Journal of Machine Learning Research 22 (2021) 1-34 Submitted 10/20; Revised 8/21; Published 9/21

Hamilton–Jacobi Deep Q-Learning for Deterministic Continuous-Time Systems with Lipschitz Continuous Controls

Jeongho Kim JHKIM206@SNU.AC.KR
Institute of New Media and Communications
Seoul National University
Seoul 08826, South Korea

Jaeuk Shin SJU5379@SNU.AC.KR
Department of Electrical and Computer Engineering
Automation and Systems Research Institute
Seoul National University
Seoul 08826, South Korea

Insoon Yang INSOONYANG@SNU.AC.KR
Department of Electrical and Computer Engineering
Automation and Systems Research Institute
Seoul National University
Seoul 08826, South Korea

Editor: Jan Peters

Abstract

In this paper, we propose Q-learning algorithms for continuous-time deterministic optimal control problems with Lipschitz continuous controls. A new class of Hamilton–Jacobi–Bellman (HJB) equations is derived from applying the dynamic programming principle to continuous-time Q-functions. Our method is based on a novel semi-discrete version of the HJB equation, which is proposed to design a Q-learning algorithm that uses data collected in discrete time without discretizing or approximating the system dynamics. We identify the conditions under which the Q-function estimated by this algorithm converges to the optimal Q-function. For practical implementation, we propose the *Hamilton–Jacobi DQN*, which extends the idea of deep Q-networks (DQN) to our continuous control setting. This approach does not require actor networks or numerical solutions to optimization problems for greedy actions since the HJB equation provides a simple characterization of optimal controls via ordinary differential equations. We empirically demonstrate the performance of our method through benchmark tasks and high-dimensional linear-quadratic problems.

Keywords: Q-learning, Deep Q-networks, Continuous-time dynamical systems, Optimal control, Hamilton–Jacobi–Bellman equations

1. Introduction

Model-free reinforcement learning (RL) algorithms provide an effective data-driven solution to sequential decision-making problems—in particular, to problems in the discrete-time setting (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998; Szepesvari, 2010). Recently, there has been a growing interest in and demand for applying these techniques to complex physical control tasks, motivated by robotic and autonomous systems. However, many physical processes evolve in continuous time, requiring RL methods that can systematically handle continuous-time dynamical systems. These systems are often described by deterministic ordinary differential equations (ODEs). Classical approaches first estimate the model parameters using system identification techniques and then design a suitable model-based controller (for example, (Ljung, 1998)). However, we do not often have such a luxury of having a separate training period for parameter identification, which often requires large-scale high-resolution data. Furthermore, when the model parameters change over time, the classical techniques have fundamental limitations in terms of adaptivity. The focus of this work is to study a control-theoretic model-free RL method that extends the popular Q-learning (Watkins and Dayan, 1992) and deep Q-networks (DQN) (Mnih et al., 2015) to the continuous-time deterministic optimal control setting.

One of the most straightforward ways to tackle such continuous-time control problems is to discretize time, state, and action, and then employ an RL algorithm for discrete Markov decision processes (MDPs). However, this approach could easily be rendered ineffective when a fine discretization is used (Doya, 2000). To avoid the explicit discretization of state and action, several methods have been proposed using function approximators (Gordon, 1995). Among those, algorithms that use deep neural networks as function approximators provide strong empirical evidence for learning high-performance policies on a range of benchmark tasks (Todorov et al., 2012; Brockman et al., 2016; Duan et al., 2016; Tassa et al., 2018). To deal with continuous action spaces, such discrete-time model-free deep RL methods numerically solve optimization problems for greedy actions (Ryu et al., 2020) or use parameterized policies and learn the network parameters via policy gradient (Schulman et al., 2015, 2017), actor-critic methods (Lillicrap et al., 2015; Mnih et al., 2016; Haarnoja et al., 2018; Fujimoto et al., 2018; Tessler et al., 2019), or normalized advantage functions (Gu et al., 2016). However, these algorithms do not exploit the structures and characteristics of continuous-time system dynamics. Moreover, it is often ignored to analyze how the size of sampling intervals affects such discrete-time methods.

The literature regarding continuous-time RL is relatively limited; most of them have tried to avoid explicit discretization using the structural properties of limited classes of system dynamics (for example, see (Palanisamy et al., 2015; Bian and Jiang, 2016; Vamvoudakis, 2017; Jiang and Jiang, 2015; Kim and Yang, 2020a; Bhasin et al., 2013; Modares and Lewis, 2014; Vamvoudakis and Lewis, 2010) for linear or control-affine systems, and see (Bradtke and Duff, 1995) for semi-MDPs with finite state and action spaces). We also refer to (Munos, 2006), where the policy gradient method in continuous time is introduced. However, in this framework the reward function does not depend on the control signal.

In general continuous-time cases, the dynamic programming equation is expressed as a Hamilton–Jacobi–Bellman (HJB) equation that provides a sound theoretical framework. Previous methods use HJB equations for learning the optimal *state-value function* or its gradient via convergent discretization (Munos, 2000), barycentric interpolation (Munos and Moore, 1999), advantage functions (Dayan and Singh, 1996), temporal difference algorithms (Doya, 2000), kernel-based approximations (Ohnishi et al., 2018), adaptive dynamic programming (Yang et al., 2017), path integrals (Theodorou et al., 2010; Rajagopal et al., 2017), and neural network approximation (Tassa and Erez, 2007; Lutter et al., 2020).

However, to our knowledge, HJB equations have not been studied to admit Q-functions (or state-action value functions) as solutions in the previous methods, although there have been a few attempts to construct variants of Q-functions for continuous-time dynamical systems. In (Kontoudis and Vamvoudakis, 2019), a Q-function for linear time-invariant systems is defined as the sum of the optimal state-value function and the Hamiltonian. Another variant of Q-functions is introduced as the sum of the running cost and the directional derivative of the state-value function (Mehta and Meyn, 2009), which is then approximated by a parameterized family of functions. However, the definitions of Q-functions in these works are different from the standard state-action value function that is defined as the maximum expected cumulative reward incurred after starting from a particular state with a specific action. Moreover, they have only used HJB equations for state-value functions, without introducing HJB equations for the constructed Q-functions. The practical performance of these methods has been demonstrated only through low-dimensional tasks. More recently, (Tallec et al., 2019) devises a new method combining advantage updating (Baird, 1994) and existing off-policy RL algorithms to propose continuous-time RL algorithms that are robust to time discretization. However, to tackle problems with continuous action spaces, this method uses actor-critic methods rather than relying solely on state-value functions.

In this work, we consider continuous-time deterministic optimal control problems with Lipschitz continuous controls in the infinite-horizon discounted setting. We show that the standard Q-function coincides with the state-value function without any particular constraints on control trajectories. This observation motivates us to introduce Lipschitz constraints on controls. Applying the dynamic programming principle to the continuous-time Q-function, we derive a novel class of HJB equations. The HJB equation is shown to admit a unique viscosity solution, which corresponds to the optimal Q-function. To the best of our knowledge, this is the first attempt to rigorously characterize the HJB equations for Q-functions in continuous-time control. The HJB equations provide a simple model-free characterization of optimal controls via ODEs and a theoretical basis for our Q-learning method. We propose a new semi-discrete version of the HJB equation to obtain a Q-learning algorithm that uses sample data collected in discrete time without discretizing or approximating the continuous-time dynamics. By design, it attains the flexibility to choose the sampling interval to take into account the features of continuous-time systems, but without the need for sophisticated ODE discretization methods. We provide a convergence analysis that suggests a sufficient condition on the sampling interval and stepsizes for the Q-functions generated by our method to converge to the optimal Q-function. This study may open a new avenue of research that connects HJB equations and Q-learning domain.

For a practical implementation of our HJB-based Q-learning, we combine it with the idea of DQN. This new model-free deep RL algorithm, which we call the *Hamilton-Jacobi DQN* (HJ DQN), is as simple as DQN but capable of solving continuous-time problems without discretizing the system dynamics or the action space. Instead of using any parameterized policy or numerically optimizing the estimated Q-functions to compute greedy actions, HJ DQN benefits from the simple ODE characterization of optimal controls, which is obtained in our theoretical analysis of the HJB equations. Thus, our algorithm is computationally light and easy to implement, thereby requiring less hyperparameter tuning compared to actor-critic methods for continuous control. We evaluate our algorithm on OpenAI benchmark tasks and high-dimensional linear-quadratic (LQ) control problems. The results of our experiments suggest that actor networks in actor-critic methods may be replaced by the optimal control obtained via our HJB equation.

This paper is significantly expanded from a preliminary conference version (Kim and Yang, 2020b). A novel semi-discrete HJB equation is proposed and analyzed in this paper to provide a theoretical basis for our method. A Q-learning algorithm and its DQN variant are newly designed in a principled manner to use transition data collected in discrete time with a theoretically consistent target. Furthermore, convergence properties of our Q-learning method are carefully studied in this paper. This paper also contains the results of more thorough numerical experiments for several benchmark tasks and high-dimensional LQ control problems, as well as design evaluations.

The remainder of this paper is organized as follows. In Section 2, we define the Q-functions for continuous-time optimal control problems with Lipschitz continuous controls and derive the associated HJB equations. We also characterize optimal control dynamics via an ODE. In Section 3, we propose a Q-learning method based on the semi-discrete HJB equation and analyze its convergence properties. In Section 4, we introduce the HJ DQN algorithm and discuss its features. Section 5 provides the results of our experiments on benchmark problems as well as LQ control problems. All the mathematical proofs are contained in Appendix B.

2. Hamilton–Jacobi–Bellman Equations for Q-Functions

Consider a continuous-time dynamical system of the form

$$\dot{x}(t) = f(x(t), a(t)), \quad t > 0, \tag{1}$$

where $x(t) \in \mathbb{R}^n$ and $a(t) \in \mathbb{R}^m$ are the system state and the control action, respectively.[1] Here, the vector field $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is an unknown function. The standard infinite-horizon discounted optimal control problem can be formulated as

$$\sup_{a \in \mathcal{A}} J_{\boldsymbol{x}}(a) := \int_0^\infty e^{-\gamma t} r(x(t), a(t))\, \mathrm{d}t, \tag{2}$$

with $x(0) = \boldsymbol{x}$, where $r : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is an unknown reward function of interest and $\gamma > 0$ is a discount factor.[2] We follow the convention in continuous-time deterministic optimal control that considers *control trajectory, instead of control policy*, as the optimization variable (Bardi and Capuzzo-Dolcetta, 1997).

The (continuous-time) Q-function of (2) is defined as

$$Q(\boldsymbol{x}, \boldsymbol{a}) := \sup_{a \in \mathcal{A}} \left\{ \int_0^\infty e^{-\gamma t} r(x(t), a(t))\, \mathrm{d}t \mid x(0) = \boldsymbol{x}, a(0) = \boldsymbol{a} \right\}, \tag{3}$$

which represents the maximal reward incurred from time 0 when starting from $x(0) = \boldsymbol{x}$ with $a(0) = \boldsymbol{a}$. Suppose for a moment that the set of admissible controls $\mathcal{A}$ has no particular constraints, that is, $\mathcal{A} := \big\{ a : \mathbb{R}_{\geq 0} \to \mathbb{R}^m \mid a \text{ measurable} \big\}$. Then, $Q(\boldsymbol{x}, \boldsymbol{a})$ is reduced to the

[1] Here, $\dot{x}$ denotes $\mathrm{d}x/\mathrm{d}t$.

[2] Although the focus of this work is deterministic control, one may also consider its stochastic counterpart. We briefly discuss the extension of our method to the stochastic control setting in Appendix C.

standard optimal value function $v(\boldsymbol{x}) := \sup_{a \in \mathcal{A}} \big\{ \int_0^\infty e^{-\gamma t} r(x(t), a(t))\, \mathrm{d}t \mid x(0) = \boldsymbol{x} \big\}$ for all $\boldsymbol{a} \in \mathbb{R}^m$ since the action can be switched immediately from $\boldsymbol{a}$ to an optimal control and in this case $\boldsymbol{a}$ does not affect the total reward or the system trajectory in the continuous-time setting.

Proposition 1 *Suppose that $\mathcal{A} := \big\{a : \mathbb{R}_{\geq 0} \to \mathbb{R}^m \mid a$ is measurable$\big\}$. Then, the optimal Q-function* (3) *corresponds to the optimal value function v for each $\boldsymbol{a} \in \mathbb{R}^m$, that is, $Q(\boldsymbol{x}, \boldsymbol{a}) = v(\boldsymbol{x})$ for all $(\boldsymbol{x}, \boldsymbol{a}) \in \mathbb{R}^n \times \mathbb{R}^m$.*

Thus, if $\mathcal{A}$ is chosen as above, the Q-function has no additional interesting property under the standard choice of $\mathcal{A}$. This observation is consistent with the previously reported result on the continuous time limit of Q-functions (Baird, 1994; Tallec et al., 2019). Motivated by the observation, we restrict the control $a(t)$ to be a Lipschitz continuous function in t. Since any Lipschitz continuous function is differentiable almost everywhere, we choose the set of admissible controls as

$$\mathcal{A} := \big\{a : \mathbb{R}_{\geq 0} \to \mathbb{R}^m \mid a \text{ is measurable, } |\dot{a}(t)| \leq L \text{ a.e.}\big\},$$

where $|\cdot|$ denotes the standard Euclidean norm, and L is a fixed constant. From now on, we will focus on the optimal control problem (2) with Lipschitz continuous controls, and the corresponding Q-function (3).

Our first step is to study the structural properties of the optimality equation and the optimal control via dynamic programming. Using the discovered structural properties, a DQN-like algorithm is then designed to solve the optimal control problem (2) in a model-free manner.

