Journal of Machine Learning Research 27 (2026) 1-37 Submitted 11/24; Revised 8/26; Published 10/26

Regret of exploratory policy improvement and q-learning

Wenpin Tang WT2319@COLUMBIA.EDU
Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027, USA

Xun Yu Zhou XZ2574@COLUMBIA.EDU
Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027, USA

Editor: Csaba Szepesvari

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. For exploratory policy improvement, we establish exponential convergence under growth and regularity assumptions on the model parameters. For q-learning, we derive quantitative error and regret bounds under additional assumptions on the function approximation and the associated stochastic approximation dynamics.

Keywords: 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 (Silver et al. (2016, 2017)) and video games (Mnih et al. (2015)), driving autonomously (Levine et al. (2016); Mirowski et al. (2017)), and more recently, aligning large language models and text-to-image generative models with human preference (Black et al. (2023); Fan et al. (2024); Ouyang et al. (2022)). RL research has predominantly focused on Markov decision processes (MDPs) in discrete time and space; see Sutton and Barto (2018) for a detailed account of theory and applications for MDPs.

Wang et al. (2020) 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 (Jia and Zhou (2022a,b, 2023)), and by policy optimization (Zhao et al. (2023, 2026)). 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 Jia and Zhou

©2026 Wenpin Tang and Xun Yu Zhou.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/24-1928.html.

(2022a,b, 2023) 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 and/or ad hoc 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 Huang et al. (2024) for continuous-time mean–variance portfolio selection and Huang et al. (2025a); Huang and Zhou (2026) for a class of stochastic linear–quadratic (LQ) control problems, all of which apply/adapt the policy gradient algorithms developed in Jia and Zhou (2022a) and exploit heavily the special structures of the problems.

The purpose of the paper is to fill this gap by providing a quantitative analysis of (little) *q-learning* introduced in Jia and Zhou (2023) 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. Jia and Zhou (2023) proposes the notion of the q-function, which is the first-order derivative of the Q-function with respect to time discretization. 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/computable 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 (Gao et al. (2025); Zhao et al. (2025)).

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 (Bellman (1957); Howard (1960)) improves the policies gradually in order to approximate the optimal one. So it is also called policy improvement; see Kerimkulov et al. (2020); Jia and Zhou (2023); Tang et al. (2025) for recent studies. In the RL setting, an entropy regularizer is added to encourage exploration, which gives rise to the exploratory control problem (Tang et al. (2022)). 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 Ma et al. (2026); Sethi et al. (2025); Tran et al. (2025) 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 2);
- an explicit error bound of q-learning, depending on the regularity of the model parameters and the learning rate (Theorems 13 and 14).

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 8), 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. In particular, we provide the first general quantitative framework for the mode-free q-learning based on stochastic approximation, which allows us to perform further convergence and regret analysis. 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 (Gao et al. (2025); Zhao et al. (2024, 2025); Sheng et al. (2025); Huang et al. (2026); Zhang et al. (2026)).

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, $\mathrm{Tr}(A)$ 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 $0 \leq \alpha < 1$, define the parabolic Hölder norm:

$$|f|_{\mathcal{C}^{\alpha/2,\alpha}} := \sup_{[0,T]\times\mathbb{R}^d} f + \sup_{(t,x)\neq(s,y)\in[0,T]\times\mathbb{R}^d} \frac{|f(t,x)-f(s,y)|}{\left(|t-s|^{1/2}+|x-y|\right)^{\alpha}}.$$

  A function $f \in \mathcal{C}^{\frac{\alpha}{2},\alpha}([0,T]\times\mathbb{R}^d)$ if f, $\partial_t f$, ∇f and $\nabla^2 f$ have bounded parabolic Hölder norms.

- 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 $0 \leq t \leq T$, $\mathbb{H}_t^\theta := \{$predictable processes $(x_s,\, t \leq s \leq T)$ 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 $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\}_{t\geq 0})$ be a filtered probability space on which we define a d'-dimensional $\mathcal{F}_t$-adapted Brownian motion $(W_t,\, t \geq 0)$. Let $\mathcal{A}$ be a generic action space, and $u = (u_t,\, 0 \leq t \leq T)$ 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):

$$dX_t^u = b(t, X_t^u, u_t)dt + \sigma(t, X_t^u, u_t)dW_t, \tag{1}$$

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:

$$\mathring{J}(t,x) := \sup_u \mathbb{E}\left[\int_t^T e^{-\beta(s-t)} r(s, X_s^u, u_s)ds + e^{-\beta(T-t)} h(X_T^u)\middle| X_t^u = x\right], \tag{2}$$

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 $(X_t^*,\, t \geq 0)$ 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 Fleming and Soner (2006); Yong and Zhou (1999) 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 errors. This calls for an essentially and drastically different school of thoughts from that of the classical control theory.

Inspired by this, Wang et al. (2020) models exploration by a probability distribution of controls $\pi = (\pi_t(\cdot),\, 0 \leq t \leq T)$ over the action space $\mathcal{A}$ from which each trial is sampled. The exploratory state process is:

$$dX_t^\pi = \widetilde{b}(t, X_t^\pi, \pi_t)dt + \widetilde{\sigma}(t, X_t^\pi, \pi_t)d\widetilde{W}_t, \tag{3}$$

where $(\widetilde{W}_t,\, t \geq 0)$ 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

$$\widetilde{b}(t,x,\pi) := \int_{\mathcal{A}} b(t,x,a)\pi(a)da \quad \text{and} \quad \widetilde{\sigma}(t,x,\pi) := \left(\int_{\mathcal{A}} \sigma(t,x,a)\sigma(t,x,a)^T\pi(a)da\right)^{\frac{1}{2}}, \tag{4}$$

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 $\pi = (\pi_t(\cdot),\, 0 \leq t \leq T)$ is also known as the relaxed control in the control literature. The objective is to maximize the entropy-regularized problem:

$$\begin{aligned} &J^*(t,x) \\ &:= \sup_\pi J(t,x;\pi) \\ &:= \sup_\pi \mathbb{E}\bigg[\int_t^T e^{-\beta(s-t)} \int_{\mathcal{A}} [r(s, X_s^\pi, a) - \gamma \log \pi_s(a)]\pi_s(a)da\, ds + e^{-\beta(T-t)} h(X_T^\pi) \bigg| X_t^\pi = x\bigg], \end{aligned} \tag{5}$$

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

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

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* (Russell (1998); Ng and Russell (2000)).

be the generalized Hamiltonian. According to Tang et al. (2022), Section 5, under suitable conditions on the parameters (r, h, b, σ), J^* is the solution to the *exploratory* Hamilton–Jacobi–Bellman (HJB) equation:

$$\begin{cases} \partial_t J^* + \gamma \log \int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} H(t, x, a, \nabla J^*, \nabla^2 J^*)\right) da - \beta J^* = 0, \\ J^*(T, x) = h(x). \end{cases} \tag{6}$$

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

$$\pi^*(\cdot \,|\, t, x) \propto \exp\left(\frac{1}{\gamma} H(t, x, \cdot, \nabla J^*, \nabla^2 J^*)\right). \tag{7}$$

See Tang et al. (2022); Zhou (2023) for further development on the exploratory control problem (5), and its HJB equation (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

$$dX_t^\pi = \widetilde{b}(t, X_t^\pi, \pi_t)dt + \sigma(t, X_t^\pi)d\widetilde{W}_t. \tag{8}$$

2.3 q-learning

The main task of RL is to solve the entropy-regularized problem (5) by learning its optimal policy (7).

The first idea, based on the policy improvement theorems in Jia and Zhou (2023); Wang and Zhou (2020), 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:

$$\pi^{n+1}(\cdot \,|\, t, x) \propto \exp\left(\frac{1}{\gamma} H(t, x, \cdot, \nabla J^n, \nabla^2 J^n)\right). \tag{9}$$

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 (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 Jia and Zhou (2023), is to learn *directly* a term equivalent to $H(t, x, a, \nabla J^n, \nabla^2 J^n)$ for the purpose of sampling (9). That term is called the (little) q-function (analogous to the Q-function in the classical discrete-time RL for MDPs).

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 Cheridito et al. (2007); Peng (2007). This is also related to the second-order risk adjustment of stochastic maximum principle when control enters into the diffusion term; see Peng (1990) and Yong and Zhou (1999), Chapter 3. We leave this case for future work.

Specifically, given a control $\pi(\cdot)$, the q-function is

$$\begin{aligned} q(t,x,a;\pi) :&= \frac{\partial J}{\partial t}(t,x;\pi) + H\left(t,x,a,\nabla J(t,x;\pi),\nabla^2 J(t,x;\pi)\right) - \beta J(t,x;\pi) \\ &= b(t,x,a)\nabla J(t,x;\pi) + r(t,x,a) + \frac{\partial J}{\partial t}(t,x;\pi) \\ &\qquad + \frac{1}{2}\operatorname{Tr}\left(\sigma^2(t,x)\nabla^2 J(t,x;\pi)\right) - \beta J(t,x;\pi). \end{aligned} \tag{10}$$

Denote by $a^\pi = (a_t^\pi,\, 0 \le t \le T)$ a control realization from sampling the policy $\pi(\cdot)$. Conceptually, the action process a^π should be sampled independently over (continuous) time. However, as pointed out in (Szpruch et al., 2024, Remark 2.1) and (Bender and Thuan, 2024, Section 3) the resulting process a^π is not necessarily progressively measurable. This was also observed earlier in (Stoyanov, 1987, Section 19.5) and (Kallianpur, 1980, Example 1.2.5). One remedy is to consider the discretized version:

$$a_t^\pi = \sum_{k=0}^{N-1} a_{t_k}^\pi 1(t_k \le t < t_{k+1}), \quad \text{for a partition } 0 = t_0 < t_1 < \cdots < t_N = T,$$

which is termed a "discretely sampled process (DSPs)" in Jia et al. (2026). Such defined $(a_t^{\pi^\phi})$ is measurable; so all the integrals involving a^π are well-defined. Jia and Zhou (2025), an erratum of Jia and Zhou (2023), restate and reprove all the q-learning results of the latter in terms of DSPs. Note that DSPs are exactly what are actually implemented in any real application; so the theory and implementation/application align here. To simplify notation and presentation, henceforth we assume that sampling of $\pi(\cdot)$ is always in the sense of DSPs. See also Footnote 11 for a remark on the discretization error arising from DSPs.

