ONLINE BELLMAN RESIDUE MINIMIZATION VIA SADDLE POINT OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of Bellman residual minimization with nonlinear function approximation in general. Based on a nonconvex saddle point formulation of Bellman residual minimization via Fenchel duality, we propose an online first-order algorithm with two-timescale learning rates. Using tools from stochastic approximation, we establish the convergence of our problem by approximating the dynamics of the iterates using two ordinary differential equations. Moreover, as a byproduct, we establish a finite-time convergence result under the assumption that the dual problem can be solved up to some error. Finally, numerical experiments are provided to back up our theory.

1 INTRODUCTION

Reinforcement learning (RL) (Sutton & Barto, 1998) studies the problem of sequential decision making under uncertainty. In these problems, an agent aims to make optimal decisions by interacting with the environment, which is modeled as a Markov Decision Process (MDP). Thanks to the recent advancement of deep learning, reinforcement learning has demonstrated extraordinary empirical success in solving complicated decision making problems, such as the game of Go (Silver et al., 2016; 2017), navigation (Banino et al., 2018), and dialogue systems (Li et al., 2016).

However, when nonlinear function approximation such as neural networks are utilized, theoretical analysis of RL algorithms becomes intractable as it involves solving a highly nonconvex statistical optimization problems. Whereas in the tabular case or in the case with linear function approximation, using tools for convex optimization and linear regression, the statistical and computational properties of the reinforcement learning algorithms are well-understood under these settings. In consequence, although RL algorithms with nonlinear function approximation great empirical success, their theoretical understanding lags behind, which makes it difficult to design RL methods in a principled fashion.

Moreover, from a statistical perspective, with nonlinear function approximation, RL methods such as fitted value iteration (Munos & Szepesvári, 2008), fitted Q-iteration (Antos et al., 2008a), and Bellman residual minimization Antos et al. (2008b) can be cast as nonlinear regression problems. Using nonparametric regression tools, statistical properties of batch RL methods with nonlinear function approximation are established (Farahmand et al., 2016). However, when it comes to computational properties, due to the fundamental hardness of nonconvex optimization, theoretical understanding of the convergence of RL methods remains less explored, which is contrast to the case with linear function approximation, where the convergence of online algorithms based on temporal-difference (TD) learning are well studied.

In this work, we we make the first attempt to study an online algorithm for Bellman residual minimization, with nonlinear function approximation. In the batch form, Bellman residual minimization is formulated as a bilevel optimization, which cannot be solved with computational efficiency. To tackle this problem, we formulate the Bellman residual itself as the optimal value of another maximization problem. In this way, Bellman residual minimization becomes a saddle point problem,

where the value function is the primal variable, and the dual variable tracks the TD-error of the primal variable. By also parametrizing the dual variable using a parametrized function class, we propose a primal-dual subgradient method which is an online first-order method for Bellman residue minimization.

Furthermore, since the saddle-point problem is not convex-concave, the order between the inner maximization and outer minimization problems plays a significant role. Similar to the batch algorithm, ideally we would fix the primal variable and solve the inner maximization problem to its global optima, and then updated the primal variable. However, in the online setting, this approach is not tractable. To achieve computational efficiency, we apply the two-timescale updates to the primal and dual variables. Specifically, we update the primal and dual variables using two sets of learning rates, where the learning rate of the dual variable is much larger than that of the primal variable. Using stochastic approximation (Borkar, 2008; Kushner & Yin, 2003), two-timescale updating rules ensures that we could safely fix the primal variable when studying the convergence of the dual variable. In this case, the dual variable converges to a local maximum of the inner maximization problem. Moreover, the dynamics of the iterates are characterized by two ordinary differential equations (ODE) running at different timescales.