2.1 Dynamic Programming and HJB Equations

By the dynamic programming principle, we have

$$\begin{aligned} Q(\boldsymbol{x}, \boldsymbol{a}) = \sup_{a \in \mathcal{A}} \bigg\{ \int_t^{t+h} & e^{-\gamma(s-t)} r(x(s), a(s))\, \mathrm{d}s \\ & + e^{-\gamma h} \int_{t+h}^{\infty} e^{-\gamma(s-(t+h))} r(x(s), a(s))\, \mathrm{d}s \mid x(t) = \boldsymbol{x}, a(t) = \boldsymbol{a} \bigg\} \\ = \sup_{a \in \mathcal{A}} \bigg\{ \int_t^{t+h} & e^{-\gamma(s-t)} r(x(s), a(s))\, \mathrm{d}s + e^{-\gamma h} Q(x(t+h), a(t+h)) \mid x(t) = \boldsymbol{x}, a(t) = \boldsymbol{a} \bigg\} \end{aligned}$$

for any $h > 0$. Rearranging this equality, we obtain

$$\begin{aligned} 0 = \sup_{a \in \mathcal{A}} \bigg\{ \frac{1}{h} \int_t^{t+h} & e^{-\gamma(s-t)} r(x(s), a(s))\, \mathrm{d}s + \frac{1}{h}[Q(x(t+h), a(t+h)) - Q(\boldsymbol{x}, \boldsymbol{a})] \\ & + \frac{e^{-\gamma h} - 1}{h} Q(x(t+h), a(t+h)) \;\bigg|\; x(t) = \boldsymbol{x}, a(t) = \boldsymbol{a} \bigg\}. \end{aligned}$$

Letting h tend to zero and assuming for a moment that the Q-function is continuously differentiable, its Taylor expansion yields

$$\gamma Q(\boldsymbol{x}, \boldsymbol{a}) - \sup_{\boldsymbol{b} \in \mathbb{R}^m, |\boldsymbol{b}| \leq L} \{\nabla_{\boldsymbol{x}} Q \cdot f(\boldsymbol{x}, \boldsymbol{a}) + \nabla_{\boldsymbol{a}} Q \cdot \boldsymbol{b} + r(\boldsymbol{x}, \boldsymbol{a})\} = 0, \tag{4}$$

where the optimization variable $\boldsymbol{b}$ represents $\dot{a}(t)$, and the constraint $|\boldsymbol{b}| \leq L$ is due to the Lipschitz constraint on control trajectories, $|\dot{a}(t)| \leq L$ a.e. These two are the distinctive features of our HJB equation compared to the standard HJB equation in which the optimization variable is the action itself, $\boldsymbol{a}$. Since the terms $\nabla_{\boldsymbol{x}} Q \cdot f(\boldsymbol{x}, \boldsymbol{a})$ and $r(\boldsymbol{x}, \boldsymbol{a})$ are independent of $\boldsymbol{b}$, the supremum in (4) is attained at $\boldsymbol{b}^\star = L \frac{\nabla_{\boldsymbol{a}} Q}{|\nabla_{\boldsymbol{a}} Q|}$.[3] Thus, we obtain

$$\gamma Q(\boldsymbol{x}, \boldsymbol{a}) - \nabla_{\boldsymbol{x}} Q \cdot f(\boldsymbol{x}, \boldsymbol{a}) - L|\nabla_{\boldsymbol{a}} Q| - r(\boldsymbol{x}, \boldsymbol{a}) = 0, \tag{5}$$

which is the *HJB equation for the Q-function.* However, the Q-function is not continuously differentiable in general. This motivates us to consider a weak solution of the HJB equation. Among several types of weak solutions, it is shown in Appendix A that the Q-function corresponds to the unique *viscosity solution* (Crandall and Lions, 1983) of the HJB equation under the following assumption:

Assumption 1 *The functions f and r are bounded and Lipschitz continuous, that is, there exists a constant C such that $\|f\|_{L^\infty} + \|r\|_{L^\infty} < C$ and $\|f\|_{\mathrm{Lip}} + \|r\|_{\mathrm{Lip}} < C$, where $\|\cdot\|_{\mathrm{Lip}}$ denotes a Lipschitz constant of argument.*

2.2 Optimal Controls

In the derivation of the HJB equation above, we deduce that an optimal control a must satisfy $\dot{a} = L \frac{\nabla_{\boldsymbol{a}} Q}{|\nabla_{\boldsymbol{a}} Q|}$ when Q is differentiable. Since the Q-function is the unique viscosity solution of our HJB equation, the viscosity solution framework (Bardi and Capuzzo-Dolcetta, 1997) can be used to obtain the following more rigorous characterization of optimal controls when the Q-function is not differentiable.

Theorem 2 *Suppose that Assumption 1 holds. Consider a control trajectory $a^\star(s)$, $s \geq t$, defined by*

$$\dot{a}^\star(s) = L \frac{p_2}{|p_2|} \quad \forall p = (p_1, p_2) \in D^\pm Q(x^\star(s), a^\star(s)) \tag{6}$$

for a.e. $s \geq t$, and $a^\star(t) = \boldsymbol{a}$, where $\dot{x}^\star = f(x^\star, a^\star)$ for $s \geq t$ and $x^\star(t) = \boldsymbol{x}$.[4] Assume that the function Q is locally Lipschitz in a neighborhood of $(x^\star(s), a^\star(s))$ and that $D^+ Q(x^\star(s), a^\star(s)) = \partial Q(x^\star(s), a^\star(s))$ for a.e. $s \geq t$.[5] Then, $a^\star$ is optimal among those in $\mathcal{A}$ such that $a(t) = \boldsymbol{a}$, that is, it satisfies

$$a^\star \in \arg\max_{a \in \mathcal{A}} \left\{ \int_t^\infty e^{-\gamma(s-t)} r(x(s), a(s)) \,\mathrm{d}s \;\middle|\; x(t) = \boldsymbol{x}, a(t) = \boldsymbol{a} \right\}. \tag{7}$$

[3] Given a vector $c \in \mathbb{R}^m$, the optimal solution of $\max_{\boldsymbol{b} \in \mathbb{R}^m, |\boldsymbol{b}| \leq L} c \cdot \boldsymbol{b}$ is $\boldsymbol{b}^\star = L\frac{c}{|c|}$ since $c \cdot \boldsymbol{b} \leq |c||\boldsymbol{b}| \leq L|c|$ and the inequalities hold with equality when $\boldsymbol{b} = \boldsymbol{b}^\star$.

[4] Here, D^+Q and D^-Q denote the super- and sub-differentials of Q, respectively, and $D^\pm Q := D^+Q \cup D^-Q$. The superdifferential and subdifferential are defined as $D^+Q(\boldsymbol{x}, \boldsymbol{a}) := \Big\{(p_1, p_2) \in \mathbb{R}^{n+m} \mid \limsup_{(\boldsymbol{x}', \boldsymbol{a}') \to (\boldsymbol{x}, \boldsymbol{a})} \frac{Q(\boldsymbol{x}', \boldsymbol{a}') - Q(\boldsymbol{x}, \boldsymbol{a}) - (p_1, p_2) \cdot (\boldsymbol{x}' - \boldsymbol{x}, \boldsymbol{a}' - \boldsymbol{a})}{|(\boldsymbol{x}', \boldsymbol{a}') - (\boldsymbol{x}, \boldsymbol{a})|} \leq 0\Big\}$ and $D^-Q(\boldsymbol{x}, \boldsymbol{a}) := \Big\{(p_1, p_2) \in \mathbb{R}^{n+m} \mid \liminf_{(\boldsymbol{x}', \boldsymbol{a}') \to (\boldsymbol{x}, \boldsymbol{a})} \frac{Q(\boldsymbol{x}', \boldsymbol{a}') - Q(\boldsymbol{x}, \boldsymbol{a}) - (p_1, p_2) \cdot (\boldsymbol{x}' - \boldsymbol{x}, \boldsymbol{a}' - \boldsymbol{a})}{|(\boldsymbol{x}', \boldsymbol{a}') - (\boldsymbol{x}, \boldsymbol{a})|} \geq 0\Big\}$. At a point $(\boldsymbol{x}, \boldsymbol{a})$ where Q is differentiable, the super- and sub-differentials are identical to the singleton of the classical derivative of Q.

[5] Here, ∂Q denotes the Clarke's generalized gradient of Q (see, for example, p. 63 of (Bardi and Capuzzo-Dolcetta, 1997)). Note that the right-hand side of ODE (6) can be arbitrarily chosen when $p_2 = 0$.

If, in addition,

$$\boldsymbol{a} \in \arg\max_{\boldsymbol{a}' \in \mathbb{R}^m} Q(\boldsymbol{x}, \boldsymbol{a}'),$$

then $a^\star$ is an optimal control, that is, it satisfies

$$a^\star \in \arg\max_{a \in \mathcal{A}} \left\{ \int_t^\infty e^{-\gamma(s-t)} r(x(s), a(s))\, \mathrm{d}s \;\middle|\; x(t) = \boldsymbol{x} \right\}.$$

Note that at a point $(\boldsymbol{x}, \boldsymbol{a})$ where Q is differentiable, the ODE (6) is simplified to $\dot{a}^\star = L \frac{\nabla_{\boldsymbol{a}} Q(x^\star, a^\star)}{|\nabla_{\boldsymbol{a}} Q(x^\star, a^\star)|}$. A useful implication of this theorem is that for any $\boldsymbol{a} \in \mathbb{R}^m$, an optimal control in $\mathcal{A}$ such that $a(t) = \boldsymbol{a}$ can be obtained using the ODE (6) with the initial condition $a^\star(t) = \boldsymbol{a}$. Thus, when the control is initialized as an arbitrary value $\boldsymbol{a}$ at arbitrary time t in Q-learning, we can still use the ODE (6) to obtain an optimal control. Another important implication of Theorem 2 is that an optimal control can be constructed without numerically solving any optimization problem. This salient feature assists in the design of a computationally efficient DQN algorithm for continuous control without involving any explicit optimization or any actor network.

3. Hamilton–Jacobi Q-Learning

3.1 Semi-Discrete HJB Equations and Asymptotic Consistency

In practice, even though the underlying physical process evolves in continuous time, the observed data, such as sensor measurements, are collected in discrete (sample) time. To design a concrete algorithm for learning the Q-function using such discrete-time data in practical problems, we propose a novel semi-discrete version of the HJB equation (5) *without discretizing or approximating the continuous-time system.* Let $h > 0$ be a fixed *sampling interval*, and let $\mathcal{B} := \{b := \{b_k\}_{k=0}^\infty \mid b_k \in \mathbb{R}^m, |b_k| \leq L\}$, where b_k is analogous to $\dot{a}(t)$ in the continuous-time case. Given $(\boldsymbol{x}, \boldsymbol{a}) \in \mathbb{R}^n \times \mathbb{R}^m$ and a sequence $b \in \mathcal{B}$, we let

$$Q^{h,b}(\boldsymbol{x}, \boldsymbol{a}) := h \sum_{k=0}^\infty r(x_k, a_k)(1 - \gamma h)^k,$$

where $\{(x_k, a_k)\}_{k=0}^\infty$ is defined by $x_{k+1} = \xi(x_k, a_k; h)$ and $a_{k+1} = a_k + h b_k$ with $(x_0, a_0) = (\boldsymbol{x}, \boldsymbol{a})$.[6] Here, $\xi(x_k, a_k; h)$ denotes the state of (1) at time $t = h$ with initial state $x(0) = x_k$ and constant action $a(t) \equiv a_k$, $t \in [0, h)$. It is worth emphasizing that our semi-discrete approximation does *not* approximate the system dynamics and thus is more accurate than the standard semi-discrete method (Section VI, (Bardi and Capuzzo-Dolcetta, 1997)). The optimal semi-discrete Q-function $Q^{h,\star} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is then defined as

$$Q^{h,\star}(\boldsymbol{x}, \boldsymbol{a}) := \sup_{b \in \mathcal{B}} Q^{h,b}(\boldsymbol{x}, \boldsymbol{a}). \tag{8}$$

Then, $Q^{h,\star}$ satisfies a semi-discrete version of the HJB equation (5).

[6] The discounting factor $(1 - \gamma h)$ is the first-order approximation of $\exp(-\gamma h)$ as $h \to 0$. The validity of using the first-order approximation is shown in Proposition 4.

Proposition 3 *Suppose that $0 < h < \frac{1}{\gamma}$. Then, the function $Q^{h,\star}$ is a solution to the following semi-discrete HJB equation:*

$$Q^{h,\star}(\boldsymbol{x}, \boldsymbol{a}) = hr(\boldsymbol{x}, \boldsymbol{a}) + (1 - \gamma h) \sup_{|\boldsymbol{b}| \leq L} Q^{h,\star}(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + h\boldsymbol{b}). \tag{9}$$

Under Assumption 1, $Q^{h,\star}$ coincides with the unique solution of the semi-discrete HJB equation (9). Moreover, the optimal semi-discrete Q-function converges uniformly to its original counterpart in every compact subset of $\mathbb{R}^n \times \mathbb{R}^m$.

Proposition 4 *Suppose that $0 < h < \frac{1}{\gamma}$ and Assumption 1 holds. Then, the function $Q^{h,\star}$ is the unique solution to the semi-discrete HJB equation (9). Furthermore, we have*

$$\lim_{h \to 0} \sup_{(\boldsymbol{x}, \boldsymbol{a}) \in K, K \text{compact}} |Q^{h,\star}(\boldsymbol{x}, \boldsymbol{a}) - Q(\boldsymbol{x}, \boldsymbol{a})| = 0.$$

This proposition justifies the use of the semi-discrete HJB equation for small h. We aim to estimate the optimal Q-function using sample data collected in discrete time, enjoying the benefits of both the semi-discrete HJB equation (9) and the original HJB equation (5). Namely, the semi-discrete version yields to naturally make use of Q-learning and DQN, and the original version provides an optimal control via (6) without requiring a numerical solution for any optimization problems or actor networks as we will see in Section 4.

3.2 Convergence Properties

Consider the following model-free update of Q-functions using the semi-discrete HJB equation (9): In the kth iteration, for each $(\boldsymbol{x}, \boldsymbol{a})$ we collect data $(x_k := \boldsymbol{x}, a_k := \boldsymbol{a}, r_k, x_{k+1})$ and update the Q-function, with learning rate α_k, by

$$Q_{k+1}^h(\boldsymbol{x}, \boldsymbol{a}) := (1 - \alpha_k) Q_k^h(\boldsymbol{x}, \boldsymbol{a}) + \alpha_k \Big[hr_k + (1 - \gamma h) \sup_{|\boldsymbol{b}| \leq L} Q_k^h(x_{k+1}, \boldsymbol{a} + h\boldsymbol{b}) \Big], \tag{10}$$

where x_{k+1} is obtained by running (or simulating) the continuous-time system from x_k with action a_k fixed for h period without any approximation, that is, $x_{k+1} = \xi(x_k, a_k; h)$, and $r_k = r(x_k, a_k)$. We refer to this synchronous Q-learning as *Hamilton–Jacobi Q-learning* (HJ Q-learning). Note that this method is not practically useful because the update must be performed for all state-action pairs in the continuous space. In the following section, we propose a DQN-like algorithm to approximately perform HJ Q-learning by employing deep neural networks as a function approximator. Before doing so, we identify the conditions under which the Q-function updated by (10) converges to the optimal semi-discrete Q-function (8) in L^∞.

Theorem 5 *Suppose that $0 < h < \frac{1}{\gamma}$, $0 \leq \alpha_k \leq 1$ and that Assumption 1 holds. If the sequence $\{\alpha_k\}_{k=0}^\infty$ of learning rates satisfies $\sum_{k=0}^\infty \alpha_k = \infty$, then*

$$\lim_{k \to \infty} \|Q_k^h - Q^{h,\star}\|_{L^\infty} = 0.$$

Finally, by Propositions 4 and Theorem 5, we establish the following convergence result associating HJ Q-learning (10) with the optimal Q-function in the original continuous-time setting.