The key takeaway in Jia and Zhou (2023) is to observe the fact that

$$e^{-\beta s}J(s,X_s^\pi;\pi) + \int_t^s e^{-\beta u}\left[r(u,X_u^\pi,a_u^\pi) - q(u,X_u^\pi,a_u^\pi;\pi)\right]du, \quad t \le s \le T, \tag{11}$$

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:

$$\min_{(\theta,\phi)} \mathbb{E}\left[\int_0^T \left\{e^{-\beta(T-t)}h(X_T^\pi) - J^\theta(t,X_t^\pi) + \int_t^T e^{-\beta(s-t)}[r(s,X_s^\pi,a_s^\pi) - q^\phi(s,X_s^\pi,a_s^\pi)]ds\right\}^2 dt\right]. \tag{12}$$

Problem (12) can be solved by stochastic gradient descent (SGD):

$$\begin{aligned} \theta &\leftarrow \theta + \alpha_\theta \int_0^T \frac{\partial J^\theta}{\partial \theta}(t,X_t^\pi)G_{t:T}dt, \\ \phi &\leftarrow \phi + \alpha_\phi \int_0^T \int_t^T e^{-\beta(s-t)}\frac{\partial q^\phi}{\partial \phi}(s,X_s^\pi,a_s^\pi)ds G_{t:T}dt, \end{aligned}$$

where $G_{t:T} := e^{-\beta(T-t)}h(X_T^\pi) - J^\theta(t,X_t^\pi) + \int_t^T e^{-\beta(s-t)}[r(s,X_s^\pi,a_s^\pi) - q^\phi(s,X_s^\pi,a_s^\pi)]ds$, and $\alpha_\theta, \alpha_\phi$ are user-defined learning rates. Note that in actual implementation of this procedure one needs to take DSPs (described earlier) for actions while observing the resulting states

and reward signals at the corresponding discretized time points and calculating the integrals as summations using the same grid 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 (10) involves the control variable a, the policy update (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 $\pi^1(\cdot\,|\,\cdot,\cdot)$, and for $n\geq 1$,

1. Update

$$\begin{aligned}\theta_{n+1}&=\theta_n+\alpha_{\theta,n}\int_0^T\frac{\partial J^\theta}{\partial\theta}_{|\theta=\theta_n}(t,X_t^{\pi^n})G_{t:T}^n dt,\\ \phi_{n+1}&=\phi_n+\alpha_{\phi,n}\int_0^T\int_t^T e^{-\beta(s-t)}\frac{\partial q^\phi}{\partial\phi}_{|\phi=\phi_n}(s,X_s^{\pi^n},a_s^{\pi^n})ds G_{t:T}^n dt,\end{aligned}\tag{13}$$

   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

$$\pi^{n+1}(\cdot\,|\,t,x)\propto\exp\left(\frac{1}{\gamma}q^{\phi_{n+1}}(t,x,\cdot)\right).\tag{14}$$

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 2) shows that it is exponentially fast. The arguments will then be used for the subsequent (model-free) analysis of the q-learning algorithm (13)–(14); so this section is also a technical preparation for later development.

Recall (8) that defines the exploratory state process X^π. The exploratory policy improvement starts with a control policy $\pi^1(\cdot\,|\,\cdot,\cdot)$, and for $n\geq 1$,

$$\begin{aligned}J^n(t,x):=\mathbb{E}\bigg[\int_t^T e^{-\beta(s-t)}\int_{\mathcal{A}}[r(s,X_s^{\pi^n},a)-\gamma\log\pi^n(a|s,X_s^{\pi^n})]\pi^n(a|s,X_s^{\pi^n})da\,ds\\ +e^{-\beta(T-t)}h(X_T^{\pi^n})\Big|X_t^{\pi^n}=x\bigg].\end{aligned}\tag{15}$$

Sample

$$\begin{aligned}\pi^{n+1}(\cdot\,|\,t,x)&\propto\exp\left(\frac{1}{\gamma}H(t,x,\cdot,\nabla J^n(t,x),\nabla^2 J^n(t,x))\right)\\ &\propto\exp\left(\frac{1}{\gamma}\left(b(t,x,\cdot)\nabla J^n(t,x)+r(t,x,\cdot)\right)\right),\end{aligned}\tag{16}$$

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 (16).

As previously mentioned, we assume that the value function J^n defined in (15) (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 (5).

The convergence of J^n is proved in Huang et al. (2025b), but with no convergence rate given. Two concurrent papers (Ma et al., 2026; Tran et al., 2025) study the convergence rate of *model-based* policy iteration (16). Their results, the exponential convergence of J^n, are similar to Theorem 2 below. Tran et al. (2025), as Huang et al. (2025b), is based on a PDE analysis, where the model parameters are assumed to be Hölder smooth. This yields Schauder's estimates and, hence, a stronger convergence in Hölder spaces. However, these PDE methods do not track explicitly the dependence on the level of exploration γ. Ma et al. (2026) rely on Malliavin calculus, where they assume the model parameters to be smooth and bounded as in our paper below. The dependence on γ, which depends on higher-order derivatives of the Gibbs map, is not transparent there as well. Nevertheless, our BSDE approach differs from Ma et al. (2026); Tran et al. (2025), and our result provides an explicit dependence on the exploration γ, which is important in efficient sampling of the policies (16) but is missing in the aforementioned work.[3]

Our analysis for exploratory policy improvement follows the general BSDE framework developed in Kerimkulov et al. (2020). In particular, several stability and contraction estimates are adaptations of arguments from Kerimkulov et al. (2020) to the exploratory-control setting considered here. The main new ingredients are the entropy-regularized Gibbs policies and the resulting nonlinear BSDE generators. 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, where the treatment of approximation errors arising from q-learning was considered neither in Kerimkulov et al. (2020), nor in Ma et al. (2026); Tran et al. (2025).[4]

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

Assumption 1

(i) $\mathcal{A}$ *is a compact set.*

(ii) r, h, b, σ *are continuous in their respective arguments.*

3. It is worth noting that Ma et al. (2026); Tran et al. (2025) prove the convergence of J^n to J^* in Hölder norms, which is stronger than our pointwise convergence in Theorem 2. Nevertheless, our goal is to derive the regret in the reinforcement learning setting, which is the total reward loss $\sum_{k=1}^{n}(J^* - J^k)$. So the pointwise convergence is sufficient for this purpose.
4. In this sense, Kerimkulov et al. (2020); Ma et al. (2026); Tran et al. (2025) are "numerical" papers solving classical stochastic control problems, while ours is a "learning" paper solving the model-free counterparts.

(iii) There exist $\overline{b}, \overline{\sigma}, \overline{r} > 0$ *such that for any* (t, x, a)*,*

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

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

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

(v) r, b, σ *belong to* $\mathcal{C}^{\frac{\alpha}{2},\alpha}$ *in* (t, x) *uniformly in* a*, and* $h \in \mathcal{C}^{2+\alpha}(\mathbb{R}^d)$ *for some* $\alpha > 0$*.*

(vi) For the initial policy, $\log \pi^1$ *belongs to* $\mathcal{C}^{\frac{\alpha}{2},\alpha}$ *in* (t, x) *uniformly in* a*.*

We emphasize that Jia and Zhou (2023) require only the weaker linear-growth condition on b, σ for their policy improvement and q-learning framework. Here the stronger boundedness assumption (iii) is imposed to obtain further quantitative convergence and regret estimates, especially for deriving Lipschitz and contraction estimates of entropy-regularized Gibbs policies.

Our result is stated as follows.

Theorem 2 *Let Assumption 1 hold, and fix* $\eta \in (0, 1)$*. There exist* $L, C > 0$ *(independent of* γ *and* n*) such that*

$$|J^*(t,x) - J^n(t,x)|^2 \le C\eta^n e^{\theta(\gamma)(T-t)}, \tag{17}$$

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 (17) implies that the exploratory policy constructed by (16) 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 2).
- Larger γ facilitates Markov chain Monte Carlo (MCMC) sampling of the policy $\pi^n(\cdot \,|\, t, x)$ (Chewi (2021), Chapter 4 and Menz et al. (2022), Section 2.6).[5]

Second, recall that $\mathring{J}$ is the optimal value function of the original, classical control problem (2). It follows from (Tang et al., 2022, Theorem 3.7) that under suitable conditions

5. 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.

on the parameters,[6]

$$|J^*(t,x) - \mathring{J}(t,x)| \le C\gamma \ln(1/\gamma), \quad \text{as } \gamma \to 0^+. \tag{18}$$

Combining (17) and (18) yields

$$|J^n(t,x) - \mathring{J}(t,x)| \lesssim \underbrace{\gamma \ln(1/\gamma)}_{\text{bias}} + \underbrace{\eta^n e^{\theta(\gamma)(T-t)}}_{\text{policy improvement error}}, \tag{19}$$

giving rise to a tradeoff between bias (due to exploration) and policy improvement error in γ. Heuristically, minimizing the right side of (19) 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 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 8) that

$$|J^n(t,x) - \mathring{J}(t,x)| \lesssim \sum_{k=1}^{n} \eta^{n-k} \gamma_k. \tag{20}$$

Thus, γ_n needs to be fast-decaying so that the right side of (20) 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 2. The next lemma specifies the PDE satisfied by the value function J^n.

Lemma 3 *Under Assumption 1, J^n defined by* (15) *is the unique classical solution to*

$$\begin{cases} \partial_t J + \int_{\mathcal{A}} \big(H(t,x,a,\nabla J, \nabla^2 J) - \gamma \log \pi^n(a \,|\, t,x)\big)\, \pi^n(a \,|\, t,x)\, da - \beta J = 0, \\ J(T,x) = h(x). \end{cases} \tag{21}$$

6. The results in (Tang et al., 2022, Section 3) imply that for $(t,x) \in [0,T] \times B_r$, $|J^*(t,x) - \mathring{J}(t,x)|$ is controlled by the Laplace approximation error:

$$\sup_{t,x,p,q} \left| \gamma \log \int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} H(t,x,a,p,q)\right) da - \sup_{a \in \mathcal{A}} H(t,x,a,p,q) \right|,$$

where the supremum is over all $t \in [0,T]$ and all x,p,q in large balls whose radii depend on r. This bound is problem-dependent, which echoes the observation in Sethi et al. (2025). Under suitable conditions (e.g., Tang et al., 2022, Corollary 4.7 or when H is uniformly concave in a in our problem), the above bound is of order $\gamma \ln(1/\gamma)$ as $\gamma \to 0^+$, which gives (18). Note that the bound $\gamma \ln(1/\gamma)$ is universal for a class of problems, but it is possible to get better rates under specific problem structures.

