^d \times \mathcal{P}(\mathcal{A})$, with $\mathcal{P}(\mathcal{A})$ being the set of absolutely continuous probability density functions on $\mathcal{A}$. The distributional control $\pi = (\pi_t(\cdot),\, 0 \le t \le T)$ is also known as the relaxed control in the control literature. The objective is to maximize the entropy-regularized problem:

$$\begin{aligned} J^*(t,x) &:= \sup_\pi J(t,x;\pi) \\ &:= \sup_\pi \mathbb{E}\left[\int_t^T e^{-\beta(s-t)} \int_{\mathcal{A}} [r(s,X_s^\pi,a) - \gamma \log \pi_s(a)]\pi_s(a)da\, ds + e^{-\beta(T-t)} h(X_T^\pi) \middle| X_t^\pi = x\right], \end{aligned} \tag{2.5}$$

where $J(\cdot,\cdot;\pi)$ denotes the value function under the control $\pi(\cdot)$, and $\gamma > 0$ is the (temperature) parameter representing the weight on exploration (so a greater γ encourages more exploration).

Let

$$H(t,x,a,p,q) := b(t,x,a)\,p + \frac{1}{2}\operatorname{Tr}\left(\sigma(t,x,a)\sigma^T(t,x,a)q\right) + r(t,x,a)$$

be the generalized Hamiltonian. According to [36, Section 5], under suitable conditions on the parameters (r,h,b,σ), J^* is the solution to the *exploratory* Hamilton–Jacobi–Bellman (HJB) equation:

$$\begin{cases} \partial_t J^* + \gamma \log \int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} H(t,x,a,\nabla J^*, \nabla^2 J^*)\right) da - \beta J^* = 0, \\ J^*(T,x) = h(x). \end{cases} \tag{2.6}$$

The corresponding optimal (exploratory) feedback control policy is the Gibbs measure:

$$\pi^*(\cdot \mid t,x) \propto \exp\left(\frac{1}{\gamma} H(t,x,\cdot,\nabla J^*, \nabla^2 J^*)\right). \tag{2.7}$$

See [36, 45] for further development on the exploratory control problem (2.5), and its HJB equation (2.6).

In the remainder of this paper, we assume that control only appears in the drift term.[2] That is, $\sigma(t,x,a) = \sigma(t,x)$ is independent of the control; so the resulting exploratory state process is governed by

$$dX_t^\pi = \widetilde{b}(t, X_t^\pi, \pi_t)dt + \sigma(t, X_t^\pi)d\widetilde{W}_t. \tag{2.8}$$

[2]Our analysis relies on the BSDE representation of the exploratory control problem, where control being only in the drift term corresponds to a semi-linear PDE. The case where control also appears in the diffusion term requires a more complicated stochastic representation of fully nonlinear PDEs; see [4, 25]. This is also related to the second-order risk adjustment of stochastic maximum principle when control enters into the diffusion term; see [26] and [42, Chapter 3]. We leave this case for future work.

2.3. **q-learning.** The main task of RL is to solve the entropy-regularized problem (2.5) by learning its optimal policy (2.7).

The first idea, based on the policy improvement theorems in [16, 41], is to approximate $\pi^*(\cdot\,|\,\cdot,\cdot)$ by a sequence of controls $\pi^n(\cdot\,|\,\cdot,\cdot)$ such that the corresponding value functions J^n are nondecreasing, i.e. $J^{n+1}(t,x) \geq J^n(t,x)$ for all n and all (t,x). The construction of the policies $\pi^n(\cdot\,|\,\cdot,\cdot)$ is through the iterations:

$$\pi^{n+1}(\cdot\,|\,t,x) \propto \exp\left(\frac{1}{\gamma}H(t,x,\cdot,\nabla J^n,\nabla^2 J^n)\right). \tag{2.9}$$

This approach is referred to as "exploratory policy improvement" aiming to approximate the optimal policy. However, the functions (b,σ,r,h) are unknown in the RL setting, whereas the sampling (2.9) requires an oracle access to them. Moreover, the derivatives of the value functions $(\nabla J^n, \nabla^2 J^n)$ are generally not easy to evaluate even if one can learn J^n through policy evaluation, which yields further difficulty in updating the policy.

The second idea, as proposed in [16], is to learn *directly* a term equivalent to $H(t,x,a,\nabla J^n,\nabla^2 J^n)$ for the purpose of sampling (2.9). That term is called the (little) q-function (analogous to the Q-function in the classical discrete-time RL for MDPs). Specifically, given a control $\pi(\cdot)$, the q-function is

$$\begin{aligned} q(t,x,a;\pi) :&= \frac{\partial J}{\partial t}(t,x;\pi) + H\left(t,x,a,\nabla J(t,x;\pi),\nabla^2 J(t,x;\pi)\right) - \beta J(t,x;\pi) \\ &= b(t,x,a)\nabla J(t,x;\pi) + r(t,x,a) + \frac{\partial J}{\partial t}(t,x;\pi) + \frac{1}{2}\operatorname{Tr}\left(\sigma^2(t,x)\nabla^2 J(t,x;\pi)\right) - \beta J(t,x;\pi). \end{aligned} \tag{2.10}$$

Denote by $a^\pi = (a^\pi_t,\, 0 \leq t \leq T)$ a control realization from sampling the policy $\pi(\cdot)$. The key takeaway in [16] is to observe the fact that

$$e^{-\beta s} J(s, X^\pi_s;\pi) + \int_t^s e^{-\beta u}\left[r(u,X^\pi_u,a^\pi_u) - q(u,X^\pi_u,a^\pi_u;\pi)\right]du, \quad t \leq s \leq T, \tag{2.11}$$

is a martingale. By parametrizing $(J(t,x;\pi), q(t,x,a;\pi))$ with $\{(J^\theta(t,x), q^\phi(t,x,a))\}_{\theta,\phi}$, the minimization of the martingale loss function motivates the following optimization problem:

$$\min_{(\theta,\phi)} \mathbb{E}\left[\int_0^T \left\{e^{-\beta(T-t)}h(X^\pi_T) - J^\theta(t,X^\pi_t) + \int_t^T e^{-\beta(s-t)}[r(s,X^\pi_s,a^\pi_s) - q^\phi(s,X^\pi_s,a^\pi_s)]ds\right\}^2 dt\right]. \tag{2.12}$$

Problem (2.12) can be solved by stochastic gradient descent (SGD):

$$\begin{aligned} \theta &\leftarrow \theta + \alpha_\theta \int_0^T \frac{\partial J^\theta}{\partial \theta}(t,X^\pi_t)G_{t:T}dt, \\ \phi &\leftarrow \phi + \alpha_\phi \int_0^T \int_t^T e^{-\beta(s-t)}\frac{\partial q^\phi}{\partial \phi}(s,X^\pi_s,a^\pi_s)dsG_{t:T}dt, \end{aligned}$$

where $G_{t:T} := e^{-\beta(T-t)}h(X^\pi_T) - J^\theta(t,X^\pi_t) + \int_t^T e^{-\beta(s-t)}[r(s,X^\pi_s,a^\pi_s) - q^\phi(s,X^\pi_s,a^\pi_s)]ds$, and $\alpha_\theta, \alpha_\phi$ are user-defined learning rates. Note that in actual implementation of this procedure one needs to discretize the integrals involved and takes actions (while observing the resulting states and reward signals) at the corresponding discretized time points. Thus, the update rule is entirely data driven without having to know/learn any model parameters.

Now we apply the above procedure to each policy iteration. Since only the term $H\left(t,x,a,\nabla J,\nabla^2 J\right)$ in (2.10) involves the control variable a, the policy update (2.9) is equivalent to

$$\pi^{n+1}(\cdot\,|\,t,x) \propto \exp\left(\frac{1}{\gamma}q(t,x,\cdot;\pi^n)\right).$$

Using the SGD step in each policy update yields the q-learning algorithm: start with some (θ_1, ϕ_1) and a control policy $\pi^1(\cdot \mid \cdot, \cdot)$, and for $n \ge 1$,

(1) Update

$$\begin{aligned} \theta_{n+1} &= \theta_n + \alpha_{\theta,n} \int_0^T \frac{\partial J^\theta}{\partial \theta}_{|\theta=\theta_n} (t, X_t^{\pi^n}) G_{t:T}^n dt, \\ \phi_{n+1} &= \phi_n + \alpha_{\phi,n} \int_0^T \int_t^T e^{-\beta(s-t)} \frac{\partial q^\phi}{\partial \phi}_{|\phi=\phi_n} (s, X_s^{\pi^n}, a_s^{\pi^n}) ds G_{t:T}^n dt \end{aligned} \tag{2.13}$$

where $G_{t:T}^n := e^{-\beta(T-t)} h(X_T^{\pi^n}) - J^{\theta_n}(t, X_t^{\pi^n}) + \int_t^T e^{-\beta(s-t)} [r(s, X_s^{\pi^n}, a_s^{\pi^n}) - q^{\phi_n}(s, X_s^{\pi^n}, a_s^{\pi^n})] ds$.

(2) Sample

$$\pi^{n+1}(\cdot \mid t, x) \propto \exp\left(\frac{1}{\gamma} q^{\phi_{n+1}}(t, x, \cdot)\right). \tag{2.14}$$

3. Convergence of exploratory policy improvement

In this section, we study the convergence of exploratory policy improvement by assuming an oracle access to some of the model parameters as well as the value functions. This allows us to understand how policy improvement itself works, and our result (Theorem 3.2) shows that it is exponentially fast. The arguments will then be used for the subsequent (model-free) analysis of the q-learning algorithm (2.13)–(2.14); so this section is also a technical preparation for later development.

Recall (2.8) that defines the exploratory state process X^π. The exploratory policy improvement starts with a control policy $\pi^1(\cdot \mid \cdot, \cdot)$, and for $n \ge 1$,

$$J^n(t,x) := \mathbb{E}\left[\int_t^T e^{-\beta(s-t)} \int_{\mathcal{A}} [r(s, X_s^{\pi^n}, a) - \gamma \log \pi^n(a|s, X_s^{\pi^n})] \pi^n(a|s, X_s^{\pi^n}) da\, ds + e^{-\beta(T-t)} h(X_T^{\pi^n}) \middle| X_t^{\pi^n} = x\right]. \tag{3.1}$$

Sample

$$\begin{aligned} \pi^{n+1}(\cdot \mid t, x) &\propto \exp\left(\frac{1}{\gamma} H(t, x, \cdot, \nabla J^n(t,x), \nabla^2 J^n(t,x))\right) \\ &\propto \exp\left(\frac{1}{\gamma}\left(b(t,x,\cdot)\nabla J^n(t,x) + r(t,x,\cdot)\right)\right), \end{aligned} \tag{3.2}$$

where we omit to write out the dependence of γ in the expressions of J^n and π^n for simplicity. Note that since $\sigma(t,x,a) = \sigma(t,x)$, only the terms $b(t,x,a)\nabla J^n(t,x)$ and $r(t,x,a)$ in the Hamiltonian involve the control a. This yields the (simpler) policy update (3.2).