Corollary 1 *Suppose that $0 \leq \alpha_k \leq 1$ and that Assumption 1 holds. If the sequence $\{\alpha_k\}_{k=0}^{\infty}$ of learning rates satisfies $\sum_{k=0}^{\infty} \alpha_k = \infty$ then, for each $0 < h < \frac{1}{\gamma}$, there exists $k_h \in \mathbb{N}$ such that $h \sum_{\tau=0}^{k_h - 1} \alpha_\tau \to \infty$ as $h \to 0$. Moreover, for such a choice of k_h, we have*

$$\lim_{h \to 0} \sup_{k \geq k_h} \sup_{(\boldsymbol{x}, \boldsymbol{a}) \in K, K \text{compact}} |Q_k^h(\boldsymbol{x}, \boldsymbol{a}) - Q(\boldsymbol{x}, \boldsymbol{a})| = 0.$$

Note that Corollary 1 provides the double limit of our approximate Q-function as it takes into account the approximation error caused by the semi-discrete HJB in addition to the Q-learning update rule. Thus, this result confirms the validity of using our semi-discrete HJB in Q-learning with transition data sampled in discrete time.

4. Hamilton–Jacobi DQN

The convergence result in the previous section suggests that the optimal Q-function can be estimated in a model-free manner via the semi-discrete HJB equation. However, as previously mentioned, it is intractable to directly implement HJ Q-learning (10) over a continuous state-action space. As a practical function approximator, we employ deep neural networks. We then propose the *Hamilton–Jacobi DQN* that approximately performs the update (10) *without discretizing or approximating the continuous-time system.* Since our algorithm has no actor, we only consider a parameterized Q-function $Q_\theta(\boldsymbol{x}, \boldsymbol{a})$, where θ is the parameter vector of the network.

As with DQN, we use a separate target function Q_{θ^-}, where the network parameter vector θ^- is updated more slowly than θ. This allows us to update θ by solving a regression problem with an almost fixed target, resulting in consistent and stable learning (Mnih et al., 2015). We also use experience replay by storing transition data (x_k, a_k, r_k, x_{k+1}) in a buffer with fixed capacity and by randomly sampling a mini-batch of transition data $\{(x_j, a_j, r_j, x_{j+1})\}$ to update the target value. This reduces bias by breaking the correlation between sample data that are sequential states (Mnih et al., 2015).

When setting the target value in DQN, the target Q-function needs to be maximized over all admissible actions, that is, $y_j^- := hr_j + \gamma' \max_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, \boldsymbol{a})$, where $\gamma' := 1 - \gamma h$ is the corresponding semi-discrete discount factor. Evaluating the maximum is tractable in the case of discrete action spaces. However, in our case of continuous action spaces, it is computationally challenging to maximize the target Q-function with respect to the action variable. To resolve this issue, we go back to the original HJB equation and use the corresponding optimal action in Theorem 2. Specifically, we consider the action dynamics (6) with $b_j := L \frac{\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j)}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j)|}$ fixed over sampling interval h to obtain

$$a_{j+1} = a_j + hb_j := a_j + hL \frac{\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j)}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j)|}.$$

Using this optimal control action, we can approximate the maximal target Q-function value as

$$\max_{|\boldsymbol{a} - a_j| \leq hL} Q_{\theta^-}(x_{j+1}, \boldsymbol{a}) \approx Q_{\theta^-}(x_{j+1}, a_j + hb_j).$$

This approximation becomes more accurate as h decreases. In particular, the approximation error has an $O(h^2)$ bound, as shown in the following proposition:

Algorithm 1: Hamilton–Jacobi DQN

```
Initialize Q-function $Q_\theta$ with random weights $\theta$, and target Q-function $Q_{\theta^-}$ with weights $\theta^- = \theta$;
Initialize replay buffer with fixed capacity;
for episode = 1 to $M$ do
    Randomly sample initial state-action pair $(x_0, a_0)$;
    for $k = 0$ to $K$ do
        Execute action $a_k$ and observe reward $r_k$ and the next state $x_{k+1}$;
        Store $(x_k, a_k, r_k, x_{k+1})$ in buffer;
        Sample the random mini-batch $\{(x_j, a_j, r_j, x_{j+1})\}$ from buffer;
        Set $y_j^- := hr_j + (1 - \gamma h) Q_{\theta^-}\big(x_{j+1}, a_j'\big)\ \forall j$ where $a_j' := a_j + hL \frac{\nabla_{\boldsymbol{a}} Q_\theta(x_j, a_j)}{|\nabla_{\boldsymbol{a}} Q_\theta(x_j, a_j)|}$;
        Update $\theta$ by minimizing $\sum_j (y_j^- - Q_\theta(x_j, a_j))^2$;
        Update $\theta^- \leftarrow (1 - \alpha)\theta^- + \alpha\theta$ for $\alpha \ll 1$;
        Set the next action as $a_{k+1} := a_k + hL \frac{\nabla_{\boldsymbol{a}} Q_\theta(x_k, a_k)}{|\nabla_{\boldsymbol{a}} Q_\theta(x_k, a_k)|} + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I_m)$;
    end for
end for
```

Proposition 6 *Suppose that Q_{θ^-} is twice continuously differentiable with bounded first and second derivatives. If $\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j) \neq 0$, we have*

$$\lim_{h \to 0} \left| \max_{|\boldsymbol{a} - a_j| \leq hL} Q_{\theta^-}(x_{j+1}, \boldsymbol{a}) - Q_{\theta^-}(x_{j+1}, a_j + hb_j) \right| = 0.$$

Moreover, the difference above is $O(h^2)$ as $h \to 0$.

The major advantage of using the optimal action obtained in the continuous-time case is to avoid explicitly solving the nonlinear optimization problem $\max_{|\boldsymbol{a} - a_j| \leq hL} Q_{\theta^-}(x_{j+1}, \boldsymbol{a})$, which is computationally demanding. With this choice of target Q-function value and the semi-discrete HJB equation (9), we set the target value as $y_j^- := hr_j + (1 - \gamma h) Q_{\theta^-}(x_{j+1}, a_j + hb_j)$. To mitigate the overestimation of Q-functions, we can employ double Q-learning (Van Hasselt et al., 2016) by simply modifying b_j as $b_j := L \frac{\nabla_{\boldsymbol{a}} Q_\theta(x_j, a_j)}{|\nabla_{\boldsymbol{a}} Q_\theta(x_j, a_j)|}$ to use a greedy action with respect to Q_θ instead of Q_{θ^-}. In this double Q-learning version, Proposition 6 remains valid except for the $O(h^2)$ convergence rate. The network parameter θ can then be trained to minimize the loss function $\sum_j (y_j^- - Q_\theta(x_j, a_j))^2$. For exploration, we add the additional Gaussian noise $\varepsilon \sim N(0, \sigma^2 I_m)$ to generate the next action as $a_{k+1} := a_k + hL \frac{\nabla_{\boldsymbol{a}} Q_\theta(x_k, a_k)}{|\nabla_{\boldsymbol{a}} Q_\theta(x_k, a_k)|} + \varepsilon$. However, one can use other exploration mechanisms such as the solution of a stochastic differential equation (Tallec et al., 2019). The overall algorithm is presented in Algorithm 1.[7]

4.1 Discussion

We now discuss a few notable features of HJ DQN with regard to existing works:

No use of parameterized policies. Most model-free deep RL algorithms for continuous control use actor-critic methods (Lillicrap et al., 2015; Haarnoja et al., 2018; Fujimoto

[7] When $\nabla_{\boldsymbol{a}} Q_\theta(x_j, a_j) = 0$, $\frac{\nabla_{\boldsymbol{a}} Q_\theta(x_j, a_j)}{|\nabla_{\boldsymbol{a}} Q_\theta(x_j, a_j)|}$ is replaced by an arbitrary vector with norm 1 of the same size.

et al., 2018; Tessler et al., 2019) or policy gradient methods (Schulman et al., 2015; Gu et al., 2016) to deal with continuous action spaces. In these methods, by parametrizing policies, the policy improvement step is performed in the space of network weights. By doing so, they avoid solving possibly complicated optimization problems over the policy or action spaces. However, these methods are subject to the issue of being stuck at local optima in the policy (parameter) space due to the use of gradient-based algorithms, as pointed out in the literature regarding policy gradient/search (Kohl and Stone, 2004; Levine and Koltun, 2013; Fazel et al., 2018) and actor-critic methods (Silver et al., 2014). Moreover, it is reported that the policy-based methods are sensitive to hyperparameters (Quillen et al., 2018). Departing from these algorithms, HJ DQN is a value-based method for continuous control without requiring the use of an actor or a parameterized policy. Previous value-based methods for continuous control (for example, (Ryu et al., 2020)) have a computational challenge in finding a greedy action, which requires a solution to a nonlinear program. Our method avoids numerically optimizing Q-functions over the continuous action space via the optimal control (6). This is a notable benefit of using the proposed HJB framework.

Continuous-time control. Many existing RL methods for continuous-time dynamical systems have been designed for linear systems (Palanisamy et al., 2015; Bian and Jiang, 2016; Vamvoudakis, 2017) or control-affine systems (Jiang and Jiang, 2015; Bhasin et al., 2013; Modares and Lewis, 2014; Vamvoudakis and Lewis, 2010), in which value functions and optimal policies can be represented in a simple form. For general nonlinear systems, Hamilton–Jacobi–Bellman equations have been considered as the optimality equations for state-value functions $v(\boldsymbol{x})$ (Doya, 2000; Munos, 2000; Dayan and Singh, 1996; Ohnishi et al., 2018). Unlike these methods, our method uses Q-functions and thus benefits from modern deep RL techniques developed in the literature on DQN. Moreover, as opposed to discrete-time RL methods, HJ DQN does not discretize or approximate the system dynamics $\dot{x} = f(x, a)$ in its algorithm design. Our theoretical analysis in Section 3.2 suggests a sufficient condition on the sampling interval h for convergence.

4.2 Smoothing

A potential defect of our Lipschitz constrained control setting is that the rate of change in action has a constant norm $L\frac{\nabla_{\boldsymbol{a}} Q(x^\star,a^\star)}{|\nabla_{\boldsymbol{a}} Q(x^\star,a^\star)|}$. This is also observed in Algorithm 1, where the action is updated by $hL\frac{\nabla_{\boldsymbol{a}} Q_\theta(x_j,a_j)}{|\nabla_{\boldsymbol{a}} Q_\theta(x_j,a_j)|}$. Therefore, the magnitude of fluctuations in action is always fixed as hL, which may lead to the oscillatory behavior of action. Such oscillatory behaviors are not uncommon in optimal control (for example, bang-bang solutions). To alleviate this potential issue, one may introduce an additional smoothing process when updating action. Inspired by (Abu-Khalaf and Lewis, 2005), we modify the term $\frac{\nabla_{\boldsymbol{a}} Q_\theta(x_j,a_j)}{|\nabla_{\boldsymbol{a}} Q_\theta(x_j,a_j)|}$ by multiplying a smoothing function. Instead of using $hL\frac{\nabla_{\boldsymbol{a}} Q_\theta(x_j,a_j)}{|\nabla_{\boldsymbol{a}} Q_\theta(x_j,a_j)|}$ in the update of action, we suggest the use of

$$hL\frac{\phi(|\nabla_{\boldsymbol{a}} Q_\theta(x_j,a_j)|)\nabla_{\boldsymbol{a}} Q_\theta(x_j,a_j)}{|\nabla_{\boldsymbol{a}} Q_\theta(x_j,a_j)|},$$

where $\phi : [0, +\infty) \to [0, 1]$ is an increasing function with $\phi(0) = 0$ and $\lim_{r\to\infty} \phi(r) = 1$. Typical examples of such a function ϕ include $\phi(r) = \tanh\left(\frac{r}{L}\right)$ and $\phi(r) = \frac{r}{L+r}$. Indeed, it is straightforward to observe that the value $b^\star := L\frac{\phi(|\nabla_a Q|)\nabla_a Q}{|\nabla_a Q|}$ is the maximizer of $b \mapsto \nabla_a Q \cdot b - \int_0^{|b|/L} \phi^{-1}(r)dr$. Comparing this with the maximization problem in the HJB equation (4), the smoothing method can be interpreted as imposing the penalty $-\int_0^{|b|/L} \phi^{-1}(r)dr$ when selecting b. Since b represents the rate of changes in actions, the penalty discourages undesirable oscillations in action trajectories, as confirmed in Section 5.3.

5. Experiments

In this section, we present the empirical performance of our method on benchmark tasks as well as high-dimensional LQ problems. The source code of our HJ DQN implementation is available online.[8]

5.1 Actor Networks vs. Optimal Control ODE

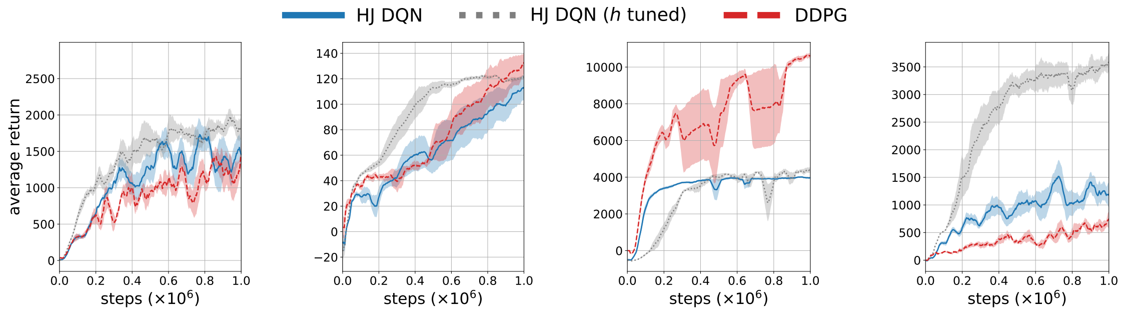
We choose deep deterministic policy gradient (DDPG) (Lillicrap et al., 2015) as a baseline for comparison since it is another variant of DQN for continuous control. DDPG is an actor-critic method that uses separate actor networks while ours is a valued-based method that does not use a parameterized policy. Although there are state-of-the-art methods built upon DDPG, such as TD3 (Fujimoto et al., 2018) and SAC (Haarnoja et al., 2018), we focus on the comparison between ours and DDPG to examine whether the role of actor networks can be replaced by the optimal control characterized through our HJB equation. The hyperparameters used in the experiments are reported in Appendix D.