Proof We first prove that under Assumption 1, the PDE (21) has a unique classical solution

$$\bar{J}^n \in \mathcal{C}^{1+\frac{\alpha}{2},2+\alpha}([0,T)\times\mathbb{R}^d)\cap\mathcal{C}([0,T]\times\mathbb{R}^d).$$

Indeed, for a given policy $\pi^n(\cdot\,|\,t,x)$, the PDE (21) is linear parabolic:

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

Since σ is positive definite (Assumption 1-(iii)) and the coefficients are Hölder ((v)–(vi)), Schauder's theory (see e.g., Friedman 1964, Chapter 3) implies this equation when $n=1$ has a unique solution $\bar{J}^1 \in \mathcal{C}^{1+\frac{\alpha}{2},2+\alpha}([0,T)\times\mathbb{R}^d)\cap\mathcal{C}([0,T]\times\mathbb{R}^d)$. A standard verification argument (by applying Itô's formula to $e^{-\beta(s-t)}\bar{J}^1(s,X_s^{\pi^n})$, integrating from t to T and then taking expectation) yields $J^1\equiv\bar{J}^1$. Now, $\nabla J^1\equiv\nabla\bar{J}^1\in\mathcal{C}^{\frac{\alpha}{2},1+\alpha}([0,T)\times\mathbb{R}^d)\cap\mathcal{C}([0,T]\times\mathbb{R}^d)$ and, hence, $\log\pi^2(a\,|\,t,x)$ belongs to $\mathcal{C}^{\frac{\alpha}{2},\alpha}$ in $(t.x)$. Repeating the above argument inductively yields $J^n\equiv\bar{J}^n$. ■

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

Lemma 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

$$Y_t=\xi+\int_t^T F_t(Y_s,z_s,Z_s)ds-\int_t^T Z_s d\widetilde{W}_s,\quad 0\le t\le T\quad a.s. \tag{22}$$

(iii) Let (Y^1,Z^1) and (Y^2,Z^2) be the solutions to (22) 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 $\theta\ge M_1+(1+\eta^{-1})(M_2+M_3)^2$,

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

Proof (i) and (ii) follow from the standard BSDE theory; see, e.g., Pham (2009), Theorem 6.3.3, and (iii) from the proof of Kerimkulov et al. (2020), Lemma A.5. ■

3.2 Proof of Theorem 2

We need a few lemmas to prove Theorem 2. Recall the definition of the optimal feedback control π^* from (7), and let X^{π^*} be the corresponding state process. Denote (with a slight abuse of notation)

$$\pi(a \,|\, s, x, z) \propto \exp\left(\frac{1}{\gamma}\left(b(s,x,a)\,\sigma^{-1}(s,x)z + r(s,x,a)\right)\right), \tag{23}$$

which is well-defined as $\sigma(s,x)$ is positive definite by Assumption 1-(iii). Then 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

$$\begin{aligned} G_s(z, Z) := \int_{\mathcal{A}} \Big[& b(s, X_s^{\pi^*}, a)\,\sigma^{-1}(s, X_s^{\pi^*})Z + r(s, X_s^{\pi^*}, a) \\ & - \gamma \log \pi(a \,|\, s, X_s^{\pi^*}, z)\Big]\pi(a \,|\, s, X_s^{\pi^*}, z)\, da, \end{aligned} \tag{24}$$

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

Lemma 5 *Let Assumption 1 hold. Then for $|z|, |Z| \le C$, there exists $L > 0$ (independent of γ) such that*

$$|G_s(z', Z') - G_s(z, Z)| \le L\left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)|z' - z| + L|Z' - Z| \quad a.s., \tag{25}$$

for all z' and Z'.

Proof Observe that

$$|G_s(z', Z') - G_s(z, Z)| \le E_1 + E_2 + E_3, \tag{26}$$

where

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

By Assumption 1-(iii), we get

$$E_1 \le \overline{r}\,\left|\pi(\cdot \,|\, s, X_s^{\pi^*}, z') - \pi(\cdot \,|\, s, X_s^{\pi^*}, z)\right|_{\mathbb{L}^1(\mathcal{A})} = 2\overline{r}\, d_{TV}\left(\pi(\cdot \,|\, s, X_s^{\pi^*}, z'),\, \pi(\cdot \,|\, s, X_s^{\pi^*}, z)\right). \tag{27}$$

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 (Sprungk, 2020, Theorem 8), [7] we get for $|z'-z| \le C$ (and hence $|z'| \le 2C$):

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

Moreover, it is true that $d_{TV}\left(\pi(\cdot \,|\, s,x,z'),\, \pi(\cdot \,|\, s,x,z)\right) \le 1$ (because the total variation distance of any two probability distributions is bounded by 1.) So for $|z'-z| > C$, we have $d_{TV}\left(\pi(\cdot \,|\, s,x,z'),\, \pi(\cdot \,|\, s,x,z)\right) \le 1 < |z'-z|/C$. Consequently,

$$d_{TV}\left(\pi(\cdot \,|\, s,x,z'),\, \pi(\cdot \,|\, s,x,z)\right) \le \left(\frac{\bar{b}\bar{\sigma}}{\gamma} \exp\left(\frac{3}{\gamma}(\bar{b}\bar{\sigma}C+\bar{r})\right) + \frac{1}{C}\right)|z'-z|. \tag{28}$$

Combining (27) and (28) yields

$$E_1 \le 2\bar{r}\left(\frac{\bar{b}\bar{\sigma}}{\gamma} \exp\left(\frac{3}{\gamma}(\bar{b}\bar{\sigma}C+\bar{r})\right) + \frac{1}{C}\right)|z'-z|. \tag{29}$$

Similarly, we get

$$\begin{aligned} E_2 & \le \bar{b}\bar{\sigma}\left(\int_{\mathcal{A}} |Z'-Z|\,\pi(a \,|\, s, X_s^{\pi^*}, z')\,da + \int_{\mathcal{A}} |Z|\,|\pi(a \,|\, s, X_s^{\pi^*}, z') - \pi(a \,|\, s, X_s^{\pi^*}, z)|da\right) \\ & \le \bar{b}\bar{\sigma}\left(|Z'-Z| + 2C\, d_{TV}\left(\pi(\cdot \,|\, s, X_s^{\pi^*}, z'),\, \pi(\cdot \,|\, s, X_s^{\pi^*}, z)\right)\right) \\ & \le \bar{b}\bar{\sigma}|Z'-Z| + \left(\frac{2\bar{b}\bar{\sigma}C}{\gamma} \exp\left(\frac{3}{\gamma}(\bar{b}\bar{\sigma}C+\bar{r})\right) + 2\bar{b}\bar{\sigma}\right)|z'-z|. \end{aligned} \tag{30}$$

7. For two Gibbs measures $\pi(da) = e^{\Psi(a)}da/Z$ and $\pi'(da) = e^{\Psi'(a)}da/Z'$, (Sprungk, 2020, 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

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

Finally, we have

$$\begin{aligned} E_3 \le & \underbrace{\left| \int_{\mathcal{A}} r(s, X_s^{\pi^*}, a) \left(\pi(a \mid s, X_s^{\pi^*}, z') - \pi(a \mid s, X_s^{\pi^*}, z) \right) da \right|}_{(a)} \\ & + \underbrace{\left| \int_{\mathcal{A}} b(s, X_s^{\pi^*}, a)\, \sigma^{-1}(s, X_s^{\pi^*}) \left(z' \pi(a \mid s, X_s^{\pi^*}, z') - z \pi(a \mid s, X_s^{\pi^*}, z) \right) da \right|}_{(b)} \\ & + \underbrace{\gamma \left| \log \left(\frac{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} (b(s, X_s^{\pi^*}, a)\, \sigma^{-1}(s, X_s^{\pi^*}) z' + r(s, X_s^{\pi^*}, a)) \right) da}{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} (b(s, X_s^{\pi^*}, a)\, \sigma^{-1}(s, X_s^{\pi^*}) z + r(s, X_s^{\pi^*}, a)) \right) da} \right) \right|}_{(c)} \\ \le & \left(3\bar{b}\bar{\sigma} + \frac{2\bar{b}\bar{\sigma}(C + \bar{r})}{\gamma} \exp\left(\frac{3}{\gamma} (\bar{b}\bar{\sigma} C + \bar{r}) \right) + \frac{2\bar{r}}{C} \right) |z' - z| + (c), \end{aligned} \tag{31}$$

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,

$$\begin{aligned} \left| \log \left(\frac{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} (b(s, x, a)\, \sigma^{-1}(s, x) z' + r(s, x, a)) \right) da}{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} (b(s, x, a)\, \sigma^{-1}(s, x) z + r(s, x, a)) \right) da} \right) \right| &\le \log \left(\frac{e^{\frac{\bar{b}\bar{\sigma}|z'| + \bar{r}}{\gamma}} |A|}{e^{-\frac{\bar{b}\bar{\sigma}C + \bar{r}}{\gamma}} |A|} \right) \\ &= \frac{\bar{b}\bar{\sigma}(|z'| + C) + 2\bar{r}}{\gamma} \\ &\le \frac{\bar{b}\bar{\sigma}|z' - z| + 2(\bar{b}\bar{\sigma} C + \bar{r})}{\gamma} \\ &\le \frac{3\bar{b}\bar{\sigma} + 2\bar{r}/C}{\gamma} |z' - z|, \end{aligned} \tag{32}$$

where the second inequality follows from $|z'| \le |z' - z| + |z| \le |z' - z| + C$, and the last inequality is due to the fact that $|z' - z| > C$.