Our contributions are three-fold. First, we formulate the problem of Bellman residual minimization as a nonconvex saddle point problem, for which we propose an online first-order algorithm using two-timescale learning rates. Second, using stochastic approximation, we show that the online algorithm converges almost surely to the asymptotically equilibria of an ODE. Third, assuming the existence of an optimization oracle which solves the dual problem up to some error, we show that the stochastic gradient method in the primal step converges to a stationary point of the squared Bellman residual up to some fundamental error.

Related Work. The statistical properties of Bellman residual minimization is studied in Antos et al. (2008b); Maillard et al. (2010); Farahmand et al. (2016) for policy evaluation, where the problem is solved using least-squares regression under the batch setting. Moreover, in these work, Bellman residue minimization is an intermediate step of least-squares policy iteration Lagoudakis & Parr (2003). These work are not comparable to our work since our study an online algorithm for Bellman residue minimization and its convergence.

In addition, our work is related to the line of research on the online algorithms for policy evaluation with function approximation. Most existing work focus on linear function approximation. Specifically, Tsitsiklis & Van Roy (1997) study the convergence of the on-policy TD(λ) algorithm based on temporal-difference (TD) error. To handle off-policy sampling, Maei et al. (2010); Sutton et al. (2009; 2016); Yu (2015); Hallak & Mannor (2017) propose various TD-learning methods with convergence guarantees. Utilizing two-timescale stochastic approximation in Borkar (2008) and strong convexity, they establish global convergence results for the proposed methods. The finite-sample analysis of these methods are recently established in Dalal et al. (2017b;a). More related works are Liu et al. (2015); Du et al. (2017), which formulate minization of the mean-squared projected Bellman error as a saddle point problem using Fenchel duality. However, since they consider linear function approximation, the corresponding saddle point problem is convex-concave, whereas our objective is nonconvex. When it comes to nonlinear function approximation, to the best of our knowledge, the only convergent algorithm is the nonlinear-GTD algorithm proposed in Bhatnagar et al. (2009). Their algorithm depends on the Hessian of the value function, and thus might be costly in practice.

Moreover, it is worth noting that Dai et al. (2017b) apply the same saddle point formulation to soft Q-learning. However, they consider a batch algorithm with the assumption that the inner maximization can be solved to the global optima. Due to nonconvexity, this assumption could stringent.

Furthermore, Chen & Wang (2016); Wang (2017) propose primal-duality of reinforcement learning based on the Lagrangian duality of linear programming for MDP (Puterman, 2014). These work establish convergence results in the tabular case. Applying neural networks to the same duality

formulation, Dai et al. (2017a); Cho & Wang (2017) propose variants of the actor-critic algorithms (Konda & Tsitsiklis, 2000), which does not have convergence guarantees.

Notation. We use the following notations throughout this paper. For any vector $x \in \mathbb{R}^n$, we use $\|x\|_2$ and $\|x\|_\infty$ to denote the Euclidean norm and the ℓ_∞-norm of x, respectively. For a finite set $\mathcal{M}$, we use $|\mathcal{M}|$ to denote its cardinality. We denote by $\mathcal{P}(\mathcal{M})$ the set of all probability measures on $\mathcal{M}$ and we write $\mathcal{B}(\mathcal{M})$ for the set of all bounded functions defined on $\mathcal{M}$. For a function $f \in \mathcal{B}(\mathcal{M})$, we define the ℓ_∞-norm of f as $\|f\|_\infty = \sup_{x\in\mathcal{M}} |f(x)|$. Moreover, for any probability measure $\mu \in \mathcal{P}(\mathcal{M})$, we write $\|f\|_\mu$ for the ℓ_2-norm with respect to μ, i.e., $\|f\|_\mu = [\int_{\mathcal{M}} |f(x)|^2 \mathrm{d}\mu(x)]^{1/2}$. Finally, we use $[n]$ to denote the set of integers $\{1, \cdots, n\}$.

2 Value Function Estimation in RL

In this section, we introduce some background on reinforcement learning that will be used in the presentation our main results.