We consider continuous control benchmark tasks in OpenAI gym (Brockman et al., 2016) simulated by MuJoCo engine (Todorov et al., 2012). Figure 1 shows the learning curves for both methods, each of which is tested with five different random seeds for 1 million steps. The solid curve represents the average of returns over 20 consecutive evaluations while the shaded regions represent half a standard deviation of the average evaluation over five trials. As shown in Figure 1, the performance of our method is comparable to that of DDPG when the default sampling interval is used. Our method outperforms DDPG on Walker2d-v2 while the opposite result is observed in the case of HalfCheetah-v2. As sampling interval h is a hyperparameter of Algorithm 1, we also identify an optimal h for each task aside from the default sampling interval. When we test the different sampling interval, we also tune the learning rate α, as suggested in (Tallec et al., 2019).[9] This additional tuning process improves the final performances and learning speed except in the case of HalfCheetah-v2. Overall, the results indicate that actor networks may be replaced by the ODE characterization (6) of optimal control obtained using our HJB framework. Without using actor networks, our method has clear advantages over DDPG in terms of hyperparameter tuning and computational burden.

Figure 2 shows the action trajectories of HalfCheetah-v2, obtained by HJ DQN and DDPG. The action trajectories obtained by HJ DQN oscillate less compared to DDPG. This confirms the fact that oscillations in action are not uncommon in optimal control. In

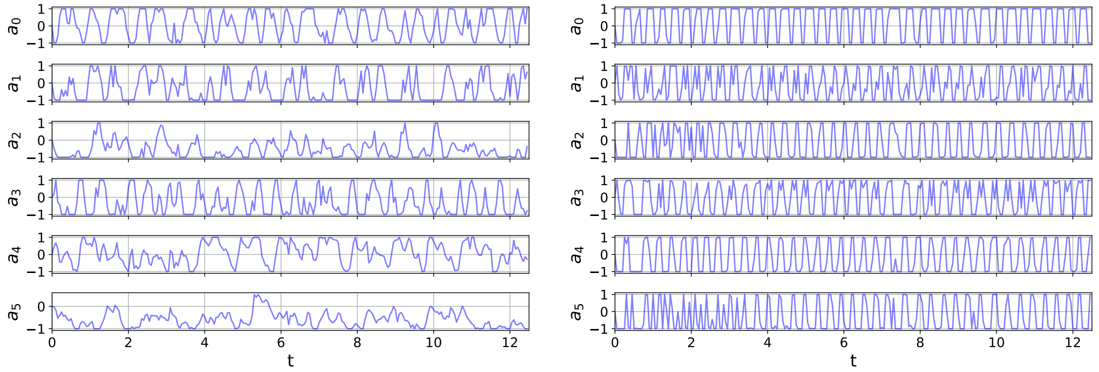
[8] `https://github.com/HJDQN/HJQ`

[9] Precisely, when the sampling interval is multiplied by a constant, the learning rate is also multiplied by the same constant.



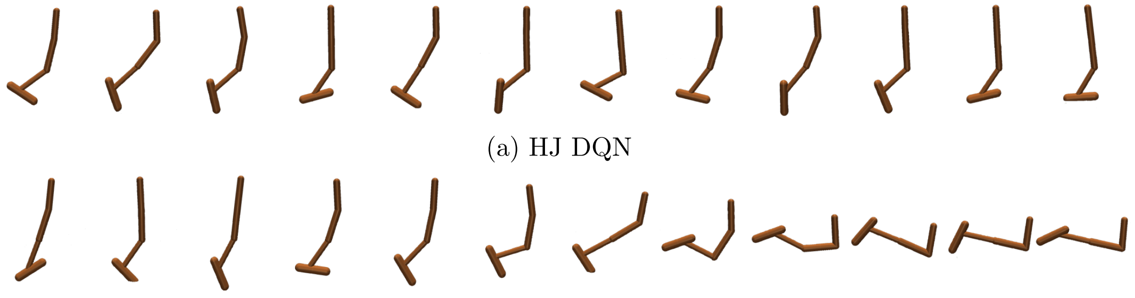
(a) Hopper-v2 (b) Swimmer-v2 (c) HalfCheetah-v2 (d) Walker2d-v2

Figure 1: Learning curves for OpenAI gym continuous control tasks.



(a) HJ DQN (b) DDPG

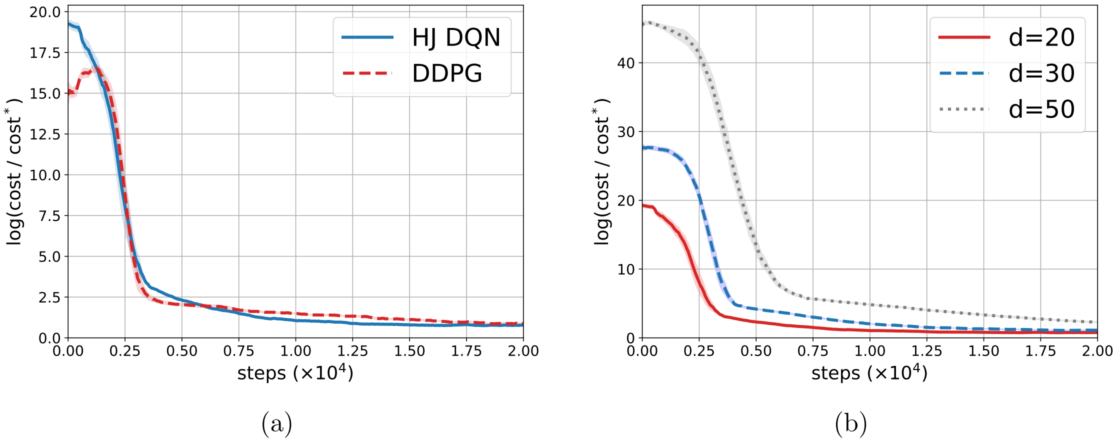
Figure 2: Action trajectories of HalfCheetah-v2, obtained by HJ DQN and DDPG.



(a) HJ DQN

(b) DDPG

Figure 3: Rendered frame sequences of length 12 for Hopper-v2, obtained by (a) HJ DQN, and (b) DDPG. The frames are selected every $20h = 0.16$ from $t = 3.6$.



(a)

(b)

Figure 4: Learning curves for the LQ problem: (a) comparing HJ DQN and DDPG; (b) the effect of problem sizes.

this particular case of HalfCheetah-v2, where DDPG outperforms HJ DQN, we suspect that fast changes in action may be needed for good performance. Oscillatory actions may be beneficial for some control tasks. On the other hand, the Lipschitz constraint in HJ DQN acts as regularizer, preventing radical changes in motion.

We estimated the success rate (that is, the proportion of the episodes where the agent proceeds until $t = 8$ without falling down) of the agent controlled by the two methods on Hopper-v2 by running 5,000 episodes of length $T = 8$. While only 12.10% of episodes were successfully solved by DDPG, the success rate of HJ DQN was 60.22% under the same conditions. Figure 3 presents sample movements of Hopper-v2 obtained by HJ DQN and DDPG. HJ DQN handled the hopper to successfully move forward, while the hopper controlled by DDPG fell in the late period of the episode. In this case, the HJ DQN agent benefited from the regularization induced by the Lipschitz constraint to generate a stable motion.

5.2 Linear-Quadratic Problems

We now consider an LQ problem with system dynamics

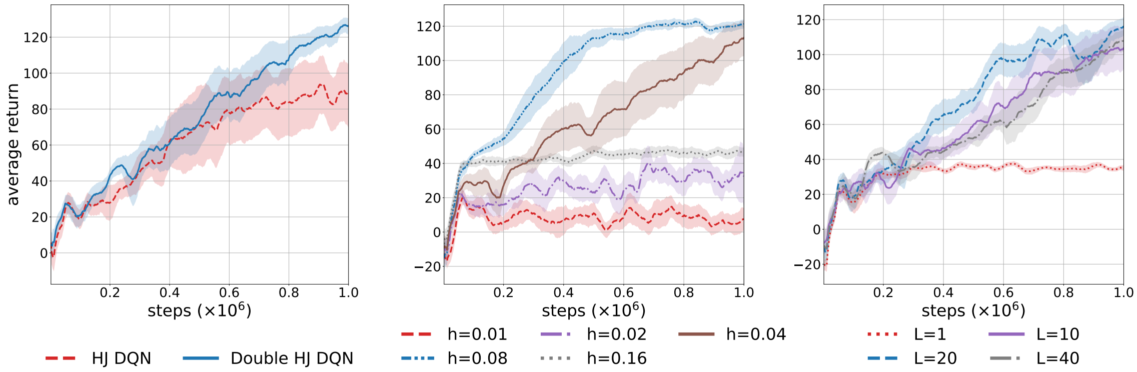
$$\dot{x}(t) = Ax(t) + Ba(t), \quad t > 0, \quad x(t), a(t) \in \mathbb{R}^d,$$

and reward function (negative cost)

$$r(\boldsymbol{x}, \boldsymbol{a}) = -(\boldsymbol{x}^\top Q\boldsymbol{x} + \boldsymbol{a}^\top R\boldsymbol{a}),$$

where $Q = Q^\top \succeq 0$ and $R = R^\top \succ 0$ (see, for example, (Anderson and Moore, 2007) for details about the theory of the classical LQ control). Note that our method solves a problem *different from the classical LQ problem* due to the Lipschitz constraint on controls. Thus, the control learned by our method must be suboptimal.

Each component of the system matrices $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times d}$ was generated uniformly from $[-0.1, 0.1]$ and $[-0.5, 0.5]$, respectively. The produced matrix A has eigenvalues



(a) Double Q-learning

(b) Sampling interval

(c) Lipschitz constraint

Figure 5: Results of the ablation study using Swimmer-v2 with respect to (a) double Q-learning, (b) sampling interval h, and (c) Lipschitz constraint on controls.

with positive real part, and thus the system is unstable. The cost weight matrices are selected as $Q = 5I$ and $R = I$. The discount factor γ and Lipschitz constant L are set to be $e^{-\gamma h} = 0.99999$ and $L = 10$. We first compared the performance of HJ DQN with DDPG for the case of $d = 20$ and reported the results in Figure 4 (a). The learning curves are plotted in the same manner as Figure 1. The y-axis of each figure represents the log of the ratio between the actual cost and the optimal cost. Therefore, the curve approaches the x-axis as the performance improves. These results imply that DDPG is unable to reduce the cost at all, whereas HJ DQN successfully learns an effective (suboptimal) policy. The result implies that HJ DQN successfully learns an effective (suboptimal) policy which is comparable to the DDPG policy, without the aid of a separate actor network. Figure 4 (b) displays the learning curves for HJ DQN with different system sizes up to 50 dimensions. Although learning speed is affected by the problem size, HJ DQN successfully solves the LQ problem with high-dimensional systems. Moreover, it is observed that the standard deviations over trials are relatively small, and the learning curves have almost no variation over trials after approximately 10^4 steps. This result indicates the stability of our method.

5.3 Design Evaluation

We make modifications to HJ DQN to understand the contribution of each component. Figure 5 presents the results for the following design evaluation experiments.

Double Q-learning. We first modify our algorithm to test whether double Q-learning contributes to the performance of our algorithm, as in DQN. Specifically, when selecting actions to update the target value, we instead use $b_j := L\frac{\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j,a_j)}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j,a_j)|}$ to remove the effects of double Q-learning. Figure 5 (a) shows that double Q-learning improves the final performance. This observation is consistent with the effect of double Q-learning in DQN. Moreover, double Q-learning reduces the variance of the average return, indicating its contribution to the stability of our algorithm.

Sampling interval. To understand the effect of sampling interval h, we run our algorithm with multiple values of h. As mentioned before, we also adjust the learning rate α

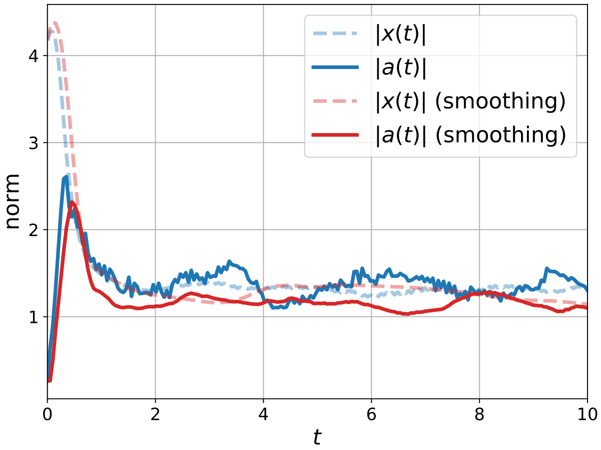


Figure 6: Effect of smoothing on the 20-dimensional LQ problem.

according to the sampling interval. As shown in Figure 5 (b), the final performance and learning speed increase as h varies from 0.01 to 0.08 and the final performance decreases as h varies from 0.08 to 0.16. When h is too small, each episode has too many sampling steps; thus, the network is trained in a small number of episodes given fixed total steps. This limits exploration, thereby diminishing the performance of our algorithm. On the other hand, as Proposition 6 implies, the target error increases with sampling interval h. This error is dominant in the case of large h. Therefore, there exists an optimal sampling interval ($h = 0.08$ in this task) that presents the best performance.

Lipschitz constraint on controls. Recall that admissible controls should satisfy the constraint $|\dot{a}(t)| \leq L$ a.e. The parameter L can be considered either a part of control problems or a design choice. We consider the latter case and display the effect of L on the learning curves in Figure 5 (c). The final reward is the lowest in the case of $L = 1$, compared to others, because the set of admissible controls is too small to allow rapid changes in control signals. With large enough L (≥ 10), HJ DQN presents a similar learning speed and performance. The final performance and learning speed slightly decrease as L varies from 20 to 40. This results from to too-large variation and frequent switching in action values, prohibiting a consistent improvement of Q-functions.

Smoothing. Finally, we present the effect of the smoothing process introduced in Section 4.2. Figure 6 shows $|x(t)|$ and $|a(t)|$ generated by the control learned with and without smoothing on the 20-dimensional LQ problem. Here, $\phi(r) = \tanh\left(\frac{r}{L}\right)$ is chosen as the smoothing function. As expected, with no smoothing process, the action trajectory shows wobbling oscillations (blue solid line). However, when the smoothing process is applied, the action trajectory has no such undesirable oscillations and presents a smooth behavior (red solid line). Regarding $|x(t)|$, the smoothing process has only a small effect. Therefore, the smoothing process can eliminate oscillations in action without significantly affecting the state trajectory.

6. Conclusions

We have presented a new theoretical and algorithmic framework that extends DQN to continuous-time deterministic optimal control for continuous action space under the Lipschitz constraint on controls. A novel class of HJB equations for Q-functions has been derived and used to construct a Q-learning method for continuous-time control. We have shown the theoretical convergence properties of this method. For practical implementation, we have combined the HJB-based method with DQN, resulting in a simple algorithm that solves continuous-time control problems without an actor network. Benefiting from our theoretical analysis of the HJB equations, this model-free off-policy algorithm does not require any numerical optimization for selecting greedy actions. The results of our experiments indicate that actor networks in DDPG may be replaced by our optimal control simply characterized via an ODE, while reducing computational effort. Our HJB framework may provide an exciting avenue for future research in continuous-time RL in terms of improving exploration capabilities with maximum entropy methods and exploiting the benefits of system models with theoretical guarantees.