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

$$\left| \log \left(\frac{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} (b(s, x, a)\, \sigma^{-1}(s, x) z' + r(s, x, a)) \right) da}{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma} (b(s, x, a)\, \sigma^{-1}(s, x) z + r(s, x, a)) \right) da} \right) \right| \le \frac{(d)}{(e)}, \tag{33}$$

where

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

By Sprungk (2020), Theorem 5, we have

$$(d) \le \frac{\bar{b}\bar{\sigma}|\mathcal{A}|}{\gamma} \exp\left(\frac{1}{\gamma}(\bar{b}\bar{\sigma}C + \bar{r})\right)|z' - z|, \quad (e) \ge |\mathcal{A}| \exp\left(-\frac{2\bar{b}\bar{\sigma}C + \bar{r}}{\gamma}\right). \tag{34}$$

The estimates (32)–(34) yield

$$(c) \le \left(\bar{b}\bar{\sigma} \exp\left(\frac{3\bar{b}\bar{\sigma}C + 2\bar{r}}{\gamma}\right) + 3\bar{b}\bar{\sigma} + \frac{2\bar{r}}{C}\right)|z' - z|. \tag{35}$$

By (31) and (35), we obtain

$$E_3 \le \left(6\bar{b}\bar{\sigma} + \frac{4\bar{r}}{C} + \bar{b}\bar{\sigma}\left(1 + \frac{2(C + \bar{r})}{\gamma}\right)\exp\left(\frac{3}{\gamma}(\bar{b}\bar{\sigma}C + \bar{r})\right)\right)|z' - z|. \tag{36}$$

Combining (26) with (29), (30) and (36) yields (25). ■

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

Lemma 6 *Let Assumption 1 hold. There exist $(\widehat{W}, \widehat{\mathbb{P}})$, with $\widehat{W}$ being $\widehat{\mathbb{P}}$-Brownian motion, such that*

(i) We have

$$J^n(t,x) = h(X_T^{\pi^*}) + \int_t^T [G_s(Z_s^{n-1}, Z_s^n) - \beta Y_s^n]\, ds - \int_t^T Z_s^n d\widehat{W}_s, \tag{37}$$

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

$$Y_s^{t,x} = h(X_T^{\pi^*}) + \int_s^T [G_u(Z_u^{t,x}, Z_u^{t,x}) - \beta Y_u^{t,x}]\, du - \int_s^T Z_u^{t,x} d\widehat{W}_u, \quad t \le s \le T, \tag{38}$$

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, we get

$$\begin{aligned}
&dJ^n(s, X_s^{\pi^*}) \\
&= \left[\partial_t J^n(s, X_s^{\pi^*}) + \frac{1}{2}\operatorname{Tr}(\sigma\sigma^T \nabla^2 J^n)(s, X_s^{\pi^*}) + \widetilde{b}(s, X_s^{\pi^*}, \pi^*)\nabla J^n(s, X_s^{\pi^*})\right] ds \\
&\qquad + (\sigma \nabla J^n)(s, X_s^{\pi^*}) d\widetilde{W}_s \\
&= \Big[\widetilde{b}(s, X_s^{\pi^*}, \pi^*)\nabla J^n(s, X_s^{\pi^*}) - \widetilde{b}(s, X_s^{\pi^*}, \pi^n)\nabla J^n(s, X_s^{\pi^*}) - \widetilde{r}(s, X_s^{\pi^*}, \pi^n) \\
&\quad + \gamma \int_{\mathcal{A}} \log \pi^n(a \,|\, s, X_s^{\pi^*})\, \pi^n(a \,|\, s, X_s^{\pi^*}) da + \beta J^n(s, X_s^{\pi^*})\Big] ds + (\sigma \nabla J^n)(s, X_s^{\pi^*}) d\widetilde{W}_s,
\end{aligned} \tag{39}$$

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 Revuz and Yor (1999), 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 (39) 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 (37).

(ii) The fact that the BSDE (38) has a unique solution follows from Lemma 4-(i) noting Lemma 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 El Karoui et al. (1997), Proposition 4.3 or Pham (2009), Section 6.3) shows that $Y_t^{t,x} = v(t,x)$ is the solution to the semi-linear PDE:[8]

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

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 (6), i.e. $Y_t^{t,x} = J^*(t,x)$. ■

Now we prove Theorem 2.

8. The semi-linear PDE can be written as $\partial_t v + \frac{1}{2}\operatorname{Tr}\left(\sigma\sigma^T \nabla^2 v\right) + f(t,x,v,p) = 0$, where

$$f(t,x,v,p) := \gamma \log \int_{\mathcal{A}} \exp\left(\frac{1}{\gamma}(b(t,x,a)p + r(t,x,a))\right) da - \beta v.$$

It follows from Assumption 1 that f is $\mathcal{C}^{\frac{\alpha}{2},\alpha}$ in (t,x) and Lipschitz in (v,p). Classical Schauder's theory (see e.g., Friedman, 1964, Chapter 7) implies that the PDE has a unique solution $\bar{v} \in \mathcal{C}^{1+\frac{\alpha}{2},2+\alpha}([0,T)\times\mathbb{R}^d) \cap \mathcal{C}([0,T]\times\mathbb{R}^d)$, which then coincides with v by a standard verification argument.

Proof (of Theorem 2) Let $F_s(Y,z,Z) := G_s(z,Z) - \beta Y$. It follows from Lemma 6 that $J^n(t,x) = Y_t^n$ and $J^*(t,x) = Y_t^{t,x}$ where

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

Note that $|\nabla J^*| \le M$ (with M independent of γ; see Tang et al., 2022), and $Z_s^{t,x} = (\sigma\nabla J^*)(s, X_s^{\pi^*})$. By Lemma 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 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 (17). ■

4. q-learning and regret

In this section, we study the q-learning algorithm (13)–(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 $n \ge 1$,

1. Update

$$\phi_{n+1} = \phi_n + \alpha_{\phi,n} \int_0^T \int_t^T e^{-\beta(s-t)} \frac{\partial q^\phi}{\partial \phi}_{|\phi=\phi_n} (s, X_s^{\pi^n}, a_s^{\pi^n}) ds G_{t:T}^n dt, \tag{40}$$

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

$$\pi^{n+1}(a \,|\, t, x) \propto \exp\left(\frac{1}{\gamma} q^{\phi_{n+1}}(t, x, a)\right) da. \tag{41}$$

Recall (15) that defines J^n and it follows from (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 throughout this subsection that $\nabla J^n \neq 0$ almost everywhere.[9] Let $b^{n+1}(t, x, a)$ be a measurable function such that the following holds almost everywhere[10]:

$$\begin{aligned} & q^{\phi_{n+1}}(t, x, a) \\ & = b^{n+1}(t, x, a) \nabla J^n(t, x) + r(t, x, a) + \frac{\partial J^n}{\partial t}(t, x) + \frac{1}{2} \operatorname{Tr}\left(\sigma^2(t, x) \nabla^2 J^n(t, x)\right) - \beta J^n(t, x). \end{aligned} \tag{42}$$

The policy update (41) can now be written as:

$$\pi^{n+1}(a \,|\, t, x) \propto \exp\left(\frac{1}{\gamma} (b^{n+1}(t, x, a) \nabla J^n(t, x) + r(t, x, a))\right) da. \tag{41'}$$

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 1.

Assumption 7 *There exists $\overline{b} > 0$ such that $\operatorname{ess\,sup}_{t,x,a} |b^n| \leq \overline{b}$ 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.

9. It is known (Wang and Zhou, 2020) 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 11 is another instance satisfying this assumption. In these examples, $b^n(t, x, a)$ is well-defined. Later, an additional assumption on the uniform boundedness of b^n (Assumption 7) is required in order to establish the convergence of the learning algorithm. We note, however, that LQ problems fall outside this bounded-coefficient setting.

10. 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 (42) if $d \geq 2$. The measurability follows from standard measurable selection theory, e.g., Wagner (1977).

Theorem 8 *Let Assumptions 1 and 7 hold, and assume that there is $M > 0$ such that $|\nabla J^n| \le M$ for all n. Fix $\eta \in (0,1)$. There exist L, $C > 0$ (independent of γ and n) such that*

$$|J^*(t,x) - J^n(t,x)|^2 \le C\left(\eta^n e^{\Lambda(\gamma)(T-t)} + \left(1 + \frac{1}{\gamma}e^{\frac{L}{\gamma}}\right)\sum_{k=1}^{n}\eta^{n-k}|q(\cdot;\pi^k) - q^{\phi_{k+1}}|_\infty\right), \quad (43)$$

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 8 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. Indeed, this assumption follows from a smallness condition on the model parameters. By a heat kernel estimate (or Bismut's formula, Elworthy and Li, 1994), we get $|\nabla J^n|_\infty \le C_1(1 + |f_n|_\infty)$, where $f_n(t,x) := \int_{\mathcal{A}} (r(t,x,a) - \gamma \log \pi^n(a\,|\,t,x))\,\pi^n(a\,|\,t,x)da$. Here the constant C_1 depends on the uniform ellipticity and regularity of σ, the time horizon T, the uniform bound $\bar{b}$ of $\int_{\mathcal{A}} b(t,x,a)\pi^n(a\,|\,t,x)da$, and the uniform bound K of $|\nabla h|_\infty$, but not on n. Next, recall that $\pi^n(a\,|\,t,x) \propto \exp\left(\frac{b(t,x,a)\nabla J^{n-1}(t,x) + r(t,x,a)}{\gamma}\right)$. We have $|f_n|_\infty \le C_2(1 + |\nabla J^{n-1}|_\infty)$, where C_2 depends on the boundedness of the model parameters – it is again independent of n. As a result,

$$|\nabla J^n|_\infty \le C_1(1 + C_2) + C_1C_2|\nabla J^{n-1}|_\infty.$$

If the smallness condition $C_1C_2 < 1$ holds, the above recursion implies that $\sup_n |\nabla J^n| < \infty$.