In reinforcement learning, the environment is often modeled as a Markov decision process (MDP) denoted by a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of all possible actions, $P\colon \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the Markov transition kernel, $r\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in (0, 1)$ is the discount factor. More specifically, an agent interacts with the MDP sequentially in the following way. At the t-th step for any $t \geq 0$, suppose the MDP is at state $s_t \in \mathcal{S}$ and the agent selects an action $a_t \in \mathcal{A}$; then, the agent observes reward $r(s_t, a_t)$ and the MDP evolves to the next state $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$. Here $P(\cdot \,|\, s, a)$ is the probability distribution of the next state when taking action a at state s. The discounted cumulative reward is defined as $R = \sum_{t\geq 0} \gamma^t \cdot r(s_t, a_t)$.

In addition, a policy $\pi\colon \mathcal{S} \to \mathcal{P}(\mathcal{A})$ specifies a rule of taking actions. Specifically, $\pi(a|s)$ is the probability of selecting action a at state s under policy π. We note that π induces a Markov chain on $\mathcal{S} \times \mathcal{A}$ with transition probability $p(s', a' \,|\, s, a) = \pi(a' \,|\, s') \cdot P(s' \,|\, s, a)$. We define the (action) value function of policy π as $Q^\pi(s, a) = \mathbb{E}(R \,|\, s_0 = s, a_0 = a, \pi)$, which is the expected value of the discounted cumulative reward when the agent takes action a at state s, and follows policy π afterwards. Moreover, we define the optimal value function $Q^*\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ by letting $Q^*(s, a) = \sup_\pi Q^\pi(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where the supremum is taken over all possible policies. By definition, $Q^*(s, a)$ is the largest reward obtainable by the agent when starting from (s, a). It is well-known that Q^π and Q^* are the unique fixed points of the Bellman evaluation operator $\mathcal{T}^\pi$ and the Bellman optimality operator $\mathcal{T}^*$, respectively. Specifically, Bellman operators $\mathcal{T}^\pi\colon \mathcal{B}(\mathcal{S}\times\mathcal{A}) \to \mathcal{B}(\mathcal{S} \times \mathcal{A})$ and $\mathcal{T}^*\colon \mathcal{B}(\mathcal{S} \times \mathcal{A}) \to \mathcal{B}(\mathcal{S} \times \mathcal{A})$ are defined respectively by

$$(\mathcal{T}^\pi Q)(s, a) = r(s, a) + \gamma \cdot \mathbb{E}\big[Q(s_{t+1}, a_{t+1}) \,\big|\, s_t = s, a_t = a, a_{t+1} \sim \pi(\cdot \,|\, s_{t+1})\big], \tag{2.1}$$

$$(\mathcal{T}^* Q)(s, a) = r(s, a) + \gamma \cdot \mathbb{E}\big[\max_{a\in\mathcal{A}} Q(s_{t+1}, a) \,\big|\, s_t = s, a_t = a\big], \tag{2.2}$$

where $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$. The problems of estimating Q^π and Q^* are usually referred to as policy evaluation and optimal control, respectively. Both of these problems lie at the core of reinforcement learning. Specifically, policy evaluation is the pivotal step of dynamic programming methods. Moreover, estimating Q^* by applying the Bellman optimality operator in (2.2) repeatedly gives the classical Q-learning algorithm (Watkins & Dayan, 1992), based on which a number of new algorithms are developed.

In this work, we propose a saddle framework for stochastic fixed-point problems in general.

3 A Saddle Point Formulation of Bellman Residue Minimization

In this section, we formulate the problem of estimating the value functions introduced in §2 as saddle point optimization problems. Before going into the details, we first introduce a standard assumption on the MDP.

Assumption 3.1 (MDP Regularity)**.** The MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ satisfies the following conditions.

(i). The action space $\mathcal{A}$ is a finite set, and there exists a constant $R_{\max} > 0$ such that $|r(s,a)| \leq R_{\max}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.