Acknowledgments

This work was supported in part by the Creative-Pioneering Researchers Program through SNU, the National Research Foundation of Korea funded by the MSIT(2020R1C1C1009766), the Information and Communications Technology Planning and Evaluation (IITP) grant funded by MSIT(2020-0-00857), and Samsung Electronics (Corresponding author: Insoon Yang).

Appendix A. Viscosity Solution of the HJB Equations

The Hamilton–Jacobi equation is a partial differential equation of the form

$$F(\boldsymbol{z}, u(\boldsymbol{z}), \nabla_{\boldsymbol{z}} u(\boldsymbol{z})) = 0, \quad \boldsymbol{z} \in \mathbb{R}^k, \tag{11}$$

where $F : \mathbb{R}^k \times \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}$. A function $u : \mathbb{R}^k \to \mathbb{R}$ that solves the HJ equation is called a (strong) solution. However, such a strong solution exists only in limited cases. To consider a broad class of HJ equations, it is typical to adopt the concept of weak solutions. Among these, the *viscosity solution* is the most relevant to dynamic programming and optimal control problems (Crandall and Lions, 1983; Bardi and Capuzzo-Dolcetta, 1997). Specifically, under a technical condition, the viscosity solution is unique and corresponds to the value function of a continuous-time optimal control problem. In the following definition, $C(\mathbb{R}^k)$ and $C^1(\mathbb{R}^k)$ denote the set of continuous functions and the set of continuously differentiable functions, respectively.

Definition 1 *A function $u \in C(\mathbb{R}^k)$ is called the* viscosity solution *of* (11) *if it satisfies the following conditions:*

1. *For any $\phi \in C^1(\mathbb{R}^k)$ such that $u - \phi$ attains a local maximum at $\boldsymbol{z}_0$,*

$$F(\boldsymbol{z}_0, u(\boldsymbol{z}_0), \nabla_{\boldsymbol{z}} \phi(\boldsymbol{z}_0)) \leq 0;$$

2. *For any $\phi \in C^1(\mathbb{R}^k)$ such that $u - \phi$ attains a local minimum at $\boldsymbol{z}_0$,*

$$F(\boldsymbol{z}_0, u(\boldsymbol{z}_0), \nabla_{\boldsymbol{z}} \phi(\boldsymbol{z}_0)) \geq 0.$$

Note that the viscosity solution does not need to be differentiable. In our case, the HJB equation (5)

$$\gamma Q(\boldsymbol{x}, \boldsymbol{a}) - \nabla_{\boldsymbol{x}} Q(\boldsymbol{x}, \boldsymbol{a}) \cdot f(\boldsymbol{x}, \boldsymbol{a}) - L|\nabla_{\boldsymbol{a}} Q(\boldsymbol{x}, \boldsymbol{a})| - r(\boldsymbol{x}, \boldsymbol{a}) = 0$$

can be expressed as (11) with

$$F(\boldsymbol{z}, q, \boldsymbol{p}) = \gamma q - \boldsymbol{p}_1 \cdot f(\boldsymbol{z}) - L|\boldsymbol{p}_2| - r(\boldsymbol{z}),$$

where $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{a}) \in \mathbb{R}^n \times \mathbb{R}^m$ and $\boldsymbol{p} = (\boldsymbol{p}_1, \boldsymbol{p}_2) \in \mathbb{R}^n \times \mathbb{R}^m$. We can show that the HJB equation admits a unique viscosity solution, which coincides with the optimal Q-function.

Theorem 7 *Suppose that Assumption 1 holds.*[10] *Then, the optimal continuous-time Q-function is the unique viscosity solution to the HJB equation* (5)*.*

Proof First, recall that our control trajectory satisfies the constraint $|\dot{a}| \leq L$. Therefore, our dynamical system can be written in the following extended form:

$$\dot{x}(t) = f(x(t), a(t)), \quad \dot{a}(t) = b(t), \quad t > 0, \quad |b(t)| \leq L,$$

by viewing $x(t)$ and $a(t)$ as state variables. More precisely, the dynamics of the extended state variable $z(t) = (x(t), a(t))$ can be written as

$$\dot{z}(t) = G(z(t), b(t)), \quad t > 0, \quad |b(t)| \leq L, \tag{12}$$

where $G(\boldsymbol{z}, \boldsymbol{b}) = (f(\boldsymbol{z}), \boldsymbol{b})$. Applying the dynamic programming principle to the Q-function, we have

$$Q(\boldsymbol{z}) = \sup_{|b(s)| \leq L} \left\{ \int_t^{t+h} e^{-\gamma(s-t)} r(z(s)) \, \mathrm{d}s + e^{-\gamma h} Q(z(t+h)) \mid z(t) = \boldsymbol{z} \right\}.$$

The remaining proof is almost the same as the proof of Proposition 2.8, Chapter 3 in (Bardi and Capuzzo-Dolcetta, 1997). However, for the self-completeness of the paper, we provide a detailed proof. In the following, we show that the Q-function satisfies the two conditions in Definition 1.

First, let $\phi \in C^1(\mathbb{R}^{n+m})$ such that $Q - \phi$ attains a local maximum at $\boldsymbol{z}$. Then, there exists $\delta > 0$ such that $Q(\boldsymbol{z}) - Q(\boldsymbol{z}') \geq \phi(\boldsymbol{z}) - \phi(\boldsymbol{z}')$ for $|\boldsymbol{z}' - \boldsymbol{z}| < \delta$. Since f and r are bounded Lipschitz continuous, there exists $h_0 > 0$, which is independent of $b(s)$, such that $|z(s) - \boldsymbol{z}| \leq \delta$, $|r(z(s)) - r(\boldsymbol{z})| \leq C(s-t)$ and $|f(z(s)) - f(\boldsymbol{z})| \leq C(s-t)$ for $t \leq s \leq t + h_0$, where $z(s)$ is a solution to (12) for $s \geq t$ with $z(t) = \boldsymbol{z}$. Now, the dynamic programming

[10] Assumption 1 can be relaxed by using a modulus associated with each function as in Chapter III.1–3 in (Bardi and Capuzzo-Dolcetta, 1997).

principle for the Q-function implies that, for any $0 < h < h_0$ and $\varepsilon > 0$, there exists $b(s)$ with $|b(s)| \leq L$ such that

$$Q(\boldsymbol{z}) \leq \int_t^{t+h} e^{-\gamma(s-t)} r(z(s))\,\mathrm{d}s + e^{-\gamma h} Q(z(t+h)) + h\varepsilon,$$

where $z(s)$ is now a solution to (12) with $z(t) = \boldsymbol{z}$ under the particular choice of b. On the other hand, it follows from our choice of h that

$$\int_t^{t+h} e^{-\gamma(s-t)} r(z(s))\,\mathrm{d}s = \int_t^{t+h} e^{-\gamma(s-t)} r(\boldsymbol{z})\,\mathrm{d}s + o(h),$$

which implies that

$$Q(\boldsymbol{z}) \leq \int_t^{t+h} e^{-\gamma(s-t)} r(\boldsymbol{z})\,\mathrm{d}s + e^{-\gamma h} Q(z(t+h)) + h\varepsilon + o(h).$$

Therefore, we have

$$\begin{aligned}\phi(\boldsymbol{z}) - \phi(z(t+h)) &\leq Q(\boldsymbol{z}) - Q(z(t+h)) \\ &\leq \int_t^{t+h} e^{-\gamma(s-t)} r(\boldsymbol{z})\,\mathrm{d}s + (e^{-\gamma h} - 1) Q(z(t+h)) + h\varepsilon + o(h).\end{aligned}$$

Since the left-hand side of the inequality above is equal to $-\int_t^{t+h} \frac{\mathrm{d}}{\mathrm{d}s}\phi(z(s))\,\mathrm{d}s = -\int_t^{t+h} \nabla_{\boldsymbol{z}}\phi(z(s)) \cdot G(z(s), b(s))\,\mathrm{d}s$, we obtain that

$$\begin{aligned}0 \leq &\int_t^{t+h} \nabla_{\boldsymbol{z}}\phi(z(s)) \cdot G(z(s), b(s))\,\mathrm{d}s \\ &+ \int_t^{t+h} e^{-\gamma(s-t)} r(\boldsymbol{z})\,\mathrm{d}s + \left(e^{-\gamma h} - 1\right) Q(z(t+h)) + h\varepsilon + o(h) \\ \leq &\int_t^{t+h} (\nabla_{\boldsymbol{x}}\phi(z(s)) \cdot f(\boldsymbol{z}) + L|\nabla_{\boldsymbol{a}}\phi(z(s))|)\,\mathrm{d}s \\ &+ \int_t^{t+h} e^{-\gamma(s-t)} r(\boldsymbol{z})\,\mathrm{d}s + \left(e^{-\gamma h} - 1\right) Q(z(t+h)) + h\varepsilon + o(h).\end{aligned}$$

By dividing both sides by h and letting $h \to 0$, we conclude that

$$\nabla_{\boldsymbol{x}}\phi(\boldsymbol{z}) \cdot f(\boldsymbol{z}) + L|\nabla_{\boldsymbol{a}}\phi(\boldsymbol{z})| + r(\boldsymbol{z}) - \gamma Q(\boldsymbol{z}) + \varepsilon \geq 0.$$

Since ε was arbitrarily chosen, we confirm that the Q-function satisfies the first condition in Definition 1, that is,

$$\gamma Q(\boldsymbol{z}) - \nabla_{\boldsymbol{x}}\phi(\boldsymbol{z}) \cdot f(\boldsymbol{z}) - L|\nabla_{\boldsymbol{a}}\phi(\boldsymbol{z})| - r(\boldsymbol{z}) \leq 0.$$

We now consider the second condition. Let $\phi \in C^1(\mathbb{R}^{n+m})$ such that $Q - \phi$ attains a local minimum at $\boldsymbol{z}$, that is, there exists δ such that $Q(\boldsymbol{z}) - Q(\boldsymbol{z}') \leq \phi(\boldsymbol{z}) - \phi(\boldsymbol{z}')$ for $|\boldsymbol{z}' - \boldsymbol{z}| < \delta$. Fix an arbitrary $\boldsymbol{b} \in \mathbb{R}^m$ such that $|\boldsymbol{b}| \leq L$ and let $b(s) \equiv \boldsymbol{b}$ be a constant function. Let $z(s)$ be a solution to (12) for $s \geq t$ with $z(t) = \boldsymbol{z}$ under the particular choice of $b(s) \equiv \boldsymbol{b}$. Then, for sufficiently small h, $|z(t+h) - \boldsymbol{z}| \leq \delta$, and therefore we have

$$\begin{aligned} Q(\boldsymbol{z}) - Q(z(t+h)) \leq \phi(\boldsymbol{z}) - \phi(z(t+h)) &= -\int_t^{t+h} \frac{\mathrm{d}}{\mathrm{d}s}\phi(z(s))\,\mathrm{d}s \\ &= -\int_t^{t+h} \nabla_{\boldsymbol{z}}\phi(z(s)) \cdot G(z(s), \boldsymbol{b})\,\mathrm{d}s. \end{aligned} \tag{13}$$

On the other hand, the dynamic programming principle yields

$$Q(\boldsymbol{z}) - Q(z(t+h)) \geq \int_t^{t+h} e^{-\gamma(s-t)} r(z(s))\,\mathrm{d}s + (e^{-\gamma h} - 1)Q(z(t+h)). \tag{14}$$

By (13) and (14), we have

$$(e^{-\gamma h} - 1)Q(z(t+h)) + \int_t^{t+h} e^{-\gamma(s-t)} r(z(s))\,\mathrm{d}s \leq -\int_t^{t+h} \nabla_{\boldsymbol{z}}\phi(z(s)) \cdot G(z(s), \boldsymbol{b})\,\mathrm{d}s.$$

Dividing both sides by h and letting $h \to 0$, we obtain that

$$-\gamma Q(\boldsymbol{z}) + r(\boldsymbol{z}) \leq -\nabla_{\boldsymbol{z}}\phi(\boldsymbol{z}) \cdot (f(\boldsymbol{z}), \boldsymbol{b}),$$

or equivalently

$$\gamma Q(\boldsymbol{z}) - \nabla_{\boldsymbol{x}}\phi(\boldsymbol{z}) \cdot f(\boldsymbol{z}) - \nabla_{\boldsymbol{a}}\phi(\boldsymbol{z}) \cdot \boldsymbol{b} - r(\boldsymbol{z}) \geq 0.$$

Since $\boldsymbol{b}$ was arbitrarily chosen from $\{\boldsymbol{b} \in \mathbb{R}^m : |\boldsymbol{b}| \leq L\}$, we have

$$\gamma Q(\boldsymbol{z}) - \nabla_{\boldsymbol{x}}\phi(\boldsymbol{z}) \cdot f(\boldsymbol{z}) - L|\nabla_{\boldsymbol{a}}\phi(\boldsymbol{z})| - r(\boldsymbol{z}) \geq 0,$$

which confirms that the Q-function satisfies the second condition in Definition 1. Therefore, we conclude that the Q-function is a viscosity solution of the HJB equation (5).

Lastly, the uniqueness of the viscosity solution can be proved by using Theorem 2.12, Chapter 3 in (Bardi and Capuzzo-Dolcetta, 1997). ■

Appendix B. Proofs

B.1 Proposition 1

Proof Fix $(\boldsymbol{x}, \boldsymbol{a}) \in \mathbb{R}^n \times \mathbb{R}^m$. Let ε be an arbitrary positive constant. Then, there exists $a \in \mathcal{A}$ such that $\int_t^\infty e^{-\gamma(s-t)} r(x(s), a(s))\,\mathrm{d}s < v(\boldsymbol{x}) + \varepsilon$, where $x(s)$ satisfies (1) with $x(t) = \boldsymbol{x}$ in the Carathéodory sense: $x(s) = \boldsymbol{x} + \int_t^s f(x(\tau), a(\tau))\,\mathrm{d}\tau$. We now construct a new control $\tilde{a} \in \mathcal{A}$ as $\tilde{a}(s) := \boldsymbol{a}$ if $s = t$; $\tilde{a}(s) := a(s)$ if $s > t$. Such a modification of controls at a single point does not affect the trajectory or the total reward. Therefore, we have

$$v(\boldsymbol{x}) \leq Q(\boldsymbol{x}, \boldsymbol{a}) \leq \int_t^\infty e^{-\gamma(s-t)} r(x(s), \tilde{a}(s))\,\mathrm{d}s < v(\boldsymbol{x}) + \varepsilon.$$

Since ε was arbitrarily chosen, we conclude that $v(\boldsymbol{x}) = Q(\boldsymbol{x}, \boldsymbol{a})$ for any $(\boldsymbol{x}, \boldsymbol{a}) \in \mathbb{R}^n \times \mathbb{R}^m$. ■

B.2 Theorem 2

Proof The classical theorem for the necessary and sufficient condition of optimality (for example, Theorem 2.54, Chapter III in (Bardi and Capuzzo-Dolcetta, 1997)) implies that $a^\star$ is optimal among those in $\mathcal{A}$ such that $a(t) = \boldsymbol{a}$ if and only if

$$\begin{aligned}&p_1 \cdot f(x^\star(s), a^\star(s)) + p_2 \cdot \dot{a}^\star(s) + r(x^\star(s), a^\star(s))\\ &= \max_{|\boldsymbol{b}|\leq L}\{p_1 \cdot f(x^\star(s), a^\star(s)) + p_2 \cdot \boldsymbol{b} + r(x^\star(s), a^\star(s))\}\end{aligned}$$

for all $p = (p_1, p_2) \in D^{\pm}Q(x^\star(s), a^\star(s))$. This optimality condition can be expressed as the desired ODE (6). Thus, its solution $a^\star$ with $a^\star(t) = \boldsymbol{a}$ satisfies (7).