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 (43) 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 (43) 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:

$$\sum_{k=1}^{n}|J^*(t,x) - J^k(t,x)| \lesssim \begin{cases} n^{1-\frac{\alpha}{2}}(\ln n)^{\frac{1}{2}} & \text{if} \quad \alpha < 2, \\ (\ln n)^{\frac{3}{2}} & \text{if} \quad \alpha = 2, \\ \mathcal{O}(1) & \text{if} \quad \alpha > 2. \end{cases} \quad (44)$$

Next we consider the convergence of the q values via the iteration (40). This is an instance of the Robbins–Monro algorithm, which can be written as:

$$\phi_{n+1} = \phi_n + \alpha_{\phi,n}\mathcal{H}(\phi_n, \underbrace{(X_t^{\pi^n}), (a_t^{\pi^n})}_{U_{n+1}}), \quad (45)$$

where

$$\mathcal{H}(\phi, U) := \int_0^T \int_t^T e^{-\beta(s-t)} \frac{\partial q^\phi}{\partial \phi}(s, X_s^{\pi^\phi}, a_s^{\pi^\phi}) ds G_{t:T}^\phi dt, \tag{46}$$

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 $G_{t:T}^\phi := e^{-\beta(T-t)} h(X_T^{\pi^\phi}) - J(t, X_t^{\pi^\phi}; \pi^\phi) + \int_t^T e^{-\beta(s-t)} [r(s, X_s^{\pi^\phi}, a_s^{\pi^\phi}) - q^\phi(s, X_s^{\pi^\phi}, a_s^{\pi^\phi})] ds$.[11] Clearly,

$$(U_{n+1} \,|\, U_n, \ldots, U_1, \phi_n, \ldots, \phi_1) \stackrel{d}{=} (U_{n+1} \,|\, \phi_n).$$

Also define

$$l(\phi) := \mathbb{E}\mathcal{H}(\phi, U). \tag{47}$$

We need a few more assumptions to study the convergence rate of the stochastic approximation (45) (or (40)).

Assumption 9

1. *The ODE $\phi'(t) = l(\phi(t))$ 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 $(\phi - \phi^*) l(\phi) \le -\kappa |\phi - \phi^*|^2$ 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 (45), (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 (40)–(41).

Theorem 10 *Let the assumptions in Theorem 8 and Assumption 9 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 $1 - \varepsilon$,*

$$|J^*(t, x) - J^n(t, x)| \le C \left(\Delta + \frac{1}{\varepsilon^{\rho_\phi/2}} n^{-\frac{\nu \rho_\phi}{4}} (\ln n)^{\frac{1}{2}} \right), \tag{48}$$

11. As commented in Subsection 2.3 the action process a^{π^ϕ} is discretely sampled to ensure the required measurability. Such defined a^{π^ϕ} will incur a discretization error of order $\max_{0 \le k \le N-1} |t_{k+1} - t_k|$ for ϕ, which vanishes as the time step size decreases. For ease of presentation, we omit to write out the effect of discretization in the results below. If time is equipartitioned, it will add a term $N^{-\frac{\rho_\phi}{2}}$ (resp. $nN^{-\frac{\rho_\phi}{2}}$) in (48) (resp. (49)).

for Δ sufficiently small. Consequently, the regret is:

$$\sum_{k=1}^{n} |J^*(t,x) - J^k(t,x)| \lesssim \Delta n + \frac{1}{\varepsilon^{\rho_\phi/2}} \begin{cases} n^{1-\frac{\nu\rho_\phi}{4}} (\ln n)^{\frac{1}{2}} & \text{if} \quad \nu\rho_\phi < 4, \\ (\ln n)^{\frac{3}{2}} & \text{if} \quad \nu\rho_\phi = 4, \\ \mathcal{O}(1) & \text{if} \quad \nu\rho_\phi > 4. \end{cases} \tag{49}$$

Note that in the reinforcement learning and Markov decision processes literature, the regret is typically defined cumulatively, i.e., the total reward loss incurred by not following the optimal policy. Accordingly, in our entropy-regularized continuous-time setting, it is defined by $\sum_{k=1}^{n}(J^* - J^k)$, which is bounded from above by $\sum_{k=1}^{n} |J^* - J^k|$. A proof of Theorem 10 will be given in Section 4.3. As previously mentioned, the error term Δ in (48) (or (49)) 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 11 *Under the setting of Theorem 10 with $\Delta = 0$, there exists $C > 0$ (depending on γ but not on n, ε) such that with probability $1 - \varepsilon$,*

$$|J^*(t,x) - J^n(t,x)| \leq \frac{C}{\varepsilon^{\rho_\phi/2}} n^{-\frac{\nu\rho_\phi}{4}} (\ln n)^{\frac{1}{2}}. \tag{50}$$

Consequently, the regret $\sum_{k=1}^{n} |J^*(t,x) - J^k(t,x)|$ is given by the second term in (49).

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

Now we illustrate Assumption 9 with an example.

Example 1 Set

$$b(t,x,a) = \alpha x + \rho a, \quad \sigma(t,x,a) = \sigma, \quad r(t,x,a) = rx, \quad h(x) = \ell x, \quad \text{and } \beta = 0$$

with the action space $\mathcal{A} = [0,1]$. As mentioned in Footnote 1, the dynamics parameters α, ρ and σ are unknown, while the reward parameters r and ℓ may be specified by the users. Given a policy π, denote by $E(\pi) := \int_{\mathcal{A}} a\pi(a)da$ and $\text{Ent}(\pi) := -\int_{\mathcal{A}} \pi(a)\log\pi(a)da$ its mean and differential entropy respectively. We discuss two cases based on how policies are parameterized.

Case 1 (three-parameter policies). Assume $\alpha \neq 0$. In this case

$$J(t,x;\pi) = A(t)x + B^\pi(t) \quad \text{and} \quad q(t,x,a;\pi) = \rho A(t)\left(a - E(\pi_t)\right) - \gamma\text{Ent}(\pi_t),$$

where $A(t) := e^{\alpha(T-t)}\left(\ell + \frac{r}{\alpha}\right) - \frac{r}{\alpha}$ and $B^\pi(t) := \int_t^T [\rho A(s)E(\pi_s) + \gamma\text{Ent}(\pi_s)]ds$. It is clear that $\nabla J(t,x;\pi) = A(t) \neq 0$ almost everywhere. The expressions of $q(t,x,a;\pi)$ and $A(t)$ suggest to parameterize:

$$q_\phi(t,x,a) = F_\phi(t)(a - E(\pi_t^\phi)) - \gamma\text{Ent}(\pi_t^\phi), \quad \text{with } F_\phi(t) := \phi_1 e^{\phi_2(T-t)} + \phi_3,$$

and $\pi_t^\phi(a) \propto e^{-F_\phi(t)a/\gamma}$ on $\mathcal{A} = [0,1]$. Define $w_i(t) := \partial_{\phi_i} F_\phi(t)$, and $l_i(\phi)$ to be the i^{th} coordinate of $l(\phi)$ for $i \in \{1,2,3\}$. We have:

$$w_1(t) = e^{\phi_2(T-t)}, \quad w_2(t) = \phi_1(T-t)e^{\phi_2(T-t)}, \quad w_3(t) = 1.$$

A direct computation yields:

$$l_i(\phi) = \int_0^T \int_0^T (s \wedge u)(\rho A(s) - F_\phi(s))E(\pi_s^\phi)w_i(u)E(\pi_u^\phi)dsdu.$$

Let us now check Assumption 9. Clearly, $l(\phi^*) = 0$ with $\phi^* \equiv (\phi_1^*, \phi_2^*, \phi_3^*) = \left(\rho\left(\ell + \frac{r}{\alpha}\right), \alpha, -\frac{\rho r}{\alpha}\right)$. Denote $w_i^*(t) = w_i(t)|_{\phi=\phi^*}$. Assume further that

$$\phi_1^* = \rho\left(\ell + \frac{r}{\alpha}\right) \neq 0,$$

and restrict the policy parameters to a compact set $\mathcal{K} = \{\phi : |\phi_i| < C,\ i = 1, 2, 3,\ |\phi_1| \geq c,\ \phi_2 \in I\}$ with I a compact interval containing α but not 0, for some $c, C > 0$. In this case, F_ϕ is injective on $\mathcal{K}$. Compute the Jacobian matrix $Dl(\phi^*) \equiv -G$, where

$$G_{ij} := \int_0^T \int_0^T (s \wedge u)w_j^*(s)E(\pi_s^{\phi^*})w_i^*(u)E(\pi_u^{\phi^*})dsdu.$$

Since $\phi_1^*, \phi_2^* \neq 0$, the functions (w_1^*, w_2^*, w_3^*) are linearly independent. Combining with the fact $E(\pi_t^{\phi^*}) > 0$ implies that G is positive definite. As a result, l has a unique stable equilibrium ϕ^* and is locally dissipative, i.e., the conditions (1) and (3) of Assumption 9 are satisfied. Moreover, the condition (2) holds automatically because $\mathcal{A}$ is compact and X^{π^ϕ} is Gaussian. Next, since q_ϕ is smooth in $\phi \in \mathcal{K}$, it is Lipschitz leading to the condition (4) with $\rho_\phi = 1$. Finally, $|q(\cdot;\pi^\phi) - q(\cdot;\pi^*)| \leq C|E(\pi^\phi) - E(\pi^*)| + |\text{Ent}(\pi^\phi) - \text{Ent}(\pi^*)| \leq 2\left(C + \gamma \max \log \pi^* + \gamma \max \frac{1}{c \wedge \pi^*}\right) d_{TV}(\pi^\phi, \pi^*)$ for some constants $C > 0$, and $c > 0$ (depending on the neighborhood of ϕ^*). This yields the condition (5).