(ii). For policy evaluation problem, we assume that the Markov chain on $\mathcal{S} \times \mathcal{A}$ induced by policy π has a stationary distribution $d^\pi \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$. For estimating Q^*, we consider the off-policy setting where we collect data using a behavioral policy $\pi_b \colon \mathcal{S} \to \mathcal{A}$. Moreover, we assume that π_b induces a stationary distribution d^{π_b} over $\mathcal{S} \times \mathcal{A}$. Moreover, we assume that it is possible to draw i.i.d. samples from d^π and d^{π_b}.

The first condition in Assumption 3.1 ensures that both Q^* and Q^π are bounded by $R_{\max}/(1-\gamma)$. The stationary distributions d^π and d^{π_b} in the second condition is are the natural measures to evaluate the error of estimating Q^π and Q^*. This condition holds if the Markov chains induced by π and π_b are irreducible and aperiodic. Moreover, when these two Markov chains possess the rapid mixing property, the observations are nearly independent, which justifies the assumption of i.i.d. sampling from d^π and d^{π_b}.

In the following, we first focus on estimating Q^*; the results for Q^π can be similarly obtained by replacing $\mathcal{T}^*$ by $\mathcal{T}^\pi$ in (2.1). To simplify the notation, we denote d^{π_b} by ρ. When the capacity of $\mathcal{S}$ is large, to estimate Q^* efficiently, we estimate Q^* using a parametrized function class $\mathcal{F} = \{Q_\theta \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \theta \in \mathbb{R}^d\}$, where θ is the parameter. Then the problem is reduced to finding a parameter $\theta \in \mathbb{R}^d$ such that Q_θ is close to Q^*.

Since Q^* is unknown, it is impossible to minimize the mean-squared error $\|Q_\theta - Q^*\|_\rho^2$. Since Q^* is the unique fixed point of $\mathcal{T}^*$, a direct idea is to minimize the mean-squared Bellman residual

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}}\, J(\theta) = \|Q_\theta - \mathcal{T}^* Q_\theta\|_\rho^2 = \mathbb{E}_{(s,a)\sim\rho}\big\{[Q_\theta(s,a) - (\mathcal{T}^* Q_\theta)(s,a)]^2\big\}. \tag{3.1}$$

via (stochastic) subgradient descent. By definition, the subgradient of $J(\theta)$ is

$$\nabla_\theta J(\theta) = \mathbb{E}_{(s,a)\sim\rho}\big([Q_\theta(s,a) - (\mathcal{T}^* Q_\theta)(s,a)] \cdot \big\{(\nabla_\theta Q_\theta)(s,a) - [\nabla_\theta (\mathcal{T}^* Q_\theta)](s,a)\big\}\big), \tag{3.2}$$

where $\nabla_\theta (\mathcal{T}^* Q_\theta)(s,a)$ is the subgradient of $\mathcal{T}^* Q_\theta(s,a)$ with respect to θ, which is given by

$$(\nabla_\theta \mathcal{T}^* Q_\theta)(s,a) = \gamma \cdot \mathbb{E}\big[\cdot \nabla_\theta Q_\theta(s_{t+1}, a') \,\big|\, s_t = s, a_t = a\big], \tag{3.3}$$

where $a' = \operatorname{argmax}_{b \in \mathcal{A}} Q_\theta(s_{t+1}, b)$. Although combining (3.2) and (3.3) yields the closed form of $\nabla J(\theta)$, there exists a fundamental challenge when applying gradient-based methods to (3.1). Specifically, notice that both $(\mathcal{T}^* Q_\theta)(s,a)$ and $(\nabla_\theta \mathcal{T}^* Q_\theta)(s,a)$ involves conditional expectation given $(s,a) \in \mathcal{S} \times \mathcal{A}$, and these two terms are multiplied are together in $\nabla_\theta J(\theta)$ in (3.2). Thus, to construct an unbiased estimator of $\nabla_\theta J(\theta)$, given an observation $(s,a) \sim \rho$, we need to draw two independent samples from $P(\cdot \,|\, s, a)$ to ensure that the estimators of $(\mathcal{T}^* Q_\theta)(s,a)$ and $(\nabla_\theta \mathcal{T}^* Q_\theta)(s,a)$ constructed respectively using these two samples are conditionally independent given (s,a). Such an issue is called the "double sampling" problem in reinforcement learning literature (Baird et al., 1995).