Suppose now that $\boldsymbol{a} \in \arg\max_{\boldsymbol{a}'\in\mathbb{R}^m} Q(\boldsymbol{x}, \boldsymbol{a}')$. It follows from the definition of Q that

$$\begin{aligned}&\max_{a\in\mathcal{A}}\left\{\int_t^\infty e^{-\gamma(s-t)} r(x(s), a(s))\,\mathrm{d}s \mid x(t) = \boldsymbol{x}\right\}\\ &= \max_{\boldsymbol{a}'\in\mathbb{R}^m}\max_{a\in\mathcal{A}}\left\{\int_t^\infty e^{-\gamma(s-t)} r(x(s), a(s))\,\mathrm{d}s \mid x(t) = \boldsymbol{x}, a(t) = \boldsymbol{a}'\right\}\\ &= \max_{\boldsymbol{a}'\in\mathbb{R}^m} Q(\boldsymbol{x}, \boldsymbol{a}') = Q(\boldsymbol{x}, \boldsymbol{a}) = \max_{a\in\mathcal{A}}\left\{\int_t^\infty e^{-\gamma(s-t)} r(x(s), a(s))\mathrm{d}s \mid x(t) = \boldsymbol{x}, a(t) = \boldsymbol{a}\right\}\\ &= \int_t^\infty e^{-\gamma(s-t)} r(x^\star(s), a^\star(s))\,\mathrm{d}s.\end{aligned}$$

Therefore, $a^\star$ is an optimal control. ■

B.3 Proposition 3

Proof We first show that $Q^{h,\star}$ satisfies (9). Fix an arbitrary sequence $b := \{b_n\}_{n=0}^\infty \in \mathcal{B}$. It follows from the definition of $Q^{h,b}$ that

$$Q^{h,b}(\boldsymbol{x}, \boldsymbol{a}) = hr(\boldsymbol{x}, \boldsymbol{a}) + (1 - \gamma h)Q^{h,\tilde{b}}(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + hb_0).$$

where $\tilde{b} := \{b_1, b_2, \ldots\} \in \mathcal{B}$. Since $Q^{h,\tilde{b}}(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + hb_0) \leq Q^{h,\star}(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + hb_0)$, we have

$$\begin{aligned}Q^{h,b}(\boldsymbol{x}, \boldsymbol{a}) &\leq hr(\boldsymbol{x}, \boldsymbol{a}) + (1 - \gamma h)Q^{h,\star}(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + hb_0)\\ &\leq hr(\boldsymbol{x}, \boldsymbol{a}) + (1 - \gamma h)\sup_{|\boldsymbol{b}|\leq L}\left\{Q^{h,\star}(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + h\boldsymbol{b})\right\}.\end{aligned}$$

Taking supremum of both sides with respect to $b \in \mathcal{B}$ yields

$$Q^{h,\star}(\boldsymbol{x}, \boldsymbol{a}) \leq hr(\boldsymbol{x}, \boldsymbol{a}) + (1 - \gamma h)\sup_{|\boldsymbol{b}|\leq L}\left\{Q^{h,\star}(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + h\boldsymbol{b})\right\}. \tag{15}$$

To obtain the other direction of inequality, we fix an arbitrary $\boldsymbol{b} \in \mathbb{R}^m$ such that $|\boldsymbol{b}| \leq L$. Let $\boldsymbol{x}' := \xi(\boldsymbol{x}, \boldsymbol{a}; h)$ and $\boldsymbol{a}' := \boldsymbol{a} + h\boldsymbol{b}$. Fix an arbitrary $\varepsilon > 0$ and choose a sequence $c := \{c_n\}_{n=0}^\infty \in \mathcal{B}$ such that

$$Q^{h,\star}(\boldsymbol{x}', \boldsymbol{a}') \leq Q^{h,c}(\boldsymbol{x}', \boldsymbol{a}') + \varepsilon.$$

We now construct a new sequence $\tilde{c} := \{\boldsymbol{b}, c_0, c_1, \ldots\} \in \mathcal{B}$. Then,

$$Q^{h,\tilde{c}}(\boldsymbol{x}, \boldsymbol{a}) = hr(\boldsymbol{x}, \boldsymbol{a}) + (1 - \gamma h) Q^{h,c}(\boldsymbol{x}', \boldsymbol{a}') \geq hr(\boldsymbol{x}, \boldsymbol{a}) + (1 - \gamma h)(Q^{h,\star}(\boldsymbol{x}', \boldsymbol{a}') - \varepsilon),$$

which implies that

$$Q^{h,\star}(\boldsymbol{x}, \boldsymbol{a}) \geq Q^{h,\tilde{c}}(\boldsymbol{x}, \boldsymbol{a}) \geq hr(\boldsymbol{x}, \boldsymbol{a}) + (1 - \gamma h)(Q^{h,\star}(\boldsymbol{x}', \boldsymbol{a}') - \varepsilon).$$

Taking the supremum of both sides with respect to $\boldsymbol{b} \in \mathbb{R}^m$ such that $|\boldsymbol{b}| \leq L$ yields

$$Q^{h,\star}(\boldsymbol{x}, \boldsymbol{a}) \geq hr(\boldsymbol{x}, \boldsymbol{a}) - (1 - \gamma h)\varepsilon + (1 - \gamma h) \sup_{|\boldsymbol{b}| \leq L} \left\{ Q^{h,\star}(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + h\boldsymbol{b}) \right\}.$$

Since ε was arbitrarily chosen, we finally obtain that

$$Q^{h,\star}(\boldsymbol{x}, \boldsymbol{a}) \geq hr(\boldsymbol{x}, \boldsymbol{a}) + (1 - \gamma h) \sup_{|\boldsymbol{b}| \leq L} \left\{ Q^{h,\star}(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + h\boldsymbol{b}) \right\}. \tag{16}$$

Combining two estimates (15) and (16), we conclude that $Q^{h,\star}$ satisfies the semi-discrete HJB equation (9). Since the proof for the uniqueness of the solution is almost the same as the proof of Theorem 4.2, Chapter VI in (Bardi and Capuzzo-Dolcetta, 1997), we have omitted the detailed proof. ■

B.4 Proposition 4

Proof For the completeness of the paper, we provide a sketch of the proof although it is similar to the proof of Theorem 1.1, Chapter VI in (Bardi and Capuzzo-Dolcetta, 1997). We begin by defining two functions $\underline{Q}^\star$ and $\overline{Q}^\star$ as

$$\underline{Q}^\star(\boldsymbol{x}, \boldsymbol{a}) := \liminf_{(\boldsymbol{x}', \boldsymbol{a}', h) \to (\boldsymbol{x}, \boldsymbol{a}, 0+)} Q^{h,\star}(\boldsymbol{x}', \boldsymbol{a}'),$$

$$\overline{Q}^\star(\boldsymbol{x}, \boldsymbol{a}) := \limsup_{(\boldsymbol{x}', \boldsymbol{a}', h) \to (\boldsymbol{x}, \boldsymbol{a}, 0+)} Q^{h,\star}(\boldsymbol{x}', \boldsymbol{a}').$$

According to the proof of Theorem 1.1, Chapter VI in (Bardi and Capuzzo-Dolcetta, 1997), it suffices to show that $\overline{Q}^\star$ satisfies the first condition of Definition 1 and $\underline{Q}^\star$ satisfies the second condition of Definition 1. To this end, for any $\phi \in C^1$, let $(\boldsymbol{x}_0, \boldsymbol{a}_0)$ be a strict local maximum point of $\overline{Q}^\star - \phi$ and choose a small enough neighborhood $\mathcal{N}$ of $(\boldsymbol{x}_0, \boldsymbol{a}_0)$ such that $(\overline{Q}^\star - \phi)(\boldsymbol{x}_0, \boldsymbol{a}_0) = \max_{\mathcal{N}}(\overline{Q}^\star - \phi)$. Then, there exists a sequence $\{(\boldsymbol{x}_n, \boldsymbol{a}_n, h_n)\}$ with $(\boldsymbol{x}_n, \boldsymbol{a}_n) \to (\boldsymbol{x}_0, \boldsymbol{a}_0)$ and $h_n \to 0+$ such that

$$(Q^{h_n,\star} - \phi)(\boldsymbol{x}_n, \boldsymbol{a}_n) = \max_{\mathcal{N}}(Q^{h_n,\star} - \phi)$$

and

$$Q^{h_n,\star}(\boldsymbol{x}_n, \boldsymbol{a}_n) \to \overline{Q}^\star(\boldsymbol{x}_0, \boldsymbol{a}_0).$$

Recall that $Q^{h,\star}$ satisfies (9). Thus, there exists $\boldsymbol{b}_n$ with $|\boldsymbol{b}_n| \leq L$ such that

$$Q^{h_n,\star}(\boldsymbol{x}_n, \boldsymbol{a}_n) - h_n r(\boldsymbol{x}_n, \boldsymbol{a}_n) - (1 - \gamma h_n) Q^{h_n,\star}(\xi(\boldsymbol{x}_n, \boldsymbol{a}_n; h_n), \boldsymbol{a}_n + h\boldsymbol{b}_n) = 0.$$

Since $Q^{h_n,\star} - \phi$ attains a local maximum at $(\boldsymbol{x}_n, \boldsymbol{a}_n)$, we have

$$(1-\gamma h_n)(\phi(\boldsymbol{x}_n,\boldsymbol{a}_n) - \phi(\xi(\boldsymbol{x}_n,\boldsymbol{a}_n;h_n),\boldsymbol{a}_n + h\boldsymbol{b}_n) + \gamma h_n Q^{h_n,\star}(\boldsymbol{x}_n,\boldsymbol{a}_n) - h_n r(\boldsymbol{x}_n,\boldsymbol{a}_n) \leq 0 \tag{17}$$

for small enough $h_n > 0$. Since $|\boldsymbol{b}_n| \leq L$ for all $n \geq 0$, there exists a subsequence n_k and $\boldsymbol{b}$ with $|\boldsymbol{b}| \leq L$ such that $\boldsymbol{b}_{n_k} \to \boldsymbol{b}$ as $k \to \infty$. Then, we substitute n in (17) by n_k, divide both sides by h_{n_k} and let $k \to \infty$ to obtain that at $(\boldsymbol{x}_0, \boldsymbol{a}_0)$

$$-\nabla_{\boldsymbol{x}}\phi \cdot f - \nabla_{\boldsymbol{a}}\phi \cdot \boldsymbol{b} + \gamma \overline{Q}^\star - r \leq 0,$$

where we use the fact that

$$\lim_{h\to 0} \frac{\xi(\boldsymbol{x},\boldsymbol{a};h) - \boldsymbol{x}}{h} = f(\boldsymbol{x},\boldsymbol{a}).$$

This implies that the first condition of Definition 1 is satisfied. Similarly, it can be shown that $\underline{Q}^\star$ satisfies the second condition of Definition 1. ■

B.5 Theorem 5

We begin by defining an optimal Bellman operator in the semi-discrete setting, $\mathcal{T}^h : L^\infty \to L^\infty$, by

$$(\mathcal{T}^h Q)(\boldsymbol{x},\boldsymbol{a}) := hr(\boldsymbol{x},\boldsymbol{a}) + (1-\gamma h)\sup_{|\boldsymbol{b}|\leq L} Q(\xi(\boldsymbol{x},\boldsymbol{a};h), \boldsymbol{a} + h\boldsymbol{b}), \tag{18}$$

where $\xi(\boldsymbol{x},\boldsymbol{a};h)$ denotes the solution of the ODE (1) at time $t = h$ with initial state $x(0) = \boldsymbol{x}$ and constant action $a(t) \equiv \boldsymbol{a}$ for $t \in [0, h)$. Our first observation is that the Bellman operator is a monotone $(1-\gamma h)$-contraction mapping for a sufficiently small h.

Lemma 8 *Suppose that $0 < h < \frac{1}{\gamma}$. Then, the Bellman operator $\mathcal{T}^h$ is a monotone contraction mapping. More precisely, it satisfies the following properties:*

(i) $\mathcal{T}^h Q \leq \mathcal{T}^h Q'$ for all $Q, Q' \in L^\infty$ such that $Q \leq Q'$;

(ii) $\|\mathcal{T}^h Q - \mathcal{T}^h Q'\|_{L^\infty} \leq (1-\gamma h)\|Q - Q'\|_{L^\infty}$ for all $Q, Q' \in L^\infty$.

Proof *(i)* Since $Q(\boldsymbol{x},\boldsymbol{a}) \leq Q'(\boldsymbol{x},\boldsymbol{a})$ for all $(\boldsymbol{x},\boldsymbol{a}) \in \mathbb{R}^n \times \mathbb{R}^m$, we have

$$\sup_{|\boldsymbol{b}|\leq L} Q(\xi(\boldsymbol{x},\boldsymbol{a};h), \boldsymbol{a} + h\boldsymbol{b}) \leq \sup_{|\boldsymbol{b}|\leq L} Q'(\xi(\boldsymbol{x},\boldsymbol{a};h), \boldsymbol{a} + h\boldsymbol{b}).$$

Multiplying $(1-\gamma h)$ and then adding $hr(\boldsymbol{x},\boldsymbol{a})$ to both sides, we confirm the monotonicity of $\mathcal{T}^h$ as desired.