As previously mentioned, we assume that the value function J^n defined in (3.1) (and hence its gradient ∇J^n) and the functions b, r are all accessed. This corresponds to the model-based RL approach; so we do not need to estimate these functions nor any of their combinations. Here, our goal is to bound

$$|J^n(t,x) - J^*(t,x)|, \quad \text{in terms of } n \text{ (and } \gamma),$$

where J^* is the optimal value function of the exploratory control problem (2.5). The convergence of J^n was proved in [13], but with no convergence rate given. After completing the present paper, we learned that two recent working papers [19, 38] studied the convergence rate of *model-based* policy iteration (3.2). Their results, the exponential convergence of J^n, are similar to Theorem 3.2 below. Nevertheless, our BSDE approach differs from [19, 38], where [38] is completely based on a PDE analysis, and [19] relies on Malliavin calculus. Moreover, our result provides an explicit dependence on the level of exploration γ, which is important in efficient sampling of the policies (3.2) but is missing in the aforementioned work. We stress again that the result in this section is a preparation for solving the "learning" problem in Section 4 that is at the heart of model-free q-learning, which neither of [19, 38] considered.[3]

To proceed, we make the following (mild) assumptions.

Assumption 3.1.

(i) $\mathcal{A}$ *is a compact set.*
(ii) r, h, b, σ *are continuous in their respective arguments.*
(iii) There exist $\overline{b}, \overline{\sigma}, \overline{r} > 0$ *such that for any* (t, x, a),

$$|b(t,x,a)| \le \overline{b}, \quad \frac{1}{\overline{\sigma}} \le |\sigma(t,x)| \le \overline{\sigma} \quad \text{and} \quad |r(t,x,a)| \le \overline{r}.$$

(iv) There exists $K > 0$ *such that for any* (t, x, y, a),

$$|b(t,x,a) - b(t,y,a)| + |\sigma(t,x) - \sigma(t,y)| + |r(t,x,a) - r(t,y,a)| + |h(x) - h(y)| \le K|x - y|.$$

Our result is stated as follows.

Theorem 3.2. *Let Assumption 3.1 hold, and fix* $\eta \in (0,1)$. *There exist* $L, C > 0$ *(independent of* γ *and* n*) such that*

$$|J^*(t,x) - J^n(t,x)|^2 \le C\eta^n e^{\theta(\gamma)(T-t)}, \tag{3.3}$$

where $\theta(\gamma) := \beta + (1 + \eta^{-1})L^2 \left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma} e^{\frac{L}{\gamma}}\right)^2$.

We make several remarks before proving the theorem. First, the bound (3.3) implies that the exploratory policy constructed by (3.2) converges exponentially in n. The bound depends explicitly on the level of exploration γ via the term $e^{\theta(\gamma)(T-t)}$. Observe that $\theta(\gamma)$ decreases in γ, which suggests to take a large γ for a faster convergence. In fact, there are two advantages of choosing a large γ:

- Larger γ induces faster convergence of the policy improvement (Theorem 3.2).
- Larger γ facilitates Markov chain Monte Carlo (MCMC) sampling of the policy $\pi^n(\cdot \,|\, t, x)$ ([5, Chapter 4], [20, Section 2.6]).[4]

Second, recall that $\mathring{J}$ is the optimal value function of the original, classical control problem (2.2). It is known [36, Corollary 4.7] that under suitable conditions on the parameters,

$$|J^*(t,x) - \mathring{J}(t,x)| \le C\gamma \ln(1/\gamma), \quad \text{as } \gamma \to 0^+. \tag{3.4}$$

Combining (3.3) and (3.4) yields

$$|J^n(t,x) - \mathring{J}(t,x)| \lesssim \underbrace{\gamma \ln(1/\gamma)}_{\text{bias}} + \underbrace{\eta^n e^{\theta(\gamma)(T-t)}}_{\text{policy improvement error}}, \tag{3.5}$$

giving rise to a tradeoff between bias (due to exploration) and policy improvement error in γ. Heuristically, minimizing the right side of (3.5) over γ leads to the equation

$$\frac{\ln(\gamma) + 1}{\theta'(\gamma)} e^{-\theta(\gamma)(T-t)} = C\eta^n (T-t)$$

where C is a constant independent of γ and n. This suggests that we could change the temperature parameter γ over iterations, specifically an *exploratory annealing* $\gamma = \gamma_n \downarrow 0$ (as $n \to \infty$), to get a

[3] In this sense, [19, 38] are "numerical" papers solving classical stochastic control problems, while ours is a "learning" paper solving the model-free counterparts.

[4] In general, the MCMC sampler of $\pi^n(\cdot \,|\, t, x)$ converges at a rate of $\exp(-e^{-\frac{1}{\gamma}} t)$ as the level of exploration $\gamma \to 0^+$. This implies for each iteration n, sampling $\pi^n(\cdot \,|\, t, x)$ requires $\mathcal{O}(e^{\frac{1}{\gamma}})$ time complexity, which is exponentially large if γ is chosen to be small.

shaper bound for $|J^n(t,x) - \mathring{J}(t,x)|$. Moreover, we can show (via an argument similar to that in the proof of Theorem 4.2) that

$$|J^n(t,x) - \mathring{J}(t,x)| \lesssim \sum_{k=1}^{n} \eta^{n-k} \gamma_k. \tag{3.6}$$

Thus, γ_n needs to be fast-decaying so that the right side of (3.6) converges to zero (a sufficient condition is $\sum_k \gamma_k < \infty$). On the other hand, MCMC sampling $\pi^n(\cdot \,|\, t,x)$ in each iteration has $\mathcal{O}(e^{\frac{1}{\gamma_n}})$ time complexity, which requires γ_n to decay slowly to avoid the curse of dimensionality. So the annealing will benefit the convergence of policy iteration yet at the cost of sampling hardness. For this reason, in this paper we focus on the *fixed* level of exploration $\gamma > 0$, leaving the study of exploratory annealing for future work.

3.1. **Auxiliary results.** This subsection collects a few useful results for proving Theorem 3.2. The next lemma, which follows from the Feynman–Kac formula, specifies the PDE satisfied by the value function J^n (see [15, Lemma 2]).

Lemma 3.3. *Under Assumption 3.1, J^n defined by (3.1) is the solution to*

$$\begin{cases} \partial_t J^n + \int_{\mathcal{A}} \left(H(t,x,a,\nabla J^n, \nabla^2 J^n) - \gamma \log \pi^n(a \,|\, t,x) \right) \pi^n(a \,|\, t,x)\, da - \beta J^n = 0, \\ J^n(T,x) = h(x). \end{cases} \tag{3.7}$$

The following result of the BSDEs will be used to bound $|J^n(t,x) - J^*(t,x)|$.

Lemma 3.4. *Let $F_t(y,z,Z)$ be a measurable function of (t,y,z,Z,ω) such that $(F_t(y,z,Z),\, 0 \le t \le T)$ is predictable for any fixed (y,z,Z) and $(F_t(0,0,0),\, 0 \le t \le T) \in \mathbb{H}^0_0$, and there exist $M_1, M_2, M_3 > 0$,*

$$|F_t(y',z',Z') - F_t(y,z,Z)| \le M_1|y'-y| + M_2|z'-z| + M_3|Z'-Z| \quad a.s.$$

Then we have

(i) For $\xi \in \mathbb{L}^2(\mathcal{F}_T)$, there exists a unique solution (Y,Z) to

$$Y_t = \xi + \int_t^T F_t(Y_s, Z_s, Z_s) ds - \int_t^T Z_s d\widetilde{W}_s, \quad 0 \le t \le T \quad a.s.$$

(ii) For $\xi \in \mathbb{L}^2(\mathcal{F}_T)$ and $z \in \mathbb{H}^0_0$, there exists a unique solution (Y,Z) to

$$Y_t = \xi + \int_t^T F_t(Y_s, z_s, Z_s) ds - \int_t^T Z_s d\widetilde{W}_s, \quad 0 \le t \le T \quad a.s. \tag{3.8}$$

(iii) Let (Y^1, Z^1) and (Y^2, Z^2) be the solutions to (3.8) corresponding to $z^1 \in \mathbb{H}^0_0$ and $z^2 \in \mathbb{H}^0_0$ respectively. Fixing $\eta \in (0,1)$, we have for $\theta \ge M_1 + (1+\eta^{-1})(M_2+M_3)^2$,

$$e^{\theta t} \mathbb{E}|Y^1_t - Y^2_t|^2 + |Z^1 - Z^2|^2_{\mathbb{H}^\theta_t} \le \eta |z^1 - z^2|^2_{\mathbb{H}^\theta_t}, \quad 0 \le t \le T.$$

Proof. (i) and (ii) follow from the standard BSDE theory; see, e.g., [27, Theorem 6.3.3], and (iii) from the proof of [17, Lemma A.5]. ☐

3.2. **Proof of Theorem 3.2.** We need a few lemmas to prove Theorem 3.2. Recall the definition of the optimal feedback control π^* from (2.7), and let X^{π^*} be the corresponding state process. Denote (with a slight abuse of notation)

$$\pi(a \,|\, s,x,z) \propto \exp\left(\frac{1}{\gamma} \left(b(s,x,a)\, \sigma^{-1}(s,x) z + r(s,x,a) \right) \right); \tag{3.9}$$

so we can rewrite

$$\pi^n(a \,|\, s,x) = \pi(a \,|\, s,x, \left(\sigma \nabla J^{n-1}\right)(s,x)) \quad \text{and} \quad \pi^*(a \,|\, s,x) = \pi(a \,|\, s,x, \left(\sigma \nabla J^*\right)(s,x)).$$

Define

$$G_s(z,Z) := \int_{\mathcal{A}} \Big[b(s, X_s^{\pi^*}, a)\, \sigma^{-1}(s, X_s^{\pi^*}) Z + r(s, X_s^{\pi^*}, a) \\ - \gamma \log \pi(a \,|\, s, X_s^{\pi^*}, z) \Big] \pi(a \,|\, s, X_s^{\pi^*}, z)\, da, \tag{3.10}$$

which is the key to our analysis. The next lemma studies the Lipschitz property of G_s.

Lemma 3.5. *Let Assumption 3.1 hold. Then for* $|z|, |Z| \le C$, *there exists* $L > 0$ *(independent of* γ*) such that*

$$|G_s(z', Z') - G_s(z, Z)| \le L \left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma} e^{\frac{L}{\gamma}} \right) |z' - z| + L|Z' - Z| \quad a.s. \tag{3.11}$$

Proof. Observe that

$$|G_s(z', Z') - G_s(z, Z)| \le E_1 + E_2 + E_3, \tag{3.12}$$

where

$$E_1 := \left| \int_{\mathcal{A}} r(s, X_s^{\pi^*}, a) \left(\pi(a \,|\, s, X_s^{\pi^*}, z') - \pi(a \,|\, s, X_s^{\pi^*}, z) \right) da \right|,$$
$$E_2 := \left| \int_{\mathcal{A}} b(s, X_s^{\pi^*}, a)\, \sigma^{-1}(s, X_s^{\pi^*}) \left(Z' \pi(a \,|\, s, X_s^{\pi^*}, z') - Z \pi(a \,|\, s, X_s^{\pi^*}, z) \right) da \right|,$$
$$E_3 := \gamma \left| \int_{\mathcal{A}} \left(\log \pi(a \,|\, s, X_s^{\pi^*}, z')\, \pi(a \,|\, s, X_s^{\pi^*}, z') - \log \pi(a \,|\, s, X_s^{\pi^*}, z)\, \pi(a \,|\, s, X_s^{\pi^*}, z) \right) da \right|.$$