To resolve this issue, inspired by the saddle point formulation of soft Q-learning in Dai et al. (2017b), we formulate the objective function $J(\theta)$ as

$$J(\theta) = \underset{\mu \in \mathcal{B}(\mathcal{S}\times\mathcal{A})}{\text{maximize}}\, \mathbb{E}_\rho\big\{-1/2 \cdot [\mu(s,a)]^2 + [Q_\theta(s,a) - (\mathcal{T}^* Q_\theta)(s,a)] \cdot \mu(s,a)\big\}, \tag{3.4}$$

where the maximization is taken over all functions $\mu \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. In particular, for any fixed θ, the solution of the optimization problem in (3.4) is $\delta_\theta(s,a) = Q_\theta(s,a) - (T^* Q_\theta)(s,a)$, which is known as the temporal-difference (TD) error in the literature. Moreover, we parametrize μ in (3.4) using function class $\mathcal{G} = \{\mu_\omega \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \omega \in \mathbb{R}^p\}$. Then combining (3.1) and (3.4), we formulate Bellman residual minimization as a stochastic saddle point problem

$$\begin{aligned}\min_{\theta \in \mathbb{R}^d} \max_{\omega \in \mathbb{R}^p}\; L(\theta, \omega) = \mathbb{E}_{s,a,s'}\big\{&-1/2 \cdot [\mu_\omega(s,a)]^2 + \\ &\big[Q_\theta(s,a) - r(s,a) - \gamma \cdot \max_{a' \in \mathcal{A}} Q_\theta(s', a')\big] \cdot \mu_\omega(s,a)\big\},\end{aligned} \tag{3.5}$$

where $(s,a) \sim \rho$ and $s' \sim P(\cdot \,|\, s,a)$ is the next state given (s,a). Here θ and ω can be viewed as the primal and dual variables respectively. The gradients of $L(\theta,\omega)$ with respect to θ and ω are given by

$$\nabla_\theta L(\theta,\omega) = \mathbb{E}_{s,a,s'}\big\{\mu_\omega(s,a)\cdot\big[\nabla_\theta Q_\theta(s,a) - \gamma\cdot\nabla_\theta Q_\theta(s',a')\big]\big\}, \tag{3.6}$$

$$\nabla_\omega L(\theta,\omega) = \mathbb{E}_{s,a,s'}\big\{\nabla_\omega\mu_\omega(s,a)\cdot\big[Q_\theta(s,a) - R(s,a) - \gamma\cdot\max_{b\in\mathcal{A}} Q_\theta(s',b) - \mu_\omega(s,a)\big]\big\}, \tag{3.7}$$

where a' in (3.6) satisfies that $a' = \operatorname{argmax}_{b\in\mathcal{A}} Q_\theta(s',b)$. From (3.6) and (3.7), replacing $\nabla_\theta L(\theta,\omega)$ and $\nabla_\omega L(\theta,\omega)$ by the stochastic gradients based on one observation (s,a,s'), we establish an stochasitc subgradient algorithm, whose details are given in Algorithm 1. Since the saddle point problem (3.5) is nonconvex, in the algorithm we project the iterates onto compact sets $\Theta \subseteq \mathbb{R}^d$ and $\Omega \subseteq \mathbb{R}^p$ respectively. Moreover, notice that ideally we would like to solve the inner maximization problem in (3.5) for fixed θ. To achieve such a goal in an online fashion, we perform the primal and dual steps in different paces. Specifically, the learning rates α_t and β_t satisfy $\beta_t/\alpha_t \to \infty$ as t goes to infinity. Using results from stochastic approximation (Borkar, 2008; Kushner & Yin, 2003), such a condition on the learning rates ensures that the dual iterates $\{\omega_t\}_{t\geq 0}$ asymptotically track the sequence $\{\operatorname{argmax}_{\omega\in\mathbb{R}^p} L(\theta_t,\omega)\}_{t\geq 0}$, thus justifying our algorithm.