(ii) We first note that for any $\boldsymbol{b} \in \mathbb{R}^m$ with $|\boldsymbol{b}| \leq L$,

$$\begin{aligned}
&\big[hr(\boldsymbol{x},\boldsymbol{a}) + (1-\gamma h)Q(\xi(\boldsymbol{x},\boldsymbol{a};h), \boldsymbol{a} + h\boldsymbol{b})\big] - \big[hr(\boldsymbol{x},\boldsymbol{a}) + (1-\gamma h)Q'(\xi(\boldsymbol{x},\boldsymbol{a};h), \boldsymbol{a} + h\boldsymbol{b})\big] \\
&= (1-\gamma h)\big[Q(\xi(\boldsymbol{x},\boldsymbol{a};h), \boldsymbol{a} + h\boldsymbol{b}) - Q'(\xi(\boldsymbol{x},\boldsymbol{a};h), \boldsymbol{a} + h\boldsymbol{b})\big] \\
&\leq (1-\gamma h)\|Q - Q'\|_{L^\infty}.
\end{aligned}$$

By the definition of $\mathcal{T}^h Q'$, we have

$$\begin{aligned}
&hr(\boldsymbol{x}, \boldsymbol{a}) + (1-\gamma h) Q(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + h\boldsymbol{b}) \\
&\leq (1-\gamma h)\|Q - Q'\|_{L^\infty} + hr(\boldsymbol{x}, \boldsymbol{a}) + (1-\gamma h) Q'(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + h\boldsymbol{b}) \\
&\leq (1-\gamma h)\|Q - Q'\|_{L^\infty} + \mathcal{T}^h Q'(\boldsymbol{x}, \boldsymbol{a}).
\end{aligned}$$

Taking the supremum of both sides with respect to $\boldsymbol{b} \in \mathbb{R}^m$ such that $|\boldsymbol{b}| \leq L$, yields

$$\mathcal{T}^h Q(\boldsymbol{x}, \boldsymbol{a}) \leq (1-\gamma h)\|Q - Q'\|_{L^\infty} + \mathcal{T}^h Q'(\boldsymbol{x}, \boldsymbol{a}),$$

or equivalently

$$\mathcal{T}^h Q(\boldsymbol{x}, \boldsymbol{a}) - \mathcal{T}^h Q'(\boldsymbol{x}, \boldsymbol{a}) \leq (1-\gamma h)\|Q - Q'\|_{L^\infty}.$$

We now change the role of Q and Q' to obtain

$$|\mathcal{T}^h Q(\boldsymbol{x}, \boldsymbol{a}) - \mathcal{T}^h Q'(\boldsymbol{x}, \boldsymbol{a})| \leq (1-\gamma h)\|Q - Q'\|_{L^\infty}.$$

Therefore, the operator $\mathcal{T}^h$ is a $(1-\gamma h)$-contraction with respect to $\|\cdot\|_{L^\infty}$. ■

Using the Bellman operator $\mathcal{T}^h$, HJ Q-learning (10) can be expressed as

$$Q_{k+1}^h := (1-\alpha_k) Q_k^h + \alpha_k \mathcal{T}^h Q_k^h.$$

Consider the difference $\Delta_k^h := Q_k^h - Q^{h,\star}$. Note that $\|\Delta_k^h\|_{L^\infty}$ represents the optimality gap at the kth iteration. It satisfies

$$\Delta_{k+1}^h = (1-\alpha_k)\Delta_k^h + \alpha_k[\mathcal{T}^h(\Delta_k^h + Q^{h,\star}) - \mathcal{T}^h Q^{h,\star}], \tag{19}$$

where we used the semi-discrete HJB equation $Q^{h,\star} = \mathcal{T}^h Q^{h,\star}$. The contraction property of the Bellman operator $\mathcal{T}^h$ can be used to show that the optimality gap $\|\Delta_k^h\|_{L^\infty}$ decreases geometrically. More precisely, we have the following lemma:

Lemma 9 *Suppose that* $0 < h < \frac{1}{\gamma}$, $0 \leq \alpha_k \leq 1$ *and that Assumption 1 holds. Then, the following inequality holds:*

$$\|\Delta_k^h\|_{L^\infty} \leq \left(\prod_{\tau=0}^{k-1}(1-\alpha_\tau\gamma h)\right)\|\Delta_0^h\|_{L^\infty}.$$

Proof We use mathematical induction to prove the assertion. When $k = 1$, it follows from the Q-function update (10) and the contraction property of $\mathcal{T}^h$ that

$$\begin{aligned}
\|\Delta_1^h\|_{L^\infty} &\leq (1-\alpha_0)\|\Delta_0^h\|_{L^\infty} + \alpha_0\|\mathcal{T}^h(\Delta_0^h + Q^{h,\star}) - \mathcal{T}^h Q^{h,\star}\|_{L^\infty} \\
&\leq (1-\alpha_0)\|\Delta_0^h\|_{L^\infty} + \alpha_0(1-\gamma h)\|\Delta_0^h\|_{L^\infty} \\
&= (1-\alpha_0\gamma h)\|\Delta_0^h\|_{L^\infty}.
\end{aligned}$$

Therefore, the assertion holds for $k = 1$. We now assume that the assertion holds for $k = n$:

$$\|\Delta_n^h\|_{L^\infty} \leq \left(\prod_{\tau=0}^{n-1}(1-\alpha_\tau\gamma h)\right)\|\Delta_0^h\|_{L^\infty}.$$

We need to show that the inequality holds for $k = n+1$. By using the same estimate as in the case of $k = 1$ and the induction hypothesis for $k = n$, we obtain

$$\begin{aligned}\|\Delta_{n+1}^h\|_{L^\infty} &\leq (1-\alpha_n)\|\Delta_n^h\|_{L^\infty} + \alpha_n \|\mathcal{T}^h(\Delta_n^h + Q^{h,\star}) - \mathcal{T}^h Q^{h,\star}\|_{L^\infty} \\ &\leq (1-\alpha_n)\|\Delta_n^h\|_{L^\infty} + \alpha_n(1-\gamma h)\|\Delta_n^h\|_{L^\infty} \\ &= (1-\alpha_n\gamma h)\|\Delta_n^h\|_{L^\infty} \\ &\leq (1-\alpha_n\gamma h)\left(\prod_{\tau=0}^{n-1}(1-\alpha_\tau\gamma h)\right)\|\Delta_0^h\|_{L^\infty} \\ &= \left(\prod_{\tau=0}^{n}(1-\alpha_\tau\gamma h)\right)\|\Delta_0^h\|_{L^\infty}.\end{aligned}$$

This completes our mathematical induction, and thus the result follows. ■

This lemma yields a condition on the sequence of learning rates under which the Q-function updated by (10) converges to the optimal semi-discrete Q-function (8) in L^∞.
Proof of Theorem 5 It suffices to show that

$$\lim_{k\to\infty}\|\Delta_k^h\|_{L^\infty} = 0.$$

By Lemma 9 and the elementary inequality $1 - x \leq e^{-x}$, we have

$$\|\Delta_k^h\|_{L^\infty} \leq \left(\prod_{\tau=0}^{k-1}(1-\alpha_\tau\gamma h)\right)\|\Delta_0^h\|_{L^\infty} \leq \exp\left(-\gamma h\left(\sum_{\tau=0}^{k-1}\alpha_\tau\right)\right)\|\Delta_0^h\|_{L^\infty}.$$

Therefore, if $\sum_{\tau=0}^{\infty}\alpha_\tau = \infty$, the result follows. ■

B.6 Corollary 1

Proof We first observe that there exists an index k_h, depending on h, such that $\sum_{\tau=0}^{k_h-1}\alpha_\tau > \frac{1}{h^2}$ since $\sum_{\tau=0}^{\infty}\alpha_\tau = \infty$. Then, we have

$$h\left(\sum_{\tau=0}^{k_h-1}\alpha_\tau\right) > \frac{1}{h} \to \infty \quad \text{as} \quad h \to 0.$$

Moreover, by the triangle inequality, we have

$$|Q_k^h(\boldsymbol{x},\boldsymbol{a}) - Q(\boldsymbol{x},\boldsymbol{a})| \leq |Q_k^h(\boldsymbol{x},\boldsymbol{a}) - Q^{h,\star}(\boldsymbol{x},\boldsymbol{a})| + |Q^{h,\star}(\boldsymbol{x},\boldsymbol{a}) - Q(\boldsymbol{x},\boldsymbol{a})|$$

for all $(\boldsymbol{x},\boldsymbol{a}) \in \mathbb{R}^n \times \mathbb{R}^m$. By Proposition 2, the second term on the right-hand side uniformly vanishes over any compact subset K of $\mathbb{R}^n \times \mathbb{R}^m$ as $h \to 0$. The first term is nothing but $|\Delta_k^h(\boldsymbol{x},\boldsymbol{a})|$, which is bounded as follows (by Lemma 9):

$$|\Delta_k^h(\boldsymbol{x},\boldsymbol{a})| \leq \left(\prod_{\tau=0}^{k-1}(1-\alpha_\tau\gamma h)\right)\|\Delta_0^h\|_{L^\infty} \leq \exp\left(-\gamma h\left(\sum_{\tau=0}^{k-1}\alpha_\tau\right)\right)\|\Delta_0^h\|_{L^\infty}, \quad k \geq 1,$$

where the second inequality holds because $1 - x \leq e^{-x}$. Our choice of k_h then yields

$$\sup_{k \geq k_h} \|\Delta_k^h\|_{L^\infty} \leq \exp\left(-\gamma h \left(\sum_{\tau=0}^{k_h - 1} \alpha_\tau \right) \right) \|\Delta_0^h\|_{L^\infty} \to 0$$

as $h \to 0$. Therefore, we conclude that

$$\begin{aligned}
&\sup_{k \geq k_h} \sup_{\substack{(\boldsymbol{x},\boldsymbol{a}) \in K \\ K \text{compact}}} |Q_k^h(\boldsymbol{x},\boldsymbol{a}) - Q(\boldsymbol{x},\boldsymbol{a})| \\
&\leq \sup_{k \geq k_h} \sup_{\substack{(\boldsymbol{x},\boldsymbol{a}) \in K \\ K \text{compact}}} |Q_k^h(\boldsymbol{x},\boldsymbol{a}) - Q^{h,\star}(\boldsymbol{x},\boldsymbol{a})| + \sup_{k \geq k_h} \sup_{\substack{(\boldsymbol{x},\boldsymbol{a}) \in K \\ K \text{compact}}} |Q^{h,\star}(\boldsymbol{x},\boldsymbol{a}) - Q(\boldsymbol{x},\boldsymbol{a})| \\
&\leq \sup_{k \geq k_h} \|\Delta_k^h\|_{L^\infty} + \sup_{\substack{(\boldsymbol{x},\boldsymbol{a}) \in K \\ K \text{compact}}} |Q^{h,\star}(\boldsymbol{x},\boldsymbol{a}) - Q(\boldsymbol{x},\boldsymbol{a})| \to 0
\end{aligned}$$

as $h \to 0$. ■

B.7 Proposition 6

Proof We first notice that by the triangle inequality,

$$\begin{aligned}
&\left| \max_{|\boldsymbol{a} - a_j| \leq hL} Q_{\theta^-}(x_{j+1}, \boldsymbol{a}) - Q_{\theta^-}(x_{j+1}, a_j + hb_j) \right| \\
&\leq \left| \max_{|\boldsymbol{a} - a_j| \leq hL} Q_{\theta^-}(x_{j+1}, \boldsymbol{a}) - Q_{\theta^-}\left(x_{j+1}, a_j + hL \frac{\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)|} \right) \right| \\
&\quad + \left| Q_{\theta^-}\left(x_{j+1}, a_j + hL \frac{\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)|} \right) - Q_{\theta^-}(x_{j+1}, a_j + hb_j) \right| \\
&=: \Delta_1 + \Delta_2.
\end{aligned}$$

We first consider Δ_1. Let $\boldsymbol{a}^\star := \arg\max_{|\boldsymbol{a} - a_j| \leq hL} Q_{\theta^-}(x_{j+1}, \boldsymbol{a})$. By the Taylor expansion, we have

$$\begin{aligned}
\max_{|\boldsymbol{a} - a_j| \leq hL} Q_{\theta^-}(x_{j+1}, \boldsymbol{a}) &= Q_{\theta^-}(x_{j+1}, \boldsymbol{a}^\star) \\
&= Q_{\theta^-}(x_{j+1}, a_j) + \nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j) \cdot (\boldsymbol{a}^\star - a_j) + O(h^2).
\end{aligned}$$

Similarly, we again use the Taylor expansion to obtain that

$$Q_{\theta^-}\left(x_{j+1}, a_j + hL \frac{\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)|} \right) = Q_{\theta^-}(x_{j+1}, a_j) + hL |\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)| + O(h^2).$$

Subtracting one equality from another yields

$$\begin{aligned}
&Q_{\theta^-}(x_{j+1}, \boldsymbol{a}^\star) - Q_{\theta^-}\left(x_{j+1}, a_j + hL \frac{\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)|} \right) \\
&= \nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j) \cdot (\boldsymbol{a}^\star - a_j) - hL |\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)| + O(h^2) \leq O(h^2),
\end{aligned}$$

where the last inequality holds because $|\boldsymbol{a}^\star - a_j| \leq hL$. Since our choice of $\boldsymbol{a}^\star$ implies that the left-hand side of the inequality above is always non-negative, we conclude that $\Delta_1 = O(h^2)$.