By Assumption 3.1-(iii), we get

$$E_1 \le \overline{r} \left| \pi(\cdot \,|\, s, X_s^{\pi^*}, z') - \pi(\cdot \,|\, s, X_s^{\pi^*}, z) \right|_{\mathbb{L}^1(\mathcal{A})} = 2\overline{r}\, d_{TV} \left(\pi(\cdot \,|\, s, X_s^{\pi^*}, z'),\, \pi(\cdot \,|\, s, X_s^{\pi^*}, z) \right). \tag{3.13}$$

Recall that

$$\pi(a \,|\, s, x, z) = \frac{\exp\left(\frac{1}{\gamma} (b(s,x,a)\, \sigma^{-1}(s,x) z + r(s,x,a) \right)}{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} (b(s,x,a)\, \sigma^{-1}(s,x) z + r(s,x,a) \right) da}.$$

Applying [33, Theorem 8], [5] we get for $|z' - z| \le C$ (and hence $|z'| \le 2C$):

$$\begin{aligned} &d_{TV} \left(\pi(\cdot \,|\, s, x, z'),\, \pi(\cdot \,|\, s, x, z) \right) \\ &\le \frac{\exp\left(\frac{2}{\gamma} (\overline{b}\overline{\sigma} C + \overline{r}) \right)}{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} (b(s,x,a)\, \sigma^{-1}(s,x) z + r(s,x,a) \right) da} \left| \frac{1}{\gamma} b(s,x,a)\, \sigma^{-1}(t,x)(z' - z) \right|_{\mathbb{L}^1(\mathcal{A})} \\ &\le \frac{\exp\left(\frac{3}{\gamma} (\overline{b}\overline{\sigma} C + \overline{r}) \right)}{|\mathcal{A}|} \frac{\overline{b}\overline{\sigma} |z - z'| |\mathcal{A}|}{\gamma} = \frac{\overline{b}\overline{\sigma}}{\gamma} \exp\left(\frac{3}{\gamma} (\overline{b}\overline{\sigma} C + \overline{r}) \right) |z' - z|. \end{aligned}$$

For $|z' - z| > C$, it is obvious that $d_{TV} \left(\pi(\cdot \,|\, s, x, z'),\, \pi(\cdot \,|\, s, x, z) \right) \le |z' - z|/C$. Consequently,

$$d_{TV} \left(\pi(\cdot \,|\, s, x, z'),\, \pi(\cdot \,|\, s, x, z) \right) \le \left(\frac{\overline{b}\overline{\sigma}}{\gamma} \exp\left(\frac{3}{\gamma} (\overline{b}\overline{\sigma} C + \overline{r}) \right) + \frac{1}{C} \right) |z' - z|. \tag{3.14}$$

[5] For two Gibbs measures $\pi(da) = e^{\Psi(a)} da / Z$ and $\pi'(da) = e^{\Psi'(a)} da / Z'$, [33, Theorem 8] provides a bound on $d_{TV}(\pi(\cdot), \pi'(\cdot))$ under the assumption that $\sup \Psi = 0$. This assumption can be removed by modifying the bound to be

$$d_{TV}(\pi(\cdot), \pi'(\cdot)) \le \frac{e^{\sup_a (\Psi \vee \Psi')}}{Z} |\Psi - \Psi'|_{\mathbb{L}^1}.$$

Combining (3.13) and (3.14) yields

$$E_1 \le 2\overline{r}\left(\frac{\overline{b}\overline{\sigma}}{\gamma}\exp\left(\frac{3}{\gamma}(\overline{b}\overline{\sigma}C+\overline{r})\right)+\frac{1}{C}\right)|z'-z|. \tag{3.15}$$

Similarly, we get

$$\begin{aligned}E_2 &\le \overline{b}\overline{\sigma}\left(\int_{\mathcal{A}}|Z'-Z|\,\pi(a\,|\,s,X_s^{\pi^*},z')\,da+\int_{\mathcal{A}}|Z|\,|\pi(a\,|\,s,X_s^{\pi^*},z')-\pi(a\,|\,s,X_s^{\pi^*},z)|da\right)\\ &\le \overline{b}\overline{\sigma}\left(|Z'-Z|+2C\,d_{TV}\left(\pi(\cdot\,|\,s,X_s^{\pi^*},z'),\,\pi(\cdot\,|\,s,X_s^{\pi^*},z)\right)\right)\\ &\le \overline{b}\overline{\sigma}|Z'-Z|+\left(\frac{2\overline{b}\overline{\sigma}C}{\gamma}\exp\left(\frac{3}{\gamma}(\overline{b}\overline{\sigma}C+\overline{r})\right)+2\overline{b}\overline{\sigma}\right)|z'-z|.\end{aligned} \tag{3.16}$$

Finally, we have

$$\begin{aligned}E_3 &\le \underbrace{\left|\int_{\mathcal{A}}r(s,X_s^{\pi^*},a)\left(\pi(a\,|\,s,X_s^{\pi^*},z')-\pi(a\,|\,s,X_s^{\pi^*},z)\right)da\right|}_{(a)}\\ &\quad+\underbrace{\left|\int_{\mathcal{A}}b(s,X_s^{\pi^*},a)\,\sigma^{-1}(s,X_s^{\pi^*})\left(z'\pi(a\,|\,s,X_s^{\pi^*},z')-z\pi(a\,|\,s,X_s^{\pi^*},z)\right)da\right|}_{(b)}\\ &\quad+\gamma\underbrace{\left|\log\left(\frac{\int_{\mathcal{A}}\exp\left(\frac{1}{\gamma}(b(s,X_s^{\pi^*},a)\,\sigma^{-1}(s,X_s^{\pi^*})z'+r(s,X_s^{\pi^*},a)))\right)da}{\int_{\mathcal{A}}\exp\left(\frac{1}{\gamma}(b(s,X_s^{\pi^*},a)\,\sigma^{-1}(s,X_s^{\pi^*})z+r(s,X_s^{\pi^*},a)))\right)da}\right)\right|}_{(c)}\\ &\le\left(3\overline{b}\overline{\sigma}+\frac{2\overline{b}\overline{\sigma}(C+\overline{r})}{\gamma}\exp\left(\frac{3}{\gamma}(\overline{b}\overline{\sigma}C+\overline{r})\right)+\frac{2\overline{r}}{C}\right)|z'-z|+(c),\end{aligned} \tag{3.17}$$

where (a) is bounded by E_1, and (b) is bounded by E_2 with $(Z,Z')=(z,z')$. Now we proceed to bounding the term (c). For $|z'|>2C$, we have $|z'-z|>C$. Thus,

$$\begin{aligned}\left|\log\left(\frac{\int_{\mathcal{A}}\exp\left(\frac{1}{\gamma}(b(s,x,a)\,\sigma^{-1}(s,x)z'+r(s,x,a))\right)da}{\int_{\mathcal{A}}\exp\left(\frac{1}{\gamma}(b(s,x,a)\,\sigma^{-1}(s,x)z+r(s,x,a))\right)da}\right)\right| &\le \frac{\overline{b}\overline{\sigma}(|z'|+C)+2\overline{r}}{\gamma}\\ &\le \frac{\overline{b}\overline{\sigma}|z'-z|+2(\overline{b}\overline{\sigma}C+\overline{r})}{\gamma}\\ &\le \frac{3\overline{b}\overline{\sigma}+2\overline{r}/C}{\gamma}|z'-z|.\end{aligned} \tag{3.18}$$

Now consider the case when $|z'|\le 2C$. Since $|\log(u'/u)|\le|u'-u|/(u\wedge u')$ for $u,u'\ge 0$, we get

$$\left|\log\left(\frac{\int_{\mathcal{A}}\exp\left(\frac{1}{\gamma}(b(s,x,a)\,\sigma^{-1}(s,x)z'+r(s,x,a))\right)da}{\int_{\mathcal{A}}\exp\left(\frac{1}{\gamma}(b(s,x,a)\,\sigma^{-1}(s,x)z+r(s,x,a))\right)da}\right)\right|\le\frac{(d)}{(e)}, \tag{3.19}$$

where

$$(d):=\int_{\mathcal{A}}\left|\exp\left(\frac{1}{\gamma}(b(s,x,a)\,\sigma^{-1}(s,x)z'+r(s,x,a))\right)-\exp\left(\frac{1}{\gamma}(b(s,x,a)\,\sigma^{-1}(s,x)z+r(s,x,a))\right)\right|da,$$

$$(e):=\int_{\mathcal{A}}\exp\left(\frac{1}{\gamma}(b(s,x,a)\,\sigma^{-1}(s,x)z'+r(s,x,a))\right)da\wedge\int_{\mathcal{A}}\exp\left(\frac{1}{\gamma}(b(s,x,a)\,\sigma^{-1}(s,x)z+r(s,x,a))\right)da.$$

By [33, Theorem 5], we have

$$(d) \le \frac{\bar{b}\bar{\sigma}|\mathcal{A}|}{\gamma} \exp\left(\frac{1}{\gamma}(\bar{b}\bar{\sigma}C + \bar{r})\right)|z'-z|, \quad (e) \ge |\mathcal{A}| \exp\left(-\frac{2\bar{b}\bar{\sigma}C + \bar{r}}{\gamma}\right). \tag{3.20}$$

The estimates (3.18)–(3.20) yield

$$(c) \le \left(\bar{b}\bar{\sigma} \exp\left(\frac{3\bar{b}\bar{\sigma}C + 2\bar{r}}{\gamma}\right) + 3\bar{b}\bar{\sigma} + \frac{2\bar{r}}{C}\right)|z'-z|. \tag{3.21}$$

By (3.17) and (3.21), we obtain

$$E_3 \le \left(6\bar{b}\bar{\sigma} + \frac{4\bar{r}}{C} + \bar{b}\bar{\sigma}\left(1 + \frac{2(C+\bar{r})}{\gamma}\right)\exp\left(\frac{3}{\gamma}(\bar{b}\bar{\sigma}C + \bar{r})\right)\right)|z'-z|. \tag{3.22}$$

Combining (3.12) with (3.15), (3.16) and (3.22) yields (3.11). □

The following result establishes representations for J^n and J^*.

Lemma 3.6. *Let Assumption 3.1 hold. There exist $(\widehat{W}, \widehat{\mathbb{P}})$, with $\widehat{W}$ being $\widehat{\mathbb{P}}$-Brownian motion, such that*

(i) We have

$$J^n(t,x) = h(X_T^{\pi^*}) + \int_t^T [G_s(Z_s^{n-1}, Z_s^n) - \beta Y_s^n]\, ds - \int_t^T Z_s^n d\widehat{W}_s, \tag{3.23}$$

where $Y_s^n := J^n(s, X_s^{\pi^*})$ and $Z_s^n := (\sigma\,\nabla J^n)(s, X_s^{\pi^*})$.

(ii) The BSDE

$$Y_s^{t,x} = h(X_T^{\pi^*}) + \int_s^T [G_u(Z_u^{t,x}, Z_u^{t,x}) - \beta Y_u^{t,x}]\, du - \int_s^T Z_u^{t,x} d\widehat{W}_u, \quad t \le s \le T, \tag{3.24}$$

has a unique solution $(Y^{t,x}, Z^{t,x})$ (the superscript stands for $X_t^{\pi^*} = x$). Moreover, $J^*(t,x) = Y_t^{t,x}$.