Case 2 (one-parameter policies). Assume $\alpha = 0$ and $\rho \neq 0$. We have:

$$J(t, x; \pi) = A(t)x + B^\pi(t) \quad \text{and} \quad q(t, x, a; \pi) = \rho A(t)\left(a - E(\pi_t)\right) - \gamma \text{Ent}(\pi_t),$$

where $A(t) := \ell + r(T - t)$ and $B^\pi(t) := \int_t^T [\rho A(s)E(\pi_s) + \gamma \text{Ent}(\pi_s)]ds$. Note that $\nabla J(t, x; \pi) \neq 0$ almost everywhere, provided that $r \neq 0$ or $\ell \neq 0$. The main difference from Case 1 is that $A(t)$ involves only reward parameters; so we do not need to parameterize it if we assume that the reward parameters are specified in advance. In this case, we only need to parameterize the q-function with a single parameter:

$$q_\phi(t, x, a) = \phi A(t)(a - E(\pi_t^\phi)) - \gamma \text{Ent}(\pi_t^\phi),$$

where $\pi_t^\phi(a) \propto e^{-\phi A(t)a/\gamma}$ on $\mathcal{A} = [0, 1]$. A direct computation yields:

$$l(\phi) = (\rho - \phi) \int_0^T \left(\int_t^T A(s)E(\pi_s^\phi)ds\right)^2 dt.$$

Since $\int_0^T \left(\int_t^T A(s)E(\pi_s^\phi)ds\right)^2 dt \geq c > 0$ on a compact set, l is dissipative and has a unique stable equilibrium, satisying the conditions (1) and (3). Other conditions can be verified similarly to (indeed more easily than) Case 1. A special case with $r = 0$ gives the following closed-form formula:

$$l(\phi) = \frac{T^3\ell^2}{3}\left(\frac{\gamma}{\phi\ell} - \frac{1}{e^{\phi\ell/\gamma} - 1}\right)^2 (\rho - \phi).$$

It is easy to see that $l'(\rho) = -\frac{T^3\ell^2}{3}\left(\frac{\gamma}{\rho\ell} - \frac{1}{e^{\rho\ell/\gamma} - 1}\right)^2 < 0$; see Figure 1 for a plot of l.

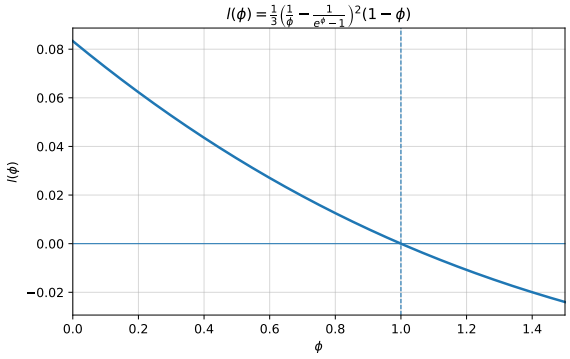


Figure 1: Plot of $l(\phi) = \frac{1}{3}(\frac{1}{\phi} - \frac{1}{e^{\phi}-1})^2(1-\phi)$ for $\alpha = r = 0$, $\rho = \gamma = \ell = 1$ and $T = 1$.

4.2 q-learning

Now we consider the q-learning (13)–(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 (13) can be written as:

$$(\theta_{n+1}, \phi_{n+1}) = (\theta_n, \phi_n) + (\alpha_{\theta,n}, \alpha_{\phi,n})\mathcal{H}((\theta_n, \phi_n), \underbrace{(X_t^{\pi^n}), (a_t^{\pi^n})}_{U_{n+1}}), \tag{51}$$

where

$$\mathcal{H}((\theta,\phi), U) := \left(\int_0^T \frac{\partial J^\theta}{\partial \theta}(t, X_t^{\pi^\phi}) G_{t:T}^{\theta,\phi} dt, \int_0^T \int_t^T e^{-\beta(s-t)} \frac{\partial q^\phi}{\partial \phi}(s, X_s^{\pi^\phi}, a_s^{\pi^\phi}) ds G_{t:T}^{\theta,\phi} dt \right), \tag{52}$$

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

$$l(\theta, \phi) := \mathbb{E}(\mathcal{H}(\phi, \theta), U). \tag{53}$$

We need the following assumptions.

Assumption 12

1. *The ODE* $(\theta'(t), \phi'(t)) = l(\theta(t), \phi(t))$ *has a unique stable equilibrium* (θ^*, ϕ^*).
2. *There exists* $C > 0$ *such that* $\mathbb{E}(|\mathcal{H}(\theta_n, \phi_n, U_{n+1})|^2 \,|\, \theta_n, \phi_n) \le C(1 + |\theta_n|^2 + |\phi_n|^2)$.
3. *There exists* $\kappa > 0$ *such that* $(\theta - \theta^*, \phi - \phi^*) l(\theta, \phi) \le -\kappa \left(|\phi - \phi^*|^2 + |\theta - \theta^*|^2 \right)$ *in a neighborhood of* (θ^*, ϕ^*).
4. *There exist* $\rho > 0$ *such that* $|J^\theta - J^{\theta'}|_{\infty,R} := \sup_{t \le T, |x| \le R} |J^\theta(t,x) - J^{\theta'}(t,x)| \le C|\theta - \theta'|^{\rho_\theta}$ *for some* $C > 0$ *(depending on* T *and* R*), and all* θ, θ'.

Set $\Delta_R := |J^*(\cdot) - J^{\theta^*}(\cdot)|_{\infty,R}$. The following result provides an error estimate of the q-learning algorithm, the proof of which is similar to that of Theorem 10.

Theorem 13 *Let Assumption 12 hold. Set* $\alpha_{\theta,n}, \alpha_{\phi,n} = \frac{A}{n^\nu + B}$ *for some* $\nu \leq 1$, $A > \frac{\alpha}{2\kappa}$ *and* $B > 0$, *and let* $\varepsilon > 0$. *Then there exists* $C > 0$ *(independent of* n, ε*) such that with probability* $1 - \varepsilon$,

$$|J^*(t,x) - J^{\theta_n}(t,x)| \leq \Delta_R + \frac{C}{\varepsilon^{\rho_\theta/2}} n^{-\frac{\nu\rho_\theta}{2}}, \tag{54}$$

for $0 \leq t \leq T$ *and* $|x| \leq R$. *Consequently, the regret is:*

$$\sum_{k=1}^{n} |J^*(t,x) - J^{\theta_k}(t,x)| \lesssim \Delta n + \frac{1}{\varepsilon^{\rho_\theta/2}} \begin{cases} n^{1-\frac{\nu\rho_\theta}{2}} & \text{if } \nu\rho_\theta < 2, \\ \ln n & \text{if } \nu\rho_\theta = 2, \\ \mathcal{O}(1) & \text{if } \nu\rho_\theta > 2. \end{cases} \tag{55}$$

We illustrate Assumption 12 with the same linear setting as in Example 1.

Example 1′ We work under the same problem setting as in Example 1.

Case 1 (six-parameter policies). Assume $\alpha \neq 0$. Recall the expression of $J(t,x;\pi)$ and $q(t,x,a;\pi)$ obtained earlier. We parameterize:

$$\begin{aligned} q_\phi(t,x,a) &= F_\phi(t)(a - E(\pi_t^\phi)) - \gamma \text{Ent}(\pi_t^\phi), \quad \text{with } F_\phi(t) := \phi_1 e^{\phi_2(T-t)} + \phi_3, \\ J_{\theta,\phi}(t,x,a) &= A_\theta(t)x + B_\phi(t), \quad \text{with } A_\theta(t) := \theta_1 e^{\theta_2(T-t)} + \theta_3, \\ & \qquad\qquad B_\phi(t) := \int_t^T [F_\phi(s)E(\pi_s^\phi) + \gamma \text{Ent}(\pi_s^\phi)]ds. \end{aligned}$$

Recall that for $i \in \{1,2,3\}$, $w_i(t) := \partial_{\phi_i} F_\phi(t)$. Define $v_i(t) := \partial_\theta A_\theta(t)$, with

$$v_1(t) = e^{\theta_2(T-t)}, \quad v_2(t) = \theta_1(T-t)e^{\theta_2(T-t)}, \quad v_3(t) = 1.$$

Denote by $l_j(\theta,\phi)$ the j^{th} coordinate of $l(\theta,\phi)$ with $j \in \{1,\ldots,6\}$. We compute for $j \in \{1,2,3\}$,

$$\begin{aligned} l_j(\theta,\phi) &= \int_0^T v_j(t) \left[(A(t) - A_\theta(t))S_\phi(t) + m_\phi(t) \int_t^T (\rho A(s) - F_\phi(s))E(\pi_s^\phi)ds \right], \\ l_{3+j}(\theta,\phi) &= \int_0^T (A(t) - A_\theta(t))m_\phi(t)R_j^\phi(t)dt + \int_0^T R_j^\phi(t) \int_t^T (\rho A(s) - F_\phi(s))E(\pi_s^\phi)dsdt, \end{aligned}$$

where $m_\phi(t) := \mathbb{E}X_t^{\pi^\phi}$, $S_\phi(t) := \mathbb{E}(X_t^{\pi^\phi})^2$, and $R_j^\phi(t) := \int_t^T w_j(s)E(\pi_s^\phi)ds$.

It is easy to check that the zero of l is

$$(\theta^*, \phi^*) \equiv (\theta_1^*, \theta_2^*, \theta_3^*, \phi_1^*, \phi_2^*, \phi_3^*) = \left(\ell + \frac{r}{\alpha}, \alpha, -\frac{r}{\alpha}, \rho\left(\ell + \frac{r}{\alpha}\right), \alpha, -\frac{\rho r}{\alpha} \right).$$

Similar to Example 1, we assume $\rho\left(\ell + \frac{r}{\alpha}\right) \neq 0$, and restrict the parameters to a compact set

$$\mathcal{K} := \{(\theta,\phi) : |\theta_i| \vee |\phi_i| < C,\ i = 1,2,3,\ |\theta_1| \wedge |\phi_1| \geq c,\ \theta_2, \phi_2 \in I\},$$

with I a compact interval containing α but not 0, for some $c, C > 0$. In this case, (A_θ, F_ϕ) is injective on $\mathcal{K}$. Note that $A(t) - A_{\theta^*}(t) = \rho A(t) - F_{\phi^*}(t) = 0$; so the derivatives of

$m_\phi, S_\phi, E(\pi^\phi), R_\phi^j$ do not contribute at (θ^*, ϕ^*). We get for $i, j \in \{1, 2, 3\}$,

$$\partial_{\theta_i} l_j(\theta^*, \phi^*) = -\int_0^T v_i^*(t) v_j^*(t) S_{\phi^*}(t) dt, \quad \partial_{\phi_i} l_j(\theta^*, \phi^*) = -\int_0^T v_j^*(t) m_{\phi^*}(t) R_i^{\phi^*}(t) dt,$$
$$\partial_{\phi_i} l_{3+j}(\theta^*, \phi^*) = -\int_0^T R_i^{\phi^*}(t) R_j^{\phi^*}(t) dt, \quad \partial_{\theta_i} l_{3+j}(\theta^*, \phi^*) = -\int_0^T v_i^*(t) m_{\phi^*}(t) R_j^{\phi^*}(t) dt.$$

Writing the Jacobian $Dl(\theta^*, \phi^*) \equiv -G$, we have for $x, y \in \mathbb{R}^3$,

$$(x, y)^T G(x, y) = \int_0^T \mathbb{E}\left(\sum_{i=1}^3 x_i v_i^*(t) X_t^{\pi^{\phi^*}} + \sum_{j=1}^3 y_j R_i^{\phi^*}(t) \right)^2 dt.$$

The quadratic form is zero if and only if $\sum_i x_i v_i^*(t) = 0$ and $\sum_j y_j R_i^{\phi^*}(t) = 0$ almost everywhere (because $\sigma > 0$, and hence $X^{\pi^{\phi^*}}$ is non-degenerate). The independence of (v_1^*, v_2^*, v_3^*) and (w_1^*, w_2^*, w_3^*) implies that $x = y = 0$. As a result, K is positive definite. Thus, l has a unique stable equilibrium (θ^*, ϕ^*), and is locally dissipative, leading to the conditions (1) and (3). The other conditions follow the same arguments as in Example 1.