Regarding Δ_2, we have

$$\begin{aligned}\Delta_2 &= \left| Q_{\theta^-}\left(x_{j+1}, a_j + hL\frac{\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)|}\right) - Q_{\theta^-}\left(x_{j+1}, a_j + hL\frac{\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j)}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j)|}\right)\right| \\ &\leq Lh\|\nabla_{\boldsymbol{a}} Q_{\theta^-}\|_{L^\infty}\left|\frac{\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)|} - \frac{\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j)}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j)|}\right|.\end{aligned}$$

Note that for any two non-zero vectors v, w,

$$\left|\frac{v}{|v|} - \frac{w}{|w|}\right| \leq \frac{|v-w|}{|v|} + |w|\left(\frac{1}{|v|} - \frac{1}{|w|}\right) = \frac{|v-w|}{|v|} + \frac{|w|-|v|}{|v|} \leq \frac{2|v-w|}{|v|}.$$

On the other hand, we have

$$|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j) - \nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j)| \leq \|\nabla^2_{\boldsymbol{x}\boldsymbol{a}} Q_{\theta^-}\|_{L^\infty}|x_{j+1} - x_j| = O(h).$$

Since we assume that Q_{θ^-} is twice differentiable and $|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j)| =: C > 0$, we have $|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)| > C/2$ for sufficiently small h. Therefore, we obtain that

$$\Delta_2 \leq 2Lh\|\nabla_{\boldsymbol{a}} Q_{\theta^-}\|_{L^\infty}\frac{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j) - \nabla_{\boldsymbol{a}} Q_{\theta^-}(x_j, a_j)|}{|\nabla_{\boldsymbol{a}} Q_{\theta^-}(x_{j+1}, a_j)|} = O(h^2).$$

Combining the estimates of Δ_1 and Δ_2 yields

$$\left|\max_{|\boldsymbol{a}-a_j|\leq hL} Q_{\theta^-}(x_{j+1}, \boldsymbol{a}) - Q_{\theta^-}(x_{j+1}, a_j + hb_j)\right| = O(h^2)$$

as desired. ■

Appendix C. Brief Discussion on Extension to Stochastic Systems

The Hamilton-Jacobi Q-learning can be extended to the continuous-time stochastic control setting with controlled diffusion processes. Consider the following stochastic counterpart of the system (1):

$$\mathrm{d}x_t = f(x_t, a_t)\mathrm{d}t + \sigma(x_t, a_t)\mathrm{d}W_t, \quad t > 0, \tag{20}$$

where $\sigma : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{n\times k}$ is the diffusion coefficient and W_t is the k-dimensional standard Bronwian motion. We now define the Q-function as

$$Q(\boldsymbol{x}, \boldsymbol{a}) := \sup_{a\in\mathcal{A}} \mathbb{E}\left[\int_0^\infty e^{-\gamma t} r(x_t, a_t)\mathrm{d}t \;\mid\; x_0 = \boldsymbol{x}, a_0 = \boldsymbol{a}\right].$$

Again, the dynamic programming principle implies

$$\begin{aligned}0 = \sup_{a\in\mathcal{A}} \mathbb{E}\bigg[&\frac{1}{h}\int_t^{t+h} e^{-\gamma(s-t)} r(x(s), a(s))\,\mathrm{d}s + \frac{1}{h}[Q(x(t+h), a(t+h)) - Q(\boldsymbol{x}, \boldsymbol{a})] \\ &+ \frac{e^{-\gamma h} - 1}{h} Q(x(t+h), a(t+h)) \mid x(t) = \boldsymbol{x}, a(t) = \boldsymbol{a}\bigg].\end{aligned} \tag{21}$$

Then, we use the Itô formula

$$\begin{aligned}\mathrm{d}Q(x_t, a_t) &= \nabla_{\boldsymbol{x}} Q \cdot \mathrm{d}x_t + \nabla_{\boldsymbol{a}} Q \cdot \dot{a}\mathrm{d}t + \frac{1}{2}\mathrm{d}x_t^\top \nabla_{\boldsymbol{x}}^2 Q \mathrm{d}x_t \\ &= \nabla_{\boldsymbol{x}} Q \cdot (f(x_t, a_t)\mathrm{d}t + \sigma(x_t, a_t)\mathrm{d}W_t) + \nabla_{\boldsymbol{a}} Q \cdot \dot{a}\mathrm{d}t + \frac{(\mathrm{d}W_t)^\top \sigma^\top \nabla_{\boldsymbol{x}}^2 Q \sigma \mathrm{d}W_t}{2}\end{aligned}$$

to derive the following Hamilton-Jacobi-Bellman equation for the stochastic system (20):

$$\gamma Q - \nabla_{\boldsymbol{x}} Q \cdot f(\boldsymbol{x}, \boldsymbol{a}) - L|\nabla_{\boldsymbol{a}} Q| - r(\boldsymbol{x}, \boldsymbol{a}) - \frac{\mathrm{tr}(\sigma^\top \nabla_{\boldsymbol{x}}^2 Q \sigma)}{2} = 0. \tag{22}$$

Note that, in the stochastic case, the optimal control also satisfies $\dot{a} = L\frac{\nabla_{\boldsymbol{a}} Q}{|\nabla_{\boldsymbol{a}} Q|}$ when Q is differentiable.

Since in most practical systems transition samples are collected in discrete time, we also introduce the semi-discrete version of (22). We define a stochastic semi-discrete Q-function $Q^{h,\star}$ as

$$Q^{h,\star}(\boldsymbol{x}, \boldsymbol{a}) := \sup_{b \in \mathcal{B}} \mathbb{E}\left[h \sum_{k=0}^{\infty} r(x_k, a_k)(1 - \gamma h)^k\right],$$

where $\mathcal{B} := \{b := \{b_k\}_{k=0}^{\infty} \mid b_k \in \mathbb{R}^m, |b_k| \leq L\}$, $x_{k+1} = \xi(x_k, a_k; h)$ and $a_{k+1} = a_k + hb_k$. Here, $\xi(x_k, a_k; h)$ is now a solution to the stochastic differential equation (20) at time $t = h$ with initial state $\boldsymbol{x}$ and constant control $a(t) \equiv \boldsymbol{a}$, $t \in [0, h)$. Then, similar to the deterministic semi-discrete HJB equation (9), its stochastic counterpart can be written as

$$Q^{h,\star}(\boldsymbol{x}, \boldsymbol{a}) = hr(\boldsymbol{x}, \boldsymbol{a}) + (1 - \gamma h) \sup_{|\boldsymbol{b}| \leq L} \mathbb{E}\left[Q^{h,\star}(\xi(\boldsymbol{x}, \boldsymbol{a}; h), \boldsymbol{a} + h\boldsymbol{b})\right].$$

Using Robbins-Monro stochastic approximation (Robbins and Monro, 1951; Kushner and Yin, 2003), we obtain the following model-free update rule: in the kth iteration, we collect data (x_k, a_k, r_k, x_{k+1}) and update the Q-function by

$$Q_{k+1}^h(x_k, a_k) := (1 - \alpha_k) Q_k^h(x_k, a_k) + \alpha_k\left[hr_k + (1 - \gamma h) \sup_{|\boldsymbol{b}| \leq L} Q_k^h(x_{k+1}, a_k + h\boldsymbol{b})\right], \tag{23}$$

where x_{k+1} is obtained by simulating the stochastic system from x_k with action a_k fixed for h period, that is, $x_{k+1} = \xi(x_k, a_k; h)$. The corresponding HJ DQN algorithm for stochastic systems is essentially the same as Algorithm 1 although the transition samples are now collected through the stochastic system.

Appendix D. Implementation Details

All the simulations in Section 5 were conducted using Python 3.7.4 on a PC with Intel Core i9-9900X @ 3.50GHz, NVIDIA GeForce RTX 2080 Ti, and 64GB RAM.

Table 1 lists the hyperparameters used in our implementation of HJ DQN for each MuJoCo task and the LQ problem. Note that in this set of experiments we view the Lipschitz constant L as a hyperparameter. When the task has a compact action space of diameter D, the Lipschitz constant L is initially chosen around D/h to cover the entire action

Table 1: Hyperparameters for HJ DQN.

Hyperparameter	HalfCheetah-v2	Hopper-v2	Walker2d-v2	Swimmer-v2	LQ
optimizer	Adam (Kingma and Ba, 2015)			Adam (Kingma and Ba, 2015)	
learning rate	5×10^{-4}	10^{-4}	10^{-4}	5×10^{-4}	10^{-3}
Lipschitz constant (L)	30	30	30	15	10
default sampling interval (h)	0.05	0.008	0.008	0.04	0.05
tuned sampling interval (h)	0.01	0.016	0.032	0.08	-
(Continuous) discount (γ)	$-\log(0.99)/h$, where h is the sampling interval			$-\log(0.99)/h$	$-\log(0.99999)/h$
replay buffer size	10^6			10^6	2×10^4
target smoothing coefficient (α)	0.001			0.001	
Noise coefficient (σ)	0.1			0.1	
# of hidden layers	2 (fully connected)			2 (fully connected)	
# of hidden units per layer	256			256	
# of samples per minibatch	128			128	512
nonlinearity	ReLU			ReLU	

space. Additional tuning can then be performed to reduce its value until an appropriate scale of L is identified. While taking $L = D/h$ produces a reasonable learning result, the additional tuning process can further improve the result in practice since control trajectories generated with an overwhelmingly large L exhibit large fluctuations, as observed in Section 5.3. For DDPG, we list our chosen hyperparameters in Table 2, which have been taken from (Lillicrap et al., 2015) for MuJoCo tasks, except the network architecture which was used in OpenAI's implementation of DDPG.[11] The discount factor in the discrete-time algorithms is chosen as $\gamma' = 0.99$ for MuJoCo tasks and 0.99999 for the LQ problem so that it is equivalent to $e^{-\gamma h} \approx (1 - \gamma h)$ in our algorithm for continuous-time systems.

[11] https://spinningup.openai.com/en/latest/spinningup/bench.html

Table 2: Hyperparameters for DDPG.

Hyperparameter	MuJoCo tasks	LQ
optimizer	Adam (Kingma and Ba, 2015)	
actor learning rate	10^{-4}	
critic learning rate	10^{-3}	
(Discrete) discount (γ')	0.99	0.99999
replay buffer size	10^{6}	2×10^{4}
target smoothing coefficient (α)	0.001	
# of hidden layers	2 (fully connected)	
# of hidden units per layer	256	
# of samples per minibatch	128	512
nonlinearity	ReLU	

References

M. Abu-Khalaf and F. L. Lewis. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*, 41:779–791, 2005.

B. D. Anderson and John B. Moore. *Optimal Control: Linear Quadratic Methods*. Courier Corporation, 2007.

L. C. Baird. Reinforcement learning in continuous time: advantage updating. In *IEEE International Conference on Neural Networks*, pages 2448–2453, 1994.

M. Bardi and I. Capuzzo-Dolcetta. *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*. Birkhäuser, Boston, MA, 1997.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon. A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems. *Automatica*, 49:82–92, 2013.

T. Bian and Z.-P. Jiang. Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design. *Automatica*, 71:348–360, 2016.

S. J. Bradtke and M. O. Duff. Reinforcement learning methods for continuous-time Markov decision problems. In *Advances in Neural Information Processing Systems*, pages 393–400, 1995.

G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. OpenAI gym. *arXiv preprint arXiv:1606.01540*, 2016.

M. Crandall and P.-L. Lions. Viscosity solutions of Hamilton–Jacobi equations. *Transactions of the American Mathematical Society*, 277:1–42, 1983.

P. Dayan and S. P. Singh. Improving policies without measuring merits. In *Advances in Neural Information Processing Systems*, pages 1059–1065, 1996.

K. Doya. Reinforcement learning in continuous time and space. *Neural Computation*, 12: 219–245, 2000.

Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.

M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

S. Fujimoto, H. Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pages 1587–1596, 2018.

G. J. Gordon. Stable function approximation in dynamic programming. In *International Conference on Machine Learning*, pages 261–268, 1995.

S. Gu, T. Lillicrap, I. Sutskever, and S. Levine. Continuous deep Q-learning with model-based acceleration. In *International Conference on Machine Learning*, pages 2829–2838, 2016.

T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870, 2018.

Y. Jiang and Z.-P. Jiang. Global adaptive dynamic programming for continuous-time nonlinear systems. *IEEE Transactions on Automatic Control*, 60:2917–2929, 2015.

J. Kim and I. Yang. Hamilton–Jacobi–Bellman equations for maximum entropy optimal control. *arXiv preprint arXiv:2009.13097*, 2020a.

J. Kim and I. Yang. Hamilton–Jacobi–Bellman equations for Q-learning in continuous time. In *Learning for Dynamics and Control (L4DC)*, pages 739–748, 2020b.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representation*, 2015.

N. Kohl and P. Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *IEEE International Conference on Robotics and Automation*, pages 2619–2624, 2004.

G. P. Kontoudis and K. G. Vamvoudakis. Kinodynamic motion planning with continuous-time Q-learning: An online, model-free, and safe navigation framework. *IEEE Transactions on Neural Networks and Learning Systems*, 30:3803–3817, 2019.

H. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer Science & Business Media, New York, 2003.

S. Levine and V. Koltun. Guided policy search. In *International Conference on Machine Learning*, pages 1–9, 2013.

T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

L. Ljung. *System Identification: Theory for the User*. Pearson, 2nd edition, 1998.

M. Lutter, B. Belousov, K. Listmann, D. Clever, and J. Peters. HJB optimal feedback control with deep differential value functions and action constraints. In *Conference on Robot Learning*, pages 640–650, 2020.

P. Mehta and S. Meyn. Q-learning and pontryagin's minimum principle. In *IEEE Conference on Decision and Control*, pages 3598–3605, 2009.

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

V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.

H. Modares and F. L. Lewis. Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning. *Automatica*, 50:1780–1792, 2014.

R. Munos. A study of reinforcement learning in the continuous case by the means of viscosity solutions. *Machine Learning*, 40:265–299, 2000.

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

R. Munos and A. W. Moore. Barycentric interpolators for continuous space and time reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1024–1030, 1999.

M. Ohnishi, M. Yukawa, M. Johansson, and M. Sugiyama. Continuous-time value function approximation in reproducing kernel Hilbert spaces. In *Advances in Neural Information Processing Systems*, pages 2813–2824, 2018.

M. Palanisamy, H. Modares, F. L. Lewis, and M. Aurangzeb. Continuous-time Q-learning for infinite-horizon discounted cost linear quadratic regulator problems. *IEEE Transactions on Cybernetics*, 45:165–176, 2015.

D. Quillen, E. Jang, O. Nachum, C. Finn, J. Ibarz, and S. Levine. Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods. In *IEEE International Conference on Robotics and Automation*, pages 6284–6291, 2018.

K. Rajagopal, S. N. Balakrishnan, and J. R. Busemeyer. Neural network-based solutions for stochastic optimal control using path integrals. *IEEE Transactions on Neural Networks and Learning Systems*, 28:534–545, 2017.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

M. Ryu, Y. Chow, R. Anderson, C. Tjandraatmadja, and C. Boutilier. CAQL: Continuous action Q-learning. *arXiv preprint arXiv:1909.12397*, 2020.

J. Schulman, S. Levine, P. Moritz, M. Jordan, and P. Abbeel. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, pages 387–395, 2014.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.

C. Szepesvari. *Algorithms for Reinforcement Learning.* Morgan and Claypool Publishers, San Rafael, CA, 2010.

C. Tallec, L. Blier, and Y. Ollivier. Making deep Q-learning methods robust to time discretization. In *International Conference on Machine Learning*, pages 6096–6104, 2019.

Y. Tassa and T. Erez. Least squares solutions of the HJB equation with neural network value-function approximators. *IEEE Transactions on Neural Networks*, 18:1031–1041, 2007.

Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. L. Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, and T. Lillicrap. DeepMind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

C. Tessler, G. Tennenholtz, and S. Mannor. Distributional policy optimization: An alternative approach for continuous control. In *Advances in Neural Information Processing Systems*, pages 1350–1360, 2019.

E. Theodorou, J. Buchli, and S. Schaal. A generalized path integral control approach to reinforcement learning. *Journal of Machine Learning Research*, 11:3137–3181, 2010.

E. Todorov, T. Erez, and Y. Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012.

K. G. Vamvoudakis. Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach. *Systems & Control Letters*, 100:14–20, 2017.

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

H. Van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double Q-learning. In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 2094–2100, 2016.

C. J. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

Y. Yang, D. Wunsch, and Y. Yin. Hamiltonian-driven adaptive dynamic programming for continuous nonlinear dynamical systems. *IEEE Transactions on Neural Networks and Learning Systems*, 28:1929–1940, 2017.