Proof. (i) Applying Itô's formula to $J^n(s, X_s^{\pi^*})$ and noting Lemma 3.3, we get

$$\begin{aligned}
&dJ^n(s, X_s^{\pi^*}) \\
&= \left[\partial_t J^n(s, X_s^{\pi^*}) + \frac{1}{2}\operatorname{Tr}(\sigma\sigma^T\nabla^2 J^n)(s, X_s^{\pi^*}) + \widetilde{b}(s, X_s^{\pi^*}, \pi^*)\,\nabla J^n(s, X_s^{\pi^*})\right] ds + (\sigma\,\nabla J^n)(s, X_s^{\pi^*}) d\widetilde{W}_s \\
&= \Big[\widetilde{b}(s, X_s^{\pi^*}, \pi^*)\,\nabla J^n(s, X_s^{\pi^*}) - \widetilde{b}(s, X_s^{\pi^*}, \pi^n)\,\nabla J^n(s, X_s^{\pi^*}) - \widetilde{r}(s, X_s^{\pi^*}, \pi^n) \\
&\quad + \gamma \int_{\mathcal{A}} \log \pi^n(a \,|\, s, X_s^{\pi^*})\, \pi^n(a \,|\, s, X_s^{\pi^*}) da + \beta J^n(s, X_s^{\pi^*})\Big] ds + (\sigma\,\nabla J^n)(s, X_s^{\pi^*}) d\widetilde{W}_s,
\end{aligned} \tag{3.25}$$

where π^* and π^n appearing in $\widetilde{b}\left(s, X_s^{\pi^*}, \cdot\right)$, $\widetilde{r}\left(s, X_s^{\pi^*}, \cdot\right)$ stand for $\pi^*(\cdot \,|\, s, X_s^{\pi^*})$ and $\pi^n(\cdot \,|\, s, X_s^{\pi^*})$ respectively.

Let $\widehat{\mathbb{P}}$ be a change of measure of $\mathbb{P}$, with the Radon–Nikodym derivative:

$$\frac{d\widehat{\mathbb{P}}}{d\mathbb{P}} = \mathcal{E}\left(-\int_0^{\cdot} \widetilde{b}(s, X_s^{\pi^*}, \pi^*)\,\sigma^{-1}(s, X_s^{\pi^*})\, d\widetilde{W}_s\right)_T,$$

where $\mathcal{E}(\cdot)$ is the stochastic exponential (see [28, p.328]). So $\widehat{W}_t := \widetilde{W}_t + \int_0^t \widetilde{b}(s, X_s^{\pi^*}, \pi^*)\,\sigma^{-1}(s, X_s^{\pi^*})\, ds$ is a $\widehat{\mathbb{P}}$-Brownian motion.

Set $Y_s^n := J^n(s, X_s^{\pi^*})$ and $Z_s^n := (\sigma \nabla J^n)(s, X_s^{\pi^*})$; so $Y_t^n = J^n(t,x)$. By integrating (3.25) on $[t,T]$ and noting that $\pi^n(a \,|\, s, X_s^{\pi^*}) = \pi(a \,|\, s, X_s^{\pi^*}, Z_s^{n-1})$, we get

$$Y_t^n = g(X_T^{\pi^*}) - \int_t^T \left[\widetilde{b}(s, X_s^{\pi^*}, \pi^*)\, \sigma^{-1}(s, X_s^{\pi^*})\, Z_s^n - G_s(Z_s^{n-1}, Z_s^n) + \beta Y_s^n\right] ds - \int_t^T Z_s^n d\widetilde{W}_s,$$

which leads to (3.23).

(ii) The fact that the BSDE (3.24) has a unique solution follows from Lemma 3.4-(i) noting Lemma 3.5. We have then

$$\begin{aligned} dY_s^{t,x} = \Bigg(&\widetilde{b}(s, X_s^{\pi^*}, \pi^*)\sigma^{-1}(s, X_s^{\pi^*}) Z_s^{t,x} + \beta Y_s^{t,x} \\ &- \gamma \log \int_{\mathcal{A}} \exp\left(\frac{1}{\gamma}(b(u, X_u^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*}) Z_s^{t,x} + r(u, X_u^{\pi^*}, a))\right) da\Bigg) ds + Z_s^{t,x} d\widetilde{W}_s. \end{aligned}$$

A standard BSDE argument (see [6, Proposition 4.3] or [27, Section 6.3]) shows that $Y_t^{t,x} = v(t,x)$ is the solution to the semi-linear PDE:

$$\begin{cases} \partial_t v + \frac{1}{2}\operatorname{Tr}\left(\sigma\sigma^T \nabla^2 v\right) + \gamma \log \int_{\mathcal{A}} \exp\left(\frac{1}{\gamma}(b(t,x,a)\nabla v + r(t,x,a))\right) da - \beta v = 0, \\ v(T,x) = h(x). \end{cases}$$

Note that $\gamma \int_{\mathcal{A}} H(t,x,a,\nabla v, \nabla^2 v)\, da = \frac{1}{2}\operatorname{Tr}\left(\sigma\sigma^T \nabla^2 v\right) + \gamma \log \int_{\mathcal{A}} \exp\left(\frac{1}{\gamma}(b(t,x,a)\nabla v + r(t,x,a))\right) da$. Thus, $Y_t^{t,x}$ is the solution to the HJB equation (2.6), i.e. $Y_t^{t,x} = J^*(t,x)$. □

Now we prove Theorem 3.2.

Proof of Theorem 3.2. Let $F_s(Y,z,Z) := G_s(z,Z) - \beta Y$. It follows from Lemma 3.6 that $J^n(t,x) = Y_t^n$ and $J^*(t,x) = Y_t^{t,x}$ where

$$\begin{aligned} Y_s^n &= h(X_T^{\pi^*}) + \int_s^T F_u(Y_u^n, Z_u^{n-1}, Z_u^n)\, du - \int_s^T Z_u^n d\widehat{W}_u, \\ Y_s^{t,x} &= h(X_T^{\pi^*}) + \int_s^T F_u(Y_u^{t,x}, Z_u^{t,x}, Z_u^{t,x})\, du - \int_s^T Z_u^{t,x} d\widehat{W}_u. \end{aligned}$$

Note that $|\nabla J^*| \le M$ (with M independent of γ; see [36]), and $Z_s^{t,x} = (\sigma \nabla J^*)(s, X_s^{\pi^*})$. By Lemma 3.5, we get

$$\begin{aligned} &|F_u(Y_u^n, Z_u^{n-1}, Z_u^n) - F_u(Y_u^{t,x}, Z_u^{t,x}, Z_u^{t,x})| \\ &\quad \le \beta |Y_u^n - Y_u^{t,x}| + L\left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma} e^{\frac{L}{\gamma}}\right) |Z_u^{t,x} - Z_u^{n-1}| + L|Z_u^{t,x} - Z_u^n|. \end{aligned}$$

Applying Lemma 3.4-(iii) with $M_1 = \beta$, $M_2 = L\left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)$ and $M_3 = L$, we have for $\theta \ge \beta + (1+\eta^{-1})L^2\left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)^2$,

$$e^{\theta t}\widehat{\mathbb{E}}|Y_t^{t,x} - Y_t^n|^2 + |Z^{t,x} - Z^n|^2_{\widehat{\mathbb{H}}_t^\theta} \le \eta |Z^{t,x} - Z^{n-1}|^2_{\widehat{\mathbb{H}}_t^\theta},$$

where $\widehat{\mathbb{H}}_t^\theta$ is defined under $\widehat{\mathbb{P}}$. So we get $|Z^{t,x} - Z^n|^2_{\widehat{\mathbb{H}}_t^\theta} \le \eta^n |Z^{t,x} - Z^0|^2_{\widehat{\mathbb{H}}_t^\theta}$, and

$$\begin{aligned} |J^*(t,x) - J^n(t,x)|^2 &\le \eta^n \widehat{\mathbb{E}} \int_t^T e^{\theta(s-t)} |\sigma(s, X_s^{\pi^*})|^2 |\nabla J^*(s, X_s^{\pi^*}) - \nabla J^0(s, X_s^{\pi^*})|^2 ds \\ &\le \eta^n e^{\theta(T-t)} \widehat{\mathbb{E}} \int_t^T |\sigma(s, X_s^{\pi^*})|^2 |\nabla J^*(s, X_s^{\pi^*}) - \nabla J^0(s, X_s^{\pi^*})|^2 ds \end{aligned}$$

which proves (3.3). □

4. q-LEARNING AND REGRET

In this section, we study the q-learning algorithm (2.13)–(2.14) and derive its regret. As an intermediate step, we consider a "semi"-q-learning algorithm in Section 4.1, where value functions of given policies can be accessed and we only need to learn the q-function. The advantage of studying the semi-q-learning first is to present the essential idea while simplifying notation. We will show how to adapt the arguments in Section 3 to derive the convergence rate of the semi-q-learning algorithm. The result for the q-learning algorithm is then provided in Section 4.2.

4.1. **Semi-q-learning.**

Here we assume an oracle access to the value functions J^n, and the resulting algorithm is called the semi-q-learning. We start with some initial parameter value ϕ^1 for the q-function and a control policy $\pi^1(\cdot\,|\,\cdot,\cdot)$, and for $n \geq 1$,

(1) Update

$$\phi_{n+1} = \phi_n + \alpha_{\phi,n} \int_0^T \int_t^T e^{-\beta(s-t)} \frac{\partial q^\phi}{\partial \phi}_{|\phi=\phi_n}(s, X_s^{\pi^n}, a_s^{\pi^n}) ds G_{t:T}^n dt, \tag{4.1}$$

where $G_{t:T}^n := e^{-\beta(T-t)} h(X_T^{\pi^n}) - J^n(t, X_t^{\pi^n}) + \int_t^T e^{-\beta(s-t)}[r(s, X_s^{\pi^n}, a_s^{\pi^n}) - q^{\phi_n}(s, X_s^{\pi^n}, a_s^{\pi^n})] ds$.

(2) Sample

$$\pi^{n+1}(a\,|\,t,x) \propto \exp\left(\frac{1}{\gamma} q^{\phi_{n+1}}(t,x,a)\right) da. \tag{4.2}$$

Recall (3.1) that defines J^n and it follows from (2.10) that

$$q(t,x,a;\pi^n) = b(t,x,a)\nabla J^n(t,x) + r(t,x,a) + \frac{\partial J^n}{\partial t}(t,x) + \frac{1}{2}\operatorname{Tr}\left(\sigma^2(t,x)\nabla^2 J^n(t,x)\right) - \beta J^n(t,x).$$

To avoid undue technicality, we assume that $\nabla J^n \neq 0$ almost everywhere.[6] Let $b^{n+1}(t,x,a)$ be a measurable function such that the following holds almost everywhere[7]:

$$q^{\phi_{n+1}}(t,x,a) = b^{n+1}(t,x,a)\nabla J^n(t,x) + r(t,x,a) + \frac{\partial J^n}{\partial t}(t,x) + \frac{1}{2}\operatorname{Tr}\left(\sigma^2(t,x)\nabla^2 J^n(t,x)\right) - \beta J^n(t,x). \tag{4.3}$$

The policy update (4.2) can now be written as:

$$\pi^{n+1}(a\,|\,t,x) \propto \exp\left(\frac{1}{\gamma}(b^{n+1}(t,x,a)\nabla J^n(t,x) + r(t,x,a))\right) da. \tag{4.3'}$$

As $q^{\phi_{n+1}}(t,x,a)$ is expected to be close to $q(t,x,a;\pi^n)$ when n is large, so is $b^n(t,x,a)$ to $b(t,x,a)$. To present our result, we need the following condition in addition to Assumption 3.1.