Algorithm 1 A Primal-Dual Algorithm for Q-Learning

Input: Initial parameter estimates $\theta_0 \in \mathbb{R}^d$ and $\omega_0 \in \mathbb{R}^p$, primal and dual learning rates $\{\alpha_t,\beta_t\}_{t\geq 0}$.

for $t = 0, 1, 2, \ldots$ until convergence **do**

Sample $(s_t,a_t)\sim\rho$, observe reward $r(s_t,a_t)$ and the next state s_t'.

Let $a_t' = \operatorname{argmax}_{a\in\mathcal{A}} Q_{\theta_t}(s_t',a)$.

Update the parameters by

$$\omega_{t+1} \leftarrow \Pi_\Omega\big\{\omega_t + \beta_t\cdot\nabla_\omega\mu_{\omega_t}(s_t,a_t)\cdot\big[Q_{\theta_t}(s_t,a_t) - r(s_t,a_t) - \gamma\cdot Q_{\theta_t}(s_t',a_t') - \mu_{\omega_t}(s_t,a_t)\big]\big\}, \tag{3.8}$$

$$\theta_{t+1} \leftarrow \Pi_\Theta\big\{\theta_t - \alpha_t\cdot\mu_{\omega_t}(s_t,a_t)\cdot\big[\nabla_\theta Q_{\theta_t}(s_t,a_t) - \gamma\cdot\nabla_\theta Q_{\theta_t}(s_t',a_t')\big]\big\}. \tag{3.9}$$

end for

Furthermore, for policy evaluation, we replace $\mathcal{T}^*$ by $\mathcal{T}^\pi$ in (3.4) to obtain

$$\min_{\theta\in\mathbb{R}^d}\max_{\omega\in\mathbb{R}^p} \mathbb{E}_{s,a,s',a}\big\{-1/2\cdot[\mu_\omega(s,a)]^2 + \big[Q_\theta(s,a) - r(s,a) - \gamma\cdot Q_\theta(s',a')\big]\cdot\mu_\omega(s,a)\big\}, \tag{3.10}$$

where $(s,a)\sim d^\pi$, $s'\sim P(\cdot\,|\,s,a)$, and $a'\sim\pi(\cdot\,|\,s')$. Here d^π is the stationary distribution on $\mathcal{S}\times\mathcal{A}$ induced by policy π. Similarly, by computing the gradient of the objective in (3.10), we obtain a stochastic gradient algorithm for policy evaluation.

Finally, it is worth noting that, various saddle point formulations of Bellman residual minimization are proposed to avoid the issue of double sampling (Antos et al., 2008b; Farahmand et al., 2016; Dai et al., 2017b). Our formulation is the same as the one for soft Q-learning in Dai et al. (2017b), and is equivalent to that in Antos et al. (2008b); Farahmand et al. (2016) for policy evaluation. All of these work are batch algorithms, with the assumption that the global optima of the inner maximization can be reached. Whereas we propose an online first-order algorithm with nonlinear function approximation. Moreover, our analysis utilize tools from stochastic approximation of iterative algorithms (Borkar, 2008; Kushner & Yin, 2003), which is significantly different from their theory. .

4 Theoretical Results

In this section, we lay out the theoretical results. For ease of presentation, we focus on the estimation of Q^* while our theory can be easily adapted to the policy evaluation problem. We first state the assumption on function class $\mathcal{F} = \{Q_\theta\colon \theta\in\mathbb{R}^d\}$.