Case 2 (one-parameter policies). Let $\alpha = 0$. We parameterize

$$q_\phi(t, x, a) = \phi A(t)(a - E(\pi_t^\phi)) - \gamma \mathrm{Ent}(\pi_t^\phi) \quad \text{and} \quad J_\phi(t, x) = A(t)x + B_\phi(t).$$

That is, we do not need to parameterize the value function J separately. In this case, we have

$$l(\phi) = (\rho - \phi) \int_0^T \left(\int_t^T A(s) E(\pi_s^\phi) ds \right)^2 dt.$$

The remainder of the argument is identical to that of Example 1.

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) (without the entropy term in the reward functional). To this end, given a policy $\pi(\cdot)$, denfine

$$\mathring{J}(t, x; \pi) := \mathbb{E}\left[\int_t^T e^{-\beta(s-t)} r(s, X_s^\pi, a_s^\pi) ds + e^{-\beta(T-t)} h(X_T^\pi) \middle| X_t^\pi = x \right], \tag{56}$$

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 14 *Let the assumptions in Theorem 8 and Assumption 12 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 $1 - \varepsilon$,*

$$|\mathring{J}^n(t, x) - \mathring{J}(t, x)| \le C \left(\Delta + \frac{\gamma}{\beta} + \frac{1}{\varepsilon^{\rho_\phi/2}} n^{-\frac{\nu \rho_\phi}{4}} (\ln n)^{\frac{1}{2}} \right) + |J^*(t, x) - \mathring{J}(t, x)|. \tag{57}$$

Consequently, the regret is:

$$\begin{aligned}&\sum_{k=1}^{n} |\mathring{J}(t,x) - \mathring{J}^k(t,x)| \\ &\qquad \lesssim \left(\Delta + \frac{\gamma}{\beta} + |J^*(t,x) - \mathring{J}(t,x)|\right) n + \frac{1}{\varepsilon^{\rho_\phi/2}} \begin{cases} n^{1-\frac{\nu\rho_\phi}{4}} (\ln n)^{\frac{1}{2}} & \textit{if} \quad \nu\rho_\phi < 4, \\ (\ln n)^{\frac{3}{2}} & \textit{if} \quad \nu\rho_\phi = 4, \\ \mathcal{O}(1) & \textit{if} \quad \nu\rho_\phi > 4. \end{cases}\end{aligned} \tag{58}$$

A proof of Theorem 14 is deferred to Section 4.3. Recall from (18) 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 implicitly related to the constants in Assumption 12, and these constants also depend on γ in rather obscure ways. Nevertheless, since the Gibbs policy becomes less regular as γ decreases, we conjecture that it is the main contribution to the dependence $C(\gamma)$, i.e., $C(\gamma)$ increases rapidly as $\gamma \to 0^+$. If our conjecture holds, then there is the exploration–stability tradeoff: a smaller γ reduces the entropy bias, but worsens the stability of sampling and learning processes.[12] A possible approach is to decrease exploration $\gamma = \gamma_n \to \gamma_\star > 0$ (as $n \to \infty$) to mitigate the entropy bias, however, at the cost of inefficient sampling and learning 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 suitably large temperature parameter γ, the regret of the learned policies applied to the original control problem in the long run is well controlled.

4.3 Proofs

The proof of Theorem 8 requires a series of lemmas. First, similar to Lemma 6, we establish a representation for J^n given by (15). Recall the definition of $b^n(t,x,a)$ from (42). Define

$$\pi^n(a \,|\, s,x,z) \propto \exp\left(\frac{1}{\gamma}(b^n(s,x,a)\sigma^{-1}(s,x)z + r(s,x,a))\right), \tag{59}$$

and

$$\begin{aligned}G_s^n(z,Z) := \int_{\mathcal{A}} \Big[& b(s, X_s^{\pi^*}, a)\, \sigma^{-1}(s, X_s^{\pi^*}) Z + r(s, X_s^{\pi^*}, a) \\ & - \gamma \log \pi^n(a \,|\, s, X_s^{\pi^*}, z) \Big] \pi^n(a \,|\, s, X_s^{\pi^*}, z)\, da.\end{aligned} \tag{60}$$

12. This is in contrast with model-based policy improvement in Section 3, where a smaller γ fosters convergence and, hence, reducing regret.

Lemma 15 *Let Assumptions 1 and 7 hold. There exists $(\widehat{W}, \widehat{\mathbb{P}})$, with $\widehat{W}$ being $\widehat{P}$-Brownian motion, such that*

$$J^n(t,x) = h(X_T^{\pi^*}) + \int_t^T [G_s^n(Z_s^{n-1}, Z_s^n) - \beta Y_s^n]ds - \int_t^T Z_s^n d\widehat{W}_s, \tag{61}$$

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 6-(i), noting that J^n satisfies the PDE (21) 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 (24). Next we bound $|G_s^n(z',Z') - G_s(z,Z)|$.

Lemma 16 *Let Assumptions 1 and 7 hold. Then for $|z|, |Z|, |Z'| \le C$, there exists $L > 0$ (independent of γ) such that*

$$\begin{aligned} |G_s^n(z',Z') - G_s(z,Z)| \le & L\left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma} e^{\frac{L}{\gamma}}\right)|z'-z| + L|Z'-Z| \\ & + L\left(1 + \frac{1}{\gamma} e^{\frac{L}{\gamma}}\right) \sup_a |(b-b^n)(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z'| \ \ a.s. \end{aligned} \tag{62}$$

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 5,

$$|G_s(z',Z') - G_s(z,Z)| \le L\left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma} e^{\frac{L}{\gamma}}\right)|z'-z| + L|Z'-Z| \ \ a.s. \tag{63}$$

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 (23) and (59) respectively. It is easy to see that

$$|G_s^n(z',Z') - G_s(z',Z')| \le 2(\bar{b}\bar{\sigma}C + \bar{r})E_1 + E_2, \tag{64}$$

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 (28) yields

$$E_1 \le \frac{1}{\gamma} \exp\left(\frac{2}{\gamma}(\bar{b}\bar{\sigma}C + \bar{r})\right) \sup_a |(b-b^n)(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z'| \ \ a.s. \tag{65}$$

Moreover,

$$\begin{aligned}
E_2 \le & \int_{\mathcal{A}} \left| b^n(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z' - b(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z' \right| \pi^n(a \,|\, s, X^{\pi^*}, z')da \\
& + \int_{\mathcal{A}} b(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z' \left| \pi^n(a \,|\, s, X^{\pi^*}, z') - \pi(a \,|\, s, X^{\pi^*}, z') \right| da \\
& + \int_{\mathcal{A}} |r(t, x, a)||\pi^n(a \,|\, s, X^{\pi^*}, z') - \pi(a \,|\, s, X^{\pi^*}, z')|da \\
& + \gamma \left| \log \left(\frac{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma}(b^n(s, X_s^{\pi^*}, a)\, \sigma^{-1}(s, X_s^{\pi^*})z' + r(s, X_s^{\pi^*}, a)))\right) da}{\int_{\mathcal{A}} \exp\left(\frac{1}{\gamma}(b(s, X_s^{\pi^*}, a)\, \sigma^{-1}(s, X_s^{\pi^*})z' + r(s, X_s^{\pi^*}, a)))\right) da} \right) \right| \qquad (66) \\
\le & \, 2 \sup_a |(b - b^n)(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z'| + 2(\bar{b}\bar{\sigma}C + \bar{r})E_1 \\
\le & \, 2\left(1 + \frac{\bar{b}\bar{\sigma}C + \bar{r}}{\gamma} \exp\left(\frac{2}{\gamma}(\bar{b}\bar{\sigma}C + \bar{r})\right)\right) \sup_a |(b - b^n)(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z'| \ \ a.s.
\end{aligned}$$

By (64), (65) and (66), we get

$$\begin{aligned}
& |G_s^n(z', Z') - G_s(z', Z')| \\
& \quad \le 2\left(1 + \frac{2(\bar{b}\bar{\sigma}C + \bar{r})}{\gamma} \exp\left(\frac{2}{\gamma}(\bar{b}\bar{\sigma}C + \bar{r})\right)\right) \sup_a |(b - b^n)(s, X_s^{\pi^*}, a)\sigma^{-1}(s, X_s^{\pi^*})z'| \ \ a.s.
\end{aligned} \qquad (67)$$

Combining (63) and (67) yields (62). ■

We also need the following variant of Lemma 4 on the BSDE.

Lemma 17 *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 $\theta \ge M_1 + (1 + \eta^{-1})(M_2 + M_3 + \delta)^2$,

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

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 (Kerimkulov et al., 2020, Lemma A.5) permits to conclude. ■

Now we give the proof of Theorem 8.