Assumption 4.1. *There exists $\bar{b} > 0$ such that $\operatorname{ess\,sup}_{t,x,a} |b^n| \leq \bar{b}$ for all n.*

The result below shows how the convergence of the value functions depends in an explicit way on that of the q values.

Theorem 4.2. *Let Assumptions 3.1 and 4.1 hold, and assume that there is $M > 0$ such that $|\nabla J^n| \leq M$ for all n. Fix $\eta \in (0,1)$. There exist L, $C > 0$ (independent of γ and n) such that*

$$|J^*(t,x) - J^n(t,x)|^2 \leq C\left(\eta^n e^{\Lambda(\gamma)(T-t)} + \left(1 + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)\sum_{k=1}^n \eta^{n-k} |q(\cdot;\pi^k) - q^{\phi_{k+1}}|_\infty\right), \tag{4.4}$$

---

[6]It is known [41] that for linear–quadratic (LQ) problems, $\nabla J^n(t,x) = A_n(t)x + B_n(t)$ with $A_n(t) > 0$ for a suitably chosen initial policy. This satisfies $\nabla J^n(t,x) \neq 0$ almost everywhere. The example presented immediately after Corollary 4.5 is another instance satisfying this assumption.

[7]Because $\nabla J^n \neq 0$ almost everywhere, $b^n(t,x,a)$ is uniquely determined if $d = 1$, and are solutions to the underdetermined system (4.3) if $d \geq 2$. The measurability follows from standard measurable selection theory, e.g., [39].

where $\Lambda(\gamma) := \beta + (1+\eta^{-1})L^2 \left[1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma}e^{\frac{L}{\gamma}} + L\left(1+\frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)\sup_n |q(\cdot;\pi^n) - q^{\phi_{n+1}}|_\infty\right]^2$.

A proof of Theorem 4.2 is delayed to Section 4.3, which is a "perturbed" variant of the arguments in Section 3. Here let us make several comments. First, the assumption $|\nabla J^n| \le M$ for all n is again to avoid excessive technicality. Because ∇J^* is bounded and J^n is expected to be a proxy to J^*, it is reasonable to assume that ∇J^n is uniformly bounded.

Second, the theorem quantifies the performance of the algorithm by how well the q values are learned. We illustrate this with two cases:

- If $\inf_n |q(\cdot;\pi^n) - q^{\phi_{n+1}}|_\infty := \delta > 0$, i.e. the q values are not well approximated, then the right side of (4.4) is bounded from below by $\sqrt{1-\eta^n} \ge 1 - \eta^n$, yielding at least a linear regret.
- If the q values are well learned, say $|q(\cdot;\pi^n) - q^{\phi_{n+1}}|_\infty \approx n^{-\alpha}$ for some $\alpha > 0$, then (4.4) specializes to:

$$|J^*(t,x) - J^n(t,x)|^2 \lesssim \sum_{k=1}^n \eta^{n-k}k^{-\alpha} \lesssim n^{-\alpha}\ln n,$$

  where the second inequality is obtained by splitting the sum into $[1, n-\ln n]$ and $[n-\ln n, n]$ and by taking η to be sufficiently small. The regret now is sublinear:

$$\sum_{k=1}^n |J^*(t,x) - J^k(t,x)| \lesssim \begin{cases} n^{1-\frac{\alpha}{2}}(\ln n)^{\frac{1}{2}} & \text{if} \quad \alpha < 2, \\ (\ln n)^{\frac{3}{2}} & \text{if} \quad \alpha = 2, \\ \mathcal{O}(1) & \text{if} \quad \alpha > 2. \end{cases} \tag{4.5}$$

Next we consider the convergence of the q values via the iteration (4.1). This is an instance of the Robbins–Monro algorithm, which can be written as:

$$\phi_{n+1} = \phi_n + \alpha_{\phi,n}\mathcal{H}(\phi_n, \underbrace{(X_t^{\pi^n}), (a_t^{\pi^n})}_{U_{n+1}}), \tag{4.6}$$

where

$$\mathcal{H}(\phi, U) := \int_0^T \int_t^T e^{-\beta(s-t)} \frac{\partial q^\phi}{\partial \phi}(s, X_s^{\pi^\phi}, a_s^{\pi^\phi}) ds G_{t:T}^\phi dt, \tag{4.7}$$

with $U = ((X_t^{\pi^\phi}), (a_t^{\pi^\phi}))$, $\pi^\phi(a\,|\,t,x) \propto \exp\left(\frac{1}{\gamma}q^\phi(t,x,a)\right)$ and $G_{t:T}^\phi := e^{-\beta(T-t)}h(X_T^{\pi^\phi}) - J(t, X_t^{\pi^\phi}; \pi^\phi) + \int_t^T e^{-\beta(s-t)}[r(s, X_s^{\pi^\phi}, a_s^{\pi^\phi}) - q^\phi(s, X_s^{\pi^\phi}, a_s^{\pi^\phi})]ds$. Clearly, $(U_{n+1}\,|\,U_n, \dots, U_1, \phi_n, \dots, \phi_1) \overset{d}{=} (U_{n+1}\,|\,\phi_n)$. Also define

$$h(\phi) := \mathbb{E}\mathcal{H}(\phi, U). \tag{4.8}$$

We need a few more assumptions to study the convergence rate of the stochastic approximation (4.6) (or (4.1)).

Assumption 4.3.

(1) The ODE $\phi'(t) = h(\phi(t))$ has a unique stable equilibrium ϕ^*.

(2) There exists $C > 0$ such that $\mathbb{E}(|\mathcal{H}(\phi_n, U_{n+1})|^2\,|\,\phi_n) \le C(1 + |\phi_n|^2)$.

(3) There exists $\kappa > 0$ such that $(\phi - \phi^*)h(\phi) \le -\kappa|\phi - \phi^*|^2$ in a neighborhood of ϕ^*.

(4) There exist $\rho_\phi, C > 0$ such that $|q^\phi - q^{\phi'}|_\infty \le C|\phi - \phi'|^{\rho_\phi}$ for all ϕ, ϕ'.

(5) There exists $C > 0$ such that $|q(\cdot;\pi^\phi) - q(\cdot;\pi^*)|_\infty \le C d_{TV}(\pi^\phi, \pi^*)$ for ϕ in a neighborhood of ϕ^*.

Set $\Delta := |q(\cdot;\pi^*) - q^{\phi^*}(\cdot)|_\infty$. The value of Δ quantifies how close the family of functions $\{q^\phi\}_\phi$ approximates the optimal q function. When $\Delta = 0$ (if rarely the case), the family $\{q^\phi\}_\phi$ is rich enough to contain the optimal q function. The condition (1) ensures the optimal ϕ^* is the only candidate for the stochastic approximation. The conditions (2)–(5) will be used to quantify the convergence rate of the q values, where (2)–(3) are growth conditions of the SGD (4.6), (4) imposes Hölder regularity of the function approximation $\{q^\phi\}_\phi$, and (5) specifies the sensitivity of the q function with respect to the (stochastic) policies. Later we will give an example in which these conditions are satisfied.

The following result provides an error estimate of the semi-q-learning (**??**)–(4.2).

Theorem 4.4. *Let the assumptions in Theorem 4.2 and Assumption 4.3 hold. Set* $\alpha_{\phi,n} = \frac{A}{n^\nu + B}$ *for some* $\nu \le 1$, $A > \frac{\alpha}{2\kappa}$ *and* $B > 0$, *and let* $\varepsilon > 0$. *Then there exists* $C > 0$ *(independent of* n, ε*) such that with probability* $1 - \varepsilon$,

$$|J^*(t,x) - J^n(t,x)| \le C\left(\Delta + \frac{1}{\varepsilon^{\rho_\phi/2}} n^{-\frac{\nu\rho_\phi}{4}} (\ln n)^{\frac{1}{2}}\right), \tag{4.9}$$

for Δ *sufficiently small. Consequently, the regret is:*

$$\sum_{k=1}^n |J^*(t,x) - J^k(t,x)| \lesssim \Delta n + \frac{1}{\varepsilon^{\rho_\phi/2}} \begin{cases} n^{1-\frac{\nu\rho_\phi}{4}} (\ln n)^{\frac{1}{2}} & \text{if} \quad \nu\rho_\phi < 4, \\ (\ln n)^{\frac{3}{2}} & \text{if} \quad \nu\rho_\phi = 4, \\ \mathcal{O}(1) & \text{if} \quad \nu\rho_\phi > 4. \end{cases} \tag{4.10}$$

A proof of Theorem 4.4 will be given in Section 4.3. As previously mentioned, the error term Δ in (4.9) (or (4.10)) comes from function approximations, which is typically related to the quality of neural nets used, something that is not dictated or controlled by the q-learning in general. If the family of functions $\{q^\phi\}_\phi$ include the optimal q function, then we get the following corollary.

Corollary 4.5. *Under the setting of Theorem 4.4 with* $\Delta = 0$, *there exists* $C > 0$ *(depending on* γ *but not on* n, ε*) such that with probability* $1 - \varepsilon$,

$$|J^*(t,x) - J^n(t,x)| \le \frac{C}{\varepsilon^{\rho_\phi/2}} n^{-\frac{\nu\rho_\phi}{4}} (\ln n)^{\frac{1}{2}}. \tag{4.11}$$

Consequently, the regret is:

$$\sum_{k=1}^n |J^*(t,x) - J^k(t,x)| \lesssim \frac{1}{\varepsilon^{\rho_\phi/2}} \begin{cases} n^{1-\frac{\nu\rho_\phi}{4}} (\ln n)^{\frac{1}{2}} & \text{if} \quad \nu\rho_\phi < 4, \\ (\ln n)^{\frac{3}{2}} & \text{if} \quad \nu\rho_\phi = 4, \\ \mathcal{O}(1) & \text{if} \quad \nu\rho_\phi > 4. \end{cases} \tag{4.12}$$

In particular, taking $\nu = 1$ and $\rho_\phi = 1$ (Lipschitz), we get a *sublinear* $n^{\frac{3}{4}}$-regret bound.

Now we illustrate Assumption 4.3 with a simple one-dimensional linear example. Set $\beta = 0$, $\gamma = 1$,

$$b(t,x,a) = Ba \text{ for some } B, \quad \sigma(t,x,a) = 1, \quad r(t,x,a) = 0, \quad h(t,x,a) = x,$$

with the action space $\mathcal{A} = [0,1]$. Given a policy $\pi(\cdot)$, denote by $E(\pi) := \int_{\mathcal{A}} a\pi(a)da$ and $\mathrm{Ent}(\pi) := -\int_{\mathcal{A}} \log \pi(a)\pi(a)da$ its mean and differential entropy respectively. Observe that

$$J(t,x;\pi) = x + (BE(\pi) + \mathrm{Ent}(\pi))(T-t), \quad q(t,x,a;\pi) = B(a - E(\pi)) - \mathrm{Ent}(\pi);$$

so $\nabla J(t,x;\pi) = 1$ for any $\pi(\cdot)$. We parametrize $q^\phi(t,x,a) = \phi a + \log\left(\frac{\phi}{e^\phi - 1}\right)$, and $\pi^\phi(t,x,a) = \frac{\phi}{e^\phi - 1} e^{\phi a}$. A direct computation yields:

$$h(\phi) = -\frac{\phi}{2}\left(\frac{1}{\phi^2} - \frac{e^\phi}{(e^\phi - 1)^2}\right) T.$$

This function is plotted in Figure 1 (with $T = 1$) for visualization. The condition (1) is satisfied, as $h(\phi)$ has a unique zero $\phi^* = 0$ and $h'(0) = -\frac{1}{24}T < 0$. Next, $\mathbb{E}(|\mathcal{H}(\phi_n, U_{n+1})|^2 \,|\, \phi_n) \le CT(1 + |\phi_n|^2)$ for some constant $C > 0$, and $\phi h(\phi) \le -\frac{1}{48}\phi^2$ in a neighborhood of $\phi^* = 0$; so the conditions (2)–(3) hold. Note that $|q^\phi(t,x,a) - q^{\phi'}(t,x,a)| \le 2|\phi - \phi'|$, leading to the condition (4) with $\rho_\phi = 1$. Finally, $|q(\cdot;\pi^\phi) - q(\cdot;\pi^*)|_\infty \le B|E(\pi^\phi) - E(\pi^*)| + |\mathrm{Ent}(\pi^\phi) - \mathrm{Ent}(\pi^*)| \le 2\left(B + \max\log\pi^* + \max\frac{1}{c \wedge \pi^*}\right) d_{TV}(\pi^\phi, \pi^*)$ for some constant $c > 0$ (depending on the neighborhood of ϕ^*), which yields the condition (5).