Assumption 4.1. Here we assume that, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $Q_\theta(s,a)$ is a differentiable function of θ such that $|Q_\theta(s,a)| \le Q_{\max}$ $\|\nabla_\theta Q_\theta(s,a)\|_2 \le G_{\max}$, and that $\nabla_\theta Q_\theta(s,a)$ is Lipschitz continuous in θ, where $Q_{\max} \ge R_{\max}/(1-\gamma)$ and $G_{\max} > 0$ are two constants.

Here our assumption on $\mathcal{F}$ allows nonlinear function approximation of the value function in general. Moreover, we only consider bounded value functions since Q^* is bounded by $R_{\max}/(1-\gamma)$. Moreover, we assume that $\nabla_\theta Q_\theta(s,a)$ is bounded and Lipschitz continuous for regularity. This assumption can be readily satisfied if θ is restricted to a compact subset of $\mathbb{R}^d$.

In addition, we assume that $\mathcal{G} = \{\mu_\omega \colon \omega \in \mathbb{R}^p\}$ is a class of linear functions as follows.

Assumption 4.2. We assume that $\mu_\omega(s,a) = \omega^\top \phi(s,a)$ for any $\omega \in \mathbb{R}^p$, where $\phi \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^p$ is a feature mapping such that $\phi(s,a)$ is uniformly bounded for any $s \in \mathcal{S}, a \in \mathcal{A}$. Furthermore, we assume that there exists a constant $\sigma_{\min} > 0$ such that $\mathbb{E}_{(s,a)\sim\rho}[\phi(s,a)\phi(s,a)^\top] \succeq \sigma_{\min} \cdot I_p$.

Here we assume the dual function is linear for the purpose of theoretical analysis. In this case, the inner maximization problem $\max_\omega L(\theta,\omega)$ has a unique solution

$$\omega(\theta) = \big\{\mathbb{E}_{(s,a)\sim\rho}[\phi(s,a)\phi(s,a)^\top]\big\}^{-1}\mathbb{E}_{(s,a)\sim\rho}\big\{\phi(s,a)\cdot[Q_\theta(s,a) - (\mathcal{T}^*Q_\theta)(s,a)]\big\} \tag{4.1}$$

for any $\theta \in \mathbb{R}^d$. Thus, $\mu(\theta)$ is the minimizer of $\|Q_\theta - \mathcal{T}^*Q_\theta - \mu_{\omega(\theta)}\|_\rho^2$, i.e., $\mu_{\omega(\theta)}$ is the best approximation of the TD-error using function class $\mathcal{G}$. It is possible to extend our result to nonlinear μ_ω following the analysis in Heusel et al. (2017) under stronger assumptions. In addition, we note that most online TD-learning algorithms uses linear function approximation. Moreover, it is shown in Tsitsiklis & Van Roy (1997) that TD-learning with nonlinear function approximation may fail to converge. To the best of our knowledge, for nonlinear function approximation, the nonlinear GTD algorithm in Bhatnagar et al. (2009) is the only convergent online algorithm. However, their focus solely on policy evaluation, and their approach depends on the Hessian $\nabla_\theta^2 V_\theta$. As a comparison, our method also consider nonlinear function approximation and can be applied to both policy evaluation and optimal control.

Now we are ready to present the convergence result for Algorithm 1.