Proof (of Theorem 8) 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 15, we have $J^n(t,x) = Y_t^n$ and $J^*(t,x) = Y_t^{t,x}$, where

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

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 16 and the assumption that $|\nabla J^n| \le M$ for all n, we get

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

Applying Lemma 17 with $\delta = L\left(1 + \frac{1}{\gamma} e^{\frac{L}{\gamma}}\right) |q(\cdot;\pi^{n-1}) - q^{\phi_n}|_\infty$, $M_1 = \beta$, $M_2 = L\left(1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma} e^{\frac{L}{\gamma}}\right)$ and $M_3 = L$, we have for $\theta \ge \beta + (1+\eta^{-1})L^2 \left[1 + e^{\frac{L}{\gamma}} + \frac{1}{\gamma} e^{\frac{L}{\gamma}} + L\left(1 + \frac{1}{\gamma} e^{\frac{L}{\gamma}}\right) |q(\cdot;\pi^{n-1}) - q^{\phi_n}|_\infty\right]^2$,

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

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

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

This yields the bound (43). ■

We proceed to proving Theorem 10.

Proof (of Theorem 10) By the argument of (Benveniste et al., 1990, Theorem 22), Assumption 9 (1)–(3) and the condition on $\alpha_{\phi,n}$ imply

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

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 9 (4), we have:

$$|q(\cdot;\pi^*) - q^{\phi_n}|_\infty \le \Delta + C\varepsilon^{-\frac{\rho_\phi}{2}} n^{-\frac{\nu\rho_\phi}{2}}. \tag{68}$$

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 9 (5), we get for Δ sufficiently small,

$$|q(\cdot;\pi^n)-q(\cdot;\pi^*)|_\infty\le C(\Delta+\varepsilon^{-\frac{\rho_\phi}{2}}n^{-\frac{\nu\rho_\phi}{2}}). \tag{69}$$

Combining (68) and (69) 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 8 and (44) yields the bound (48). ■

Finally, we prove Theorem 14.

Proof (of Theorem 14) Note that

$$\begin{aligned}|\mathring{J}^n(t,x)-\mathring{J}(t,x)|&=|\mathring{J}^n(t,x)-J(t,x;\pi^n)+J(t,x;\pi^n)-J^*(t,x)+J^*(t,x)-\mathring{J}(t,x)|\\&\le|\mathring{J}^n(t,x)-J(t,x;\pi^n)|+|J(t,x;\pi^n)-J^*(t,x)|+|J^*(t,x)-\mathring{J}(t,x)|.\end{aligned} \tag{70}$$

It follows from Theorem 10 that

$$|J(t,x;\pi^n)-J^*(t,x)|\le C\left(\Delta+\frac{1}{\varepsilon^{\rho_\phi/2}}n^{-\frac{\nu\rho_\phi}{4}}(\ln n)^{\frac{1}{2}}\right). \tag{71}$$

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 12 (4) and that $|q^{\phi_*}|_\infty<\infty$, we conclude that $|\log\pi^n(a\,|\,t,x)|$ is uniformly bounded. As a result,

$$|\mathring{J}^n(t,x)-J(t,x;\pi^n)|\le\frac{C\gamma}{\beta}\quad\text{for some } C>0. \tag{72}$$

Combining (70), (71) and (72) yields (57). ■

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 8, and provide general sufficient conditions on the model parameters to ensure Assumptions 9 and 12 (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.

Acknowledgments

We thank the Action Editor and three reviewers for careful reading and constructive suggestions that help us improve the paper substantially. Tang gratefully acknowledges financial support through NSF CAREER Award DMS-2538791 and NSF MFAI Award DMS-2602038. Zhou gratefully acknowledges financial supports through the Nie Center for Intelligent Asset Management at Columbia University. Tang and Zhou receive support from the Columbia Innovation Hub grant, and are part of a Columbia-CityU/HK collaborative project that is supported by InnoHK Initiative, The Government of the HKSAR and the AIFT Lab.

References

Richard Bellman. *Dynamic programming*. Princeton University Press, Princeton, NJ, 1957.

Christian Bender and Nguyen Tran Thuan. On the grid-sampling limit SDE. 2024. arXiv:2410.07778.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2023.

Patrick Cheridito, H. Mete Soner, Nizar Touzi, and Nicolas Victoir. Second-order backward stochastic differential equations and fully nonlinear parabolic PDEs. *Comm. Pure Appl. Math.*, 60(7):1081–1110, 2007.

Sinho Chewi. *Log-concave sampling*. 2021. Available at `https://chewisinho.github.io/main.pdf`.

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

K. D. Elworthy and X.-M. Li. Formulae for the derivatives of heat semigroups. *J. Funct. Anal.*, 125(1):252–286, 1994.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Neurips*, volume 36, 2024.

Wendell H. Fleming and H. Mete Soner. *Controlled Markov processes and viscosity solutions*, volume 25 of *Stochastic Modelling and Applied Probability*. Springer, New York, second edition, 2006.

Avner Friedman. *Partial differential equations of parabolic type*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1964.

Xuefeng Gao, Jiale Zha, and Xun Yu Zhou. Reward-directed score-based diffusion models via q-learning. *J. Mach. Learn. Res.*, 26:Paper No. [302], 46, 2025.

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

Yilie Huang and Xun Yu Zhou. Data-driven exploration for a class of continuous-time indefinite linear–quadratic reinforcement learning problems. *IEEE Trans. Autom. Control.*, 2026.

Yilie Huang, Yanwei Jia, and Xun Yu Zhou. Mean–variance portfolio selection by continuous-time reinforcement learning: Algorithms, regret analysis, and empirical study. 2024. arXiv:2412.16175.

Yilie Huang, Yanwei Jia, and Xun Yu Zhou. Sublinear regret for a class of continuous-time linear-quadratic reinforcement learning problems. *SIAM J. Control Optim.*, 63(5): 3452–3474, 2025a.

Yilie Huang, Wenpin Tang, and Xun Yu Zhou. ART for diffusion sampling: continuous-time control and actor-critic learning. 2026. arXiv:2607.02137.

Yu-Jui Huang, Zhenhua Wang, and Zhou Zhou. Convergence of policy improvement for entropy-regularized stochastic control problems. *SIAM J. Control Optim.*, 63(2):752–777, 2025b.

Yanwei Jia and Xun Yu Zhou. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *J. Mach. Learn. Res.*, 23(154):1–55, 2022a.

Yanwei Jia and Xun Yu Zhou. Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *J. Mach. Learn. Res.*, 23(154):1–55, 2022b.

Yanwei Jia and Xun Yu Zhou. q-learning in continuous time. *J. Mach. Learn. Res.*, 24 (161):1–61, 2023.

Yanwei Jia and Xun Yu Zhou. Erratum to "q-learning in continuous time". *J. Mach. Learn. Res.*, 2025.

Yanwei Jia, Du Ouyang, and Yufei Zhang. Accuracy of discretely sampled stochastic policies in continuous-time reinforcement learning. *SIAM J. Control Optim.*, 64(3):1889–1929, 2026.

Gopinath Kallianpur. *Stochastic filtering theory*, volume 13 of *Applications of Mathematics*. Springer-Verlag, New York-Berlin, 1980.

Bekzhan Kerimkulov, David Šiška, and Lukasz Szpruch. Exponential convergence and stability of Howard's policy improvement algorithm for controlled diffusions. *SIAM J. Control Optim.*, 58(3):1314–1340, 2020.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*, 17(39):1–40, 2016.

Jin Ma, Gaozhan Wang, and Jianfeng Zhang. Convergence analysis for entropy-regularized control problems: a probabilistic approach. *SIAM J. Control Optim.*, 64(2):816–842, 2026.

Georg Menz, André Schlichting, Wenpin Tang, and Tianqi Wu. Ergodicity of the infinite swapping algorithm at low temperature. *Stochastic Process. Appl.*, 151:519–552, 2022.

Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, and Koray Kavukcuoglu. Learning to navigate in complex environments. In *ICLR*, 2017.

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

Andrew Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *ICML*, volume 17, pages 663–670, 2000.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, et al. Training language models to follow instructions with human feedback. In *Neurips*, volume 35, pages 27730–27744, 2022.

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

Shige 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.

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

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

Stuart Russell. Learning agents for uncertain environments. In *COLT*, pages 101–103, 1998.

Deven Sethi, David ˇ Siˇ ska, and Yufei Zhang. Entropy annealing for policy mirror descent in continuous time and space. *SIAM J. Control Optim.*, 63(4):3006–3041, 2025.

Jiayuan Sheng, Hanyang Zhao, Haoxian Chen, David D Yao, and Wenpin Tang. Understanding sampler stochasticity in training diffusion models for RLHF. 2025. arXiv:2510.10767.

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

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, and Adrian Bolton. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.

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

Jordan M. Stoyanov. *Counterexamples in probability*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1987.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2018.

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

Wenpin Tang, Yuming Paul Zhang, and Xun Yu Zhou. Exploratory HJB equations and their convergence. *SIAM J. Control Optim.*, 60(6):3191–3216, 2022.

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

Hung Vinh Tran, Zhenhua Wang, and Yuming Paul Zhang. Policy iteration for exploratory Hamilton-Jacobi-Bellman equations. *Appl. Math. Optim.*, 91(2):Paper No. 50, 29, 2025.

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

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

Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *J. Mach. Learn. Res.*, 21:1–34, 2020.

Jiongmin Yong and Xun Yu Zhou. *Stochastic controls – Hamiltonian systems and HJB equations*, volume 43 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1999.

Zikun Zhang, Jiayuan Sheng, David D Yao, and Wenpin Tang. A continuous-time reinforcement learning framework for fine-tuning discrete diffusion models. 2026. arXiv:2607.14522.

Hanyang Zhao, Wenpin Tang, and David D Yao. Policy optimization for continuous reinforcement learning. In *Neurips*, 2023.

Hanyang Zhao, Haoxian Chen, Ji Zhang, David D Yao, and Wenpin Tang. Scores as Actions: a framework of fine-tuning diffusion models by continuous-time reinforcement learning. 2024. arXiv:2409.08400.

Hanyang Zhao, Haoxian Chen, Ji Zhang, David D Yao, and Wenpin Tang. Score as action: Fine tuning diffusion generative models by continuous-time reinforcement learning. In *ICML*, pages 77371–77389, 2025.

Hanyang Zhao, Wenpin Tang, and David D Yao. Policy optimization for reinforcement learning in continuous time and space. *Appl. Math. Optim.*, 94(1):Paper No. 16, 41, 2026.

Xun Yu 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.