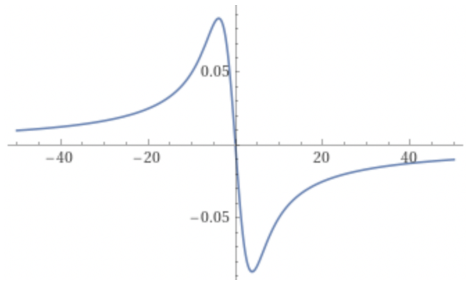

FIGURE 1. Plot of $\phi \to -\frac{\phi}{2}\left(\frac{1}{\phi^2} - \frac{e^\phi}{(e^\phi-1)^2}\right)$.

4.2. ***q*-learning.** Now we consider the q-learning (2.13)–(2.14) by incorporating the policy evaluation part for learning J^n. The idea is similar to the second half of Section 4.1, which we expand as follows.

The update (2.13) can be written as:

$$(\theta_{n+1}, \phi_{n+1}) = (\theta_n, \phi_n) + (\alpha_{\theta,n}, \alpha_{\phi,n})\mathcal{H}((\theta_n, \phi_n), \underbrace{(X_t^{\pi^n}), (a_t^{\pi^n})}_{U_{n+1}}), \tag{4.13}$$

where

$$\mathcal{H}((\theta,\phi), U) := \left(\int_0^T \frac{\partial J^\theta}{\partial \theta}(t, X_t^{\pi^\phi}) G_{t:T}^{\theta,\phi} dt, \int_0^T \int_t^T e^{-\beta(s-t)} \frac{\partial q^\phi}{\partial \phi}(s, X_s^{\pi^\phi}, a_s^{\pi^\phi}) ds G_{t:T}^{\theta,\phi} dt\right), \tag{4.14}$$

with $U = ((X_t^{\pi^\phi}), (a_t^{\pi^\phi}))$, $\pi^\phi(a \,|\, t, x) \propto \exp\left(\frac{1}{\gamma} q^\phi(t,x,a)\right)$ and $G_{t:T}^{\theta,\phi} := e^{-\beta(T-t)} h(X_T^{\pi^\phi}) - J^\theta(t, X_t^{\pi^\phi}) + \int_t^T e^{-\beta(s-t)}[r(s, X_s^{\pi^\phi}, a_s^{\pi^\phi}) - q^\phi(s, X_s^{\pi^\phi}, a_s^{\pi^\phi})] ds$. Define

$$h(\theta, \phi) := \mathbb{E}(\mathcal{H}(\phi, \theta), U). \tag{4.15}$$

We need the following assumptions.

Assumption 4.6.

(1) The ODE $(\theta'(t), \phi'(t)) = h(\theta(t), \phi(t))$ has a unique stable equilibrium (θ^*, ϕ^*).

(2) There exists $C > 0$ such that $\mathbb{E}(|\mathcal{H}(\phi_n, U_{n+1})|^2 \,|\, \phi_n) \le C(1 + |\theta_n|^2 + |\phi_n|^2)$.

(3) There exists $\kappa > 0$ such that $(\theta - \theta^*, \phi - \phi^*) h(\theta, \phi) \le -\kappa\left(|\phi - \phi^*|^2 + |\theta - \theta^*|^2\right)$ in a neighborhood of (θ^*, ϕ^*).

(4) There exist $\rho_\theta, C > 0$ such that $|J^\theta - J^{\theta'}|_\infty \le C|\theta - \theta'|^{\rho_\theta}$ and $|q^\phi - q^{\phi'}|_\infty \le C|\phi - \phi'|^{\rho_\phi}$ for all $\theta, \theta', \phi, \phi'$.

(5) There exists $C > 0$ such that $|q(\cdot; \pi^\phi) - q(\cdot; \pi^*)|_\infty \le C d_{TV}(\pi^\phi, \pi^*)$ for ϕ in a neighborhood of ϕ^*.

Set $\Delta := |J^*(\cdot) - J^{\theta^*}(\cdot)|_\infty \vee |q(\cdot; \pi^*) - q^{\phi^*}(\cdot)|_\infty$. The following result provides an error estimate of the q-learning algorithm, the proof of which is similar to that of Theorem 4.4.

Theorem 4.7. *Let Assumption 4.6 hold. Set $\alpha_{\theta,n}, \alpha_{\phi,n} = \frac{A}{n^\nu + B}$ for some $\nu \le 1$, $A > \frac{\alpha}{2\kappa}$ and $B > 0$, and let $\varepsilon > 0$. Then there exists $C > 0$ (independent of n, ε) such that with probability $1 - \varepsilon$,*

$$|J^*(t,x) - J^{\theta_n}(t,x)| \le \Delta + \frac{C}{\varepsilon^{\rho_\theta/2}} n^{-\frac{\nu\rho_\theta}{2}}. \tag{4.16}$$

Consequently, the regret is:

$$\sum_{k=1}^{n} |J^*(t,x) - J^{\theta_k}(t,x)| \lesssim \Delta n + \frac{1}{\varepsilon^{\rho_\theta/2}} \begin{cases} n^{1-\frac{\nu\rho_\theta}{2}} & \text{if} \quad \nu\rho_\theta < 2, \\ \ln n & \text{if} \quad \nu\rho_\theta = 2, \\ \mathcal{O}(1) & \text{if} \quad \nu\rho_\theta > 2. \end{cases} \tag{4.17}$$

Our ultimate goal is to derive the regret bound of the learned policies $\pi^n(a \,|\, t,x) \propto \exp(q^{\phi_n}(t,x,a))$ in terms of the *original* control problem (2.2) (without the entropy term in the reward functional). To this end, given a policy $\pi(\cdot)$, denfine

$$\mathring{J}(t,x;\pi) := \mathbb{E}\left[\int_t^T e^{-\beta(s-t)} r(s, X_s^\pi, a_s^\pi) ds + e^{-\beta(T-t)} h(X_T^\pi) \middle| X_t^\pi = x\right], \tag{4.18}$$

where $(a_s^\pi,\ t \le s \le T)$ is sampled from π and $\mathbb{E}$ is with respect to both the original probability measure $\mathbb{P}$ and the policy randomization. The following theorem gives the discrepancy between $\mathring{J}^n(t,x) := \mathring{J}(t,x;\pi^n)$ and $\mathring{J}(t,x)$.

Theorem 4.8. *Let the assumptions in Theorem 4.2 and Assumption 4.6 hold. Moreover, assume that* $|q^{\phi_*}|_\infty < \infty$. *Set* $\alpha_{\phi,n} = \frac{A}{n^\nu + B}$ *for some* $\nu \le 1$, $A > \frac{\alpha}{2\kappa}$ *and* $B > 0$, *and let* $\varepsilon > 0$. *Then there exists* $C > 0$ *(independent of* n, ε*) such that with probability* $1 - \varepsilon$,

$$|\mathring{J}^n(t,x) - \mathring{J}(t,x)| \le C\left(\Delta + \frac{\gamma}{\beta} + \frac{1}{\varepsilon^{\rho_\phi/2}} n^{-\frac{\nu\rho_\phi}{4}} (\ln n)^{\frac{1}{2}}\right) + |J^*(t,x) - \mathring{J}(t,x)|. \tag{4.19}$$

Consequently, the regret is:

$$\sum_{k=1}^{n} |\mathring{J}(t,x) - \mathring{J}^k(t,x)| \lesssim \left(\Delta + \frac{\gamma}{\beta} + |J^*(t,x) - \mathring{J}(t,x)|\right) n + \frac{1}{\varepsilon^{\rho_\phi/2}} \begin{cases} n^{1-\frac{\nu\rho_\phi}{4}} (\ln n)^{\frac{1}{2}} & \text{if} \quad \nu\rho_\phi < 4, \\ (\ln n)^{\frac{3}{2}} & \text{if} \quad \nu\rho_\phi = 4, \\ \mathcal{O}(1) & \text{if} \quad \nu\rho_\phi > 4. \end{cases} \tag{4.20}$$

A proof of Theorem 4.8 is deferred to Section 4.3. Recall from (3.4) that under additional conditions on the parameters, $|J^*(t,x) - \mathring{J}(t,x)| \lesssim \gamma \ln(1/\gamma)$. Thus, the regret is of order:

$$\left(C(\gamma)\Delta + \frac{\gamma}{\beta} + \gamma\ln(1/\gamma)\right) n + \frac{C(\gamma)}{\varepsilon^{\rho_\phi/2}} \begin{cases} n^{1-\frac{\nu\rho_\phi}{4}} (\ln n)^{\frac{1}{2}} & \text{if} \quad \nu\rho_\phi < 4, \\ (\ln n)^{\frac{3}{2}} & \text{if} \quad \nu\rho_\phi = 4, \\ \mathcal{O}(1) & \text{if} \quad \nu\rho_\phi > 4. \end{cases}$$

Note, however, that the dependence on γ of the constant $C(\gamma)$ is hard to track. This is because it is related to the constants in Assumption 4.6, and these constants depend on γ in rather obscure ways. As explained in Section 3, it is possible to use an exploratory annealing $\gamma = \gamma_n \downarrow 0$ (as $n \to \infty$) to improve the regret bound, however, at the cost of inefficient sampling of policies in each iteration. The algorithmic implications of the above result is that, for a sufficiently good neural network approximation (so that Δ is sufficiently small) and a sufficiently small temperature parameter γ, the regret of the learned policies applied to the original control problem in the long run is sufficiently small.

4.3. **Proofs.** The proof of Theorem 4.2 requires a series of lemmas. First, similar to Lemma 3.6, we establish a representation for J^n given by (3.1). Recall the definition of $b^n(t,x,a)$ from (4.3). Define

$$\pi^n(a \,|\, s,x,z) \propto \exp\left(\frac{1}{\gamma}(b^n(s,x,a)\sigma^{-1}(s,x)z + r(s,x,a))\right), \tag{4.21}$$

and

$$
\begin{aligned}
G_s^n(z, Z) := \int_{\mathcal{A}} \Big[& b(s, X_s^{\pi^*}, a)\, \sigma^{-1}(s, X_s^{\pi^*}) Z + r(s, X_s^{\pi^*}, a) \\
& - \gamma \log \pi^n(a \,|\, s, X_s^{\pi^*}, z) \Big] \pi^n(a \,|\, s, X_s^{\pi^*}, z)\, da.
\end{aligned}
\tag{4.22}
$$

Lemma 4.9. *Let Assumptions 3.1 and 4.1 hold. There exists $(\widehat{W}, \widehat{\mathbb{P}})$, with $\widehat{W}$ being $\widehat{P}$-Brownian motion, such that*

$$
J^n(t, x) = h(X_T^{\pi^*}) + \int_t^T [G_s^n(Z_s^{n-1}, Z_s^n) - \beta Y_s^n] ds - \int_t^T Z_s^n d\widehat{W}_s, \tag{4.23}
$$

where $Y_s^n := J^n(s, X_s^{\pi^*})$ and $Z_s^n := (\sigma \nabla J^n)(s, X_s^{\pi^*})$.

Proof. The proof is the same as Lemma 3.6-(i), noting that J^n satisfies the PDE (3.7) with $\pi^n(a \,|\, s, x) = \pi^n(a \,|\, s, x, (\sigma \nabla J^{n-1})(s, x))$. □

Recall the definition of $G_s(z, Z)$ from (3.10). Next we bound $|G_s^n(z', Z') - G_s(z, Z)|$.

Lemma 4.10. *Let Assumptions 3.1 and 4.1 hold. Then for $|z|, |Z|, |Z'| \le C$, there exists $L > 0$ (independent of γ) such that*

$$
\begin{aligned}
|G_s^n(z', Z') - G_s(z, Z)| \le & L \left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma} e^{\frac{L}{\gamma}}\right) |z' - z| + L|Z' - Z| \\
& + L \left(1 + \frac{1}{\gamma} e^{\frac{L}{\gamma}}\right) \sup_a |(b - b^n)(s, X_s^{\pi^*}, a) \sigma^{-1}(s, X_s^{\pi^*}) z'| \ \ a.s.
\end{aligned}
\tag{4.24}
$$

Proof. Note that $|G_s^n(z', Z') - G_s(z, Z)| \le |G_s^n(z', Z') - G_s(z', Z')| + |G_s(z', Z') - G_s(z, Z)|$. By Lemma 3.5,

$$
|G_s(z', Z') - G_s(z, Z)| \le L \left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma} e^{\frac{L}{\gamma}}\right) |z' - z| + L|Z' - Z| \ \ a.s. \tag{4.25}
$$

Now we estimate $|G_s^n(z', Z') - G_s(z', Z')|$. Recall the definitions of $\pi(a \,|\, s, x, z)$ and $\pi^n(a \,|\, s, x, z)$ from (3.9) and (4.21) respectively. It is easy to see that

$$
|G_s^n(z', Z') - G_s(z', Z')| \le 2(\bar{b}\bar{\sigma} C + \bar{r}) E_1 + E_2, \tag{4.26}
$$

where

$$
\begin{aligned}
E_1 &:= d_{TV} \left(\pi^n(\cdot \,|\, s, X_s^{\pi^*}, z'), \pi(\cdot \,|\, s, X_s^{\pi^*}, z')\right), \\
E_2 &:= \gamma \left| \int_{\mathcal{A}} \log \pi^n(a \,|\, s, X_s^{\pi^*}, z')\, \pi^n(a \,|\, s, X_s^{\pi^*}, z') - \log \pi(a \,|\, s, X_s^{\pi^*}, z')\, \pi(a \,|\, s, X_s^{\pi^*}, z') da \right|.
\end{aligned}
$$

The same argument as in (3.14) yields

$$
E_1 \le \frac{1}{\gamma} \exp\left(\frac{2}{\gamma}(\bar{b}\bar{\sigma} C + \bar{r})\right) \sup_a |(b - b^n)(s, X_s^{\pi^*}, a) \sigma^{-1}(s, X_s^{\pi^*}) z'| \ \ a.s. \tag{4.27}
$$

Moreover,

$$
\begin{aligned}
E_2 \le & \int_{\mathcal{A}} \left| b^n(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z' - b(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z' \right| \pi^n(a \,|\, s, X^{\pi^*}, z')da \\
& + \int_{\mathcal{A}} b(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z' \left| \pi^n(a \,|\, s, X^{\pi^*}, z') - \pi(a \,|\, s, X^{\pi^*}, z') \right| da \\
& + \int_{\mathcal{A}} |r(t,x,a)| |\pi^n(a \,|\, s, X^{\pi^*}, z') - \pi(a \,|\, s, X^{\pi^*}, z')| da \\
& + \gamma \left| \log \left(\frac{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma}(b^n(s, X_s^{\pi^*}, a)\,\sigma^{-1}(s, X_s^{\pi^*})z' + r(s, X_s^{\pi^*}, a))) \right) da}{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma}(b(s, X_s^{\pi^*}, a)\,\sigma^{-1}(s, X_s^{\pi^*})z' + r(s, X_s^{\pi^*}, a))) \right) da} \right) \right| \\
\le & \; 2 \sup_a |(b - b^n)(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z'| + 2(\bar{b}\bar{\sigma}C + \bar{r})E_1 \\
\le & \; 2\left(1 + \frac{\bar{b}\bar{\sigma}C + \bar{r}}{\gamma} \exp\left(\frac{2}{\gamma}(\bar{b}\bar{\sigma}C + \bar{r}) \right) \right) \sup_a |(b - b^n)(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z'| \;\; a.s.
\end{aligned}
\tag{4.28}
$$

By (4.26), (4.27) and (4.28), we get

$$
\begin{aligned}
& |G_s^n(z', Z') - G_s(z', Z')| \\
& \quad \le 2\left(1 + \frac{2(\bar{b}\bar{\sigma}C + \bar{r})}{\gamma} \exp\left(\frac{2}{\gamma}(\bar{b}\bar{\sigma}C + \bar{r}) \right) \right) \sup_a |(b - b^n)(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z'| \;\; a.s.
\end{aligned}
\tag{4.29}
$$

Combining (4.25) and (4.29) yields (4.24). □

We also need the following variant of Lemma 3.4 on the BSDE.

Lemma 4.11. *For $i = 1, 2$, let $F_t^i(y, z, Z)$ be a measurable function of (t, y, z, Z, ω) such that $(F_t^i(y, z, Z),\ 0 \le t \le T)$ is predictable for any fixed (y, z, Z) and $(F_t^i(0, 0, 0),\ 0 \le t \le T) \in \mathbb{H}_0^0$, and there exist $M_1, M_2, M_3 > 0$,*

$$
|F_t^i(y', z', Z') - F_t^i(y, z, Z)| \le M_1|y' - y| + M_2|z' - z| + M_3|Z' - Z| \quad a.s.,
$$

and that $\delta := \sup_{t,\omega} |F_t^1 - F_t^2|_\infty < \infty$. Further, let $z^i \in \mathbb{H}_0^0$ and (Y^i, Z^i) be the unique solution to

$$
Y_t = \xi + \int_t^T F_t^i(Y_s, z_s^i, Z_s)ds - \int_t^T Z_s dW_s, \quad 0 \le t \le T \quad a.s.
$$

Fixing $\eta \in (0, 1)$, we have for $\theta \ge M_1 + (1 + \eta^{-1})(M_2 + M_3 + \delta)^2$,

$$
e^{\theta t}\mathbb{E}|Y_t^1 - Y_t^2|^2 + |Z^1 - Z^2|^2_{\mathbb{H}_t^\theta} \le \eta(|z^1 - z^2|^2_{\mathbb{H}_t^\theta} + \delta).
$$

Proof. It suffices to note that $|F_t^1(y, z, Z) - F_t^2(y', z', Z')| \le \delta + M_1|y' - y| + M_2|z' - z| + M_3|Z' - Z|$ a.s., and the same argument as in [17, Lemma A.5] permits to conclude. □

Now we give the proof of Theorem 4.2.

Proof of Theorem 4.2. Let $F_s(Y, z, Z) := G_s(z, Z) - \beta Y$ and $F_s^n(Y, z, Z) := G_s^n(z, Z) - \beta Y$. By Lemma 4.9, we have $J^n(t, x) = Y_t^n$ and $J^*(t, x) = Y_t^{t,x}$, where

$$
\begin{aligned}
Y_s^n &= h(X_T^{\pi_\gamma^*}) + \int_s^T F_u^n(Y_u^n, Z_u^{n-1}, Z_u^n)\, du - \int_s^T Z_u^n d\widehat{W}_u, \\
Y_s^{t,x} &= h(X_T^{\pi_\gamma^*}) + \int_s^T F_u(Y_u^{t,x}, Z_u^{t,x}, Z_u^{t,x})\, du - \int_s^T Z_u^{t,x} d\widehat{W}_u,
\end{aligned}
$$

with $Z_s^n = (\sigma\nabla J^n)(s, X_s^{\pi^*})$ and $Z_s^{t,x} = (\sigma\nabla J^*)(s, X_s^{\pi^*})$. By Lemma 4.10 and the assumption that $|\nabla J^n| \le M$ for all n, we get

$$|F_u^n(Y_u^n, Z_u^{n-1}, Z_u^n) - F_u(Y_u^{t,x}, Z_u^{t,x}, Z_u^{t,x})|$$
$$\le \beta|Y_u^n - Y_u^{t,x}| + L\left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)|Z_u^{t,x} - Z_u^{n-1}| + L|Z_u^{t,x} - Z_u^n| + L\left(1 + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)|q(\cdot;\pi^{n-1}) - q^{\phi_n}|_\infty.$$

Applying Lemma 4.11 with $M_1 = \beta$, $M_2 = L\left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)$, $M_3 = L$ and $\delta = L\left(1 + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)|q(\cdot;\pi^{n-1}) - q^{\phi_n}|_\infty$, we have for $\theta \ge \beta + (1+\eta^{-1})L^2\left[1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma}e^{\frac{L}{\gamma}} + L\left(1 + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)|q(\cdot;\pi^{n-1}) - q^{\phi_n}|_\infty\right]^2$,

$$e^{\theta t}\widehat{\mathbb{E}}|Y_t^{t,x} - Y_t^n|^2 + |Z^{t,x} - Z^n|^2_{\widehat{\mathbb{H}}_t^\theta} \le \eta\left[|Z^{t,x} - Z^{n-1}|^2_{\widehat{\mathbb{H}}_t^\theta} + L\left(1 + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)|q(\cdot;\pi^{n-1}) - q^{\phi_n}|_\infty\right].$$

Thus, $|Z^{t,x} - Z^n|^2_{\widehat{\mathbb{H}}_t^\theta} \le \eta^n|Z^{t,x} - Z^0|^2_{\widehat{\mathbb{H}}_t^\theta} + L\left(1 + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)\sum_{k=1}^n \eta^{n+1-k}|q(\cdot;\pi^k) - q^{\phi_{k+1}}|_\infty$, and

$$|J_\gamma^*(t,x) - J_\gamma^n(t,x)|^2 \le \eta^n \int_t^T e^{\theta(T-t)}|\sigma(s, X_s^{\pi_\gamma^*})|^2|\nabla J^*(s, X_s^{\pi_\gamma^*}) - \nabla J^0(s, X_s^{\pi_\gamma^*})|^2 ds$$
$$+ L\left(1 + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)e^{-\theta t}\sum_{k=1}^n \eta^{n+1-k}|q(\cdot;\pi^k) - q^{\phi_{k+1}}|_\infty.$$

This yields the bound (4.4). □

We proceed to proving Theorem 4.4.

Proof of Theorem 4.4. By the argument of [2, Theorem 22], Assumption 4.3 (1)–(3) and the condition on $\alpha_{\phi,n}$ imply

$$\mathbb{E}|\phi^n - \phi^*|^2 \le Cn^{-\nu} \quad \text{for some } C > 0 \text{ (independent of } n).$$

Thus, $|\phi^n - \phi^*| \le C\varepsilon^{-\frac{1}{2}}n^{-\frac{\nu}{2}}$ with a probability of at least $1 - \varepsilon$. Next by Assumption 4.3 (4), we have:

$$|q(\cdot;\pi^*) - q^{\phi_n}|_\infty \le \Delta + C\varepsilon^{-\frac{\rho_\phi}{2}}n^{-\frac{\nu\rho_\phi}{2}}. \tag{4.30}$$

It is easy to deduce that $d_{TV}(\pi^n, \pi^*) \le C(\Delta + \varepsilon^{-\frac{\rho_\phi}{2}}n^{-\frac{\nu\rho_\phi}{2}})$, since $\pi^n(a\,|\,t,x) \propto \exp(\frac{1}{\gamma}q^{\phi_n}(t,x,a))$ and $\pi^*(a\,|\,t,x) \propto \exp(\frac{1}{\gamma}q(t,x,a;\pi^*))$. By Assumption 4.3 (5), we get for Δ sufficiently small,

$$|q(\cdot;\pi^n) - q(\cdot;\pi^*)|_\infty \le C(\Delta + \varepsilon^{-\frac{\rho_\phi}{2}}n^{-\frac{\nu\rho_\phi}{2}}). \tag{4.31}$$

Combining (4.30) and (4.31) gives $|q(\cdot;\pi^n) - q^{\phi_{n+1}}|_\infty \le C(\Delta + \varepsilon^{-\frac{\rho_\phi}{2}}n^{-\frac{\nu\rho_\phi}{2}})$. Applying Theorem 4.2 and (4.5) yields the bound (4.9). □

Finally, we prove Theorem 4.8.

Proof of Theorem 4.8. Note that

$$\begin{aligned}|\mathring{J}^n(t,x) - \mathring{J}(t,x)| &= |\mathring{J}^n(t,x) - J(t,x;\pi^n) + J(t,x;\pi^n) - J^*(t,x) + J^*(t,x) - \mathring{J}(t,x)| \\ &\le |\mathring{J}^n(t,x) - J(t,x;\pi^n)| + |J(t,x;\pi^n) - J^*(t,x)| + |J^*(t,x) - \mathring{J}(t,x)|.\end{aligned} \tag{4.32}$$

It follows from Theorem 4.4 that

$$|J(t,x;\pi^n) - J^*(t,x)| \le C\left(\Delta + \frac{1}{\varepsilon^{\rho_\phi/2}}n^{-\frac{\nu\rho_\phi}{4}}(\ln n)^{\frac{1}{2}}\right). \tag{4.33}$$

Moreover, $|\mathring{J}^n(t,x) - J(t,x;\pi^n)| \le \gamma \mathbb{E}\left[\int_t^T e^{-\beta(s-t)} \int_{\mathcal{A}} |\log \pi^n(a\,|\,s, X_s^{\pi^n})| \pi^n(a\,|\,s, X_s^{\pi^n}) da \middle| X_t^{\pi^n} = x\right]$. Note that $|\log \pi^n(a\,|\,t,x)| \le q^{\phi_n}(t,x,a) + |\log \int_{\mathcal{A}} \exp(q^{\phi_n}(t,x,a)) da|$. By Assumption 4.6 (4) and that $|q^{\phi_*}|_\infty < \infty$, we conclude that $|\log \pi^n(a\,|\,t,x)|$ is uniformly bounded. As a result,

$$|\mathring{J}^n(t,x) - J(t,x;\pi^n)| \le \frac{C\gamma}{\beta} \quad \text{for some } C > 0. \tag{4.34}$$

Combining (4.32), (4.33) and (4.34) yields (4.19). □

5. Conclusion

This paper studies convergence of various RL algorithms for controlled diffusions. We provide the convergence rate and the regret of exploratory policy iteration, semi-q-learning, and q-learning. The tools that we develop in this paper encompass stochastic control, partial differential equations and probability theory (BSDEs in particular).

As continuous RL is still in the early innings and to our best knowledge this paper is the first to study convergence and regret of q-learning, there are numerous open questions. First, one can relax some technical assumptions, e.g. the uniform boundedness of ∇J^n in Theorem 4.2, and provide general sufficient conditions on the model parameters to ensure Assumptions 4.3 and 4.6 (1)–(3),(5). In particular, it is interesting to inquire if these assumptions hold in the important class of linear–quadratic (LQ) controls. Second, we only exercise control (and hence exploration) on the drift, and it is important and curious, as well as challenging, to consider control-dependent diffusion coefficients. Finally, it is intriguing to investigate whether the established convergence rates and regrets are optimal. For that, considering exploratory annealing is a promising direction.

Acknowledgement: Tang gratefully acknowledges financial support through NSF grants DMS-2113779 and DMS-2206038, and by a start-up grant at Columbia University. Zhou gratefully acknowledges financial supports through the Nie Center for Intelligent Asset Management at Columbia.

References

[1] R. Bellman. *Dynamic programming*. Princeton University Press, Princeton, NJ, 1957.

[2] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990.

[3] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2023.

[4] P. Cheridito, H. M. Soner, N. Touzi, and N. Victoir. Second-order backward stochastic differential equations and fully nonlinear parabolic PDEs. *Comm. Pure Appl. Math.*, 60(7):1081–1110, 2007.

[5] S. Chewi. *Log-concave sampling*. 2021. Available at `https://chewisinho.github.io/main.pdf`.

[6] N. El Karoui, S. Peng, and M. C. Quenez. Backward stochastic differential equations in finance. *Math. Finance*, 7(1):1–71, 1997.

[7] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Neurips*, volume 36, 2024.

[8] W. H. Fleming and H. M. Soner. *Controlled Markov processes and viscosity solutions*, volume 25 of *Stochastic Modelling and Applied Probability*. Springer, New York, second edition, 2006.

[9] X. Gao, J. Zha, and X. Y. Zhou. Reward-directed score-based diffusion models via q-learning. 2024. arXiv:2409.04832.

[10] R. A. Howard. *Dynamic programming and Markov processes*. Technology Press of M.I.T., Cambridge, MA; John Wiley & Sons, Inc., New York-London, 1960.

[11] Y. Huang, Y. Jia, and X. Y. Zhou. Mean–variance portfolio selection by continuous-time reinforcement learning: Algorithms, regret analysis, and empirical study. 2024. In preparation.

[12] Y. Huang, Y. Jia, and X. Y. Zhou. Sublinear regret for an actor-critic algorithm in continuous-time linear-quadratic reinforcement learning. 2024. arXiv:2407.17226.

[13] Y.-J. Huang, Z. Wang, and Z. Zhou. Convergence of policy improvement for entropy-regularized stochastic control problems. 2022. arXiv:2209.07059.

[14] Y. Jia and X. Y. Zhou. Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *J. Mach. Learn. Res.*, 23(154):1–55, 2022.

[15] Y. Jia and X. Y. Zhou. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *J. Mach. Learn. Res.*, 23(154):1–55, 2022.

[16] Y. Jia and X. Y. Zhou. q-learning in continuous time. *J. Mach. Learn. Res.*, 24(161):1–61, 2023.

[17] B. Kerimkulov, D. Šiška, and L. Szpruch. Exponential convergence and stability of Howard's policy improvement algorithm for controlled diffusions. *SIAM J. Control Optim.*, 58(3):1314–1340, 2020.

[18] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*, 17(39):1–40, 2016.

[19] J. Ma, G. Wang, and J. Zhang. On convergence and rate of convergence of policy improvement algorithms. 2024. arXiv:2406.10959.

[20] G. Menz, A. Schlichting, W. Tang, and T. Wu. Ergodicity of the infinite swapping algorithm at low temperature. *Stochastic Process. Appl.*, 151:519–552, 2022.

[21] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, and K. Kavukcuoglu. Learning to navigate in complex environments. In *ICLR*, 2017.

[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[23] A. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *ICML*, volume 17, pages 663–670, 2000.

[24] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, et al. Training language models to follow instructions with human feedback. In *Neurips*, volume 35, pages 27730–27744, 2022.

[25] S. Peng. G-expectation, G-Brownian motion and related stochastic calculus of Itô type. In *Stochastic analysis and applications*, volume 2 of *Abel Symp.*, pages 541–567. Springer, Berlin, 2007.

[26] S. G. Peng. A general stochastic maximum principle for optimal control problems. *SIAM J. Control Optim.*, 28(4):966–979, 1990.

[27] H. Pham. *Continuous-time stochastic control and optimization with financial applications*, volume 61 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2009.

[28] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999.

[29] S. Russell. Learning agents for uncertain environments. In *COLT*, pages 101–103, 1998.

[30] D. Sethi, D. Šiška, and Y. Zhang. Entropy annealing for policy mirror descent in continuous time and space. 2024. arXiv:2405.20250.

[31] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, and M. Lanctot. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[32] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, and A. Bolton. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[33] B. Sprungk. On the local Lipschitz stability of Bayesian inverse problems. *Inverse Problems*, 36(5):055015, 31, 2020.

[34] R. S. Sutton and A. G. Barto. *Reinforcement learning: an introduction*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2018.

[35] W. Tang, H. V. Tran, and Y. P. Zhang. Policy iteration for the deterministic control problems–a viscosity approach. 2023. arXiv:2301.00419. To appear in SIAM J. Control Optim.

[36] W. Tang, Y. P. Zhang, and X. Y. Zhou. Exploratory HJB equations and their convergence. *SIAM J. Control Optim.*, 60(6):3191–3216, 2022.

[37] W. Tang and H. Zhao. Score-based diffusion models via stochastic differential equations–a technical tutorial. 2024. arXiv:2402.07487.

[38] H. V. Tran, Z. Wang, and Y. P. Zhang. Policy iteration for exploratory Hamilton–Jacobi–Bellman equations. 2024. arXiv:2406.00612.

[39] D. H. Wagner. Survey of measurable selection theorems. *SIAM J. Control Optim.*, 15(5):859–903, 1977.

[40] H. Wang, T. Zariphopoulou, and X. Y. Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *J. Mach. Learn. Res.*, 21:1–34, 2020.

[41] H. Wang and X. Y. Zhou. Continuous-time mean-variance portfolio selection: a reinforcement learning framework. *Math. Finance*, 30(4):1273–1308, 2020.

[42] J. Yong and X. Y. Zhou. *Stochastic controls – Hamiltonian systems and HJB equations*, volume 43 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1999.

[43] H. Zhao, H. Chen, J. Zhang, D. D. Yao, and W. Tang. Scores as Actions: a framework of fine-tuning diffusion models by continuous-time reinforcement learning. 2024. arXiv:2409.08400.
[44] H. Zhao, W. Tang, and D. D. Yao. Policy optimization for continuous reinforcement learning. In *Neurips*, 2023.
[45] X. Y. Zhou. The curse of optimality, and how to break it? In *Machine Learning and Data Sciences for Financial Markets: A Guide to Contemporary Practices*, pages 354–368. Cambridge University Press, 2023.

Department of Industrial Engineering and Operations Research, Columbia University.
Email address: wt2319@columbia.edu

Department of Industrial Engineering and Operations Research, Columbia University.
Email address: xz2574@columbia.edu