Theorem 4.3. We assume the learning rates $\{\alpha_t, \beta_t\}_{t\ge0}$ in (3.8) and (3.9) satisfy

$$\sum_{t\ge0}\alpha_t = \sum_{t\ge0}\beta_t = \infty, \quad \sum_{t\ge0}\alpha_t^2 + \beta_t^2 < \infty, \quad \lim_{t\to\infty}\alpha_t/\beta_t = 0. \tag{4.2}$$

In addition, let $\Theta \subseteq \mathbb{R}^d$ and $\Omega \subseteq \mathbb{R}^p$ be the Euclidean balls with radius R_θ and R_ω respectively. For function $L(\theta,\omega)$ defined in (3.5), we define

$$\mathcal{K} = \big\{\theta \in \Theta \colon \nabla_\theta L(\theta,\omega)|_{\omega=\omega(\theta)} = 0\big\} \cup \big\{\theta \in \partial\Theta \colon \nabla_\theta L(\theta,\omega)|_{\omega=\omega(\theta)} = \lambda\theta \text{ for some } \lambda \ge 0\big\}. \tag{4.3}$$

Then under Assumptions 3.1, 4.1, and 4.2, the iterates $\{(\theta_t,\omega_t)\}_{t\ge0}$ created by Algorithm 1 converges almost surely to the set $\{[\theta^*, \omega(\theta^*)], \theta^* \in \mathcal{K}\}$.

In addition to the Robbins-Monro condition, (Robbins et al., 1951), the learning rates in (4.2) also satisfies $\lim_{t\to\infty}\alpha_t/\beta_t = 0$. Intuitively, this means that the sequence $\{\omega_t\}_{t\ge1}$ tracks $\{\omega(\theta_t)\}_{t\ge1}$ asymptotically. In other words, (4.2) enables our online algorithm to approximately solve the inner maximization problem $\max_\omega L(\theta,\omega)$ with θ fixed. Using two-timescale stochastic approximation (Borkar, 2008; Kushner & Yin, 2003), when studying the convergence $\{\theta_t\}_{t\ge0}$, we could replace ω_t in (3.9) by $\omega(\theta_t)$. In this case, using ODE approximation, $\{\theta_t\}_{t\ge0}$ converges almost surely to $\mathcal{K}$ in (4.3), which is the asymptotically stable equilibria of the projected ODE $\dot\theta = \nabla_\theta L(\theta,\omega)|_{\omega=\omega(\theta)} + \xi^\theta$, where $\xi^\theta(t)$ is the correction term caused by projection onto Θ.

Furthermore, even when Θ is sufficiently large such that $\{(\omega_t,\theta_t)\}_{t\ge0}$ converges to $[\omega(\theta^*),\theta^*]$ with $\nabla_\theta L[\theta^*,\omega(\theta^*)] = 0$, due to the error of function approximation, θ^* is not a stationary point of $J(\cdot)$. Specifically, let $\delta_\theta = Q_\theta - \mathcal{T}^*Q_\theta$ be the TD-error. Then, by (3.6) we have

$$\nabla_\theta J(\theta^*) = \mathbb{E}_{(s,a)\sim\rho}[\nabla_\theta\delta_{\theta^*}(s,a)\cdot\delta_{\theta^*}(s,a)] = \mathbb{E}_{(s,a)\sim\rho}\big\{\nabla_\theta\delta_{\theta^*}(s,a)\cdot[\delta_{\theta^*}(s,a) - \mu_{\omega(\theta^*)}(s,a)]\big\}. \tag{4.4}$$

Since $\mu_{\omega(\theta^*)}$ is the best approximator of δ_{θ^*} within $\mathcal{G}$, θ^* is in a neighborhood of a stationary point of $J(\cdot)$ if the function approximation error is small.

To see error incurred in estimating the TD-error reflects the fundamental limit of our method, we also provide a finite-time analysis provided that there exists an optimization oracle which approximately solves the inner maximization problem in (3.1).

Assumption 4.4 (Optimization Oracle)**.** We assume that there exists an optimization oracle that returns a bounded function $\widetilde{\mu}_\theta \in \mathcal{B}(\mathcal{S} \times \mathcal{A})$ when queried by any $\theta \in \mathbb{R}^d$. Moreover, we assume that $|\mu_\theta(s,a)| \leq 2Q_{\max}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, and there exists a constant $\epsilon > 0$ such that $\|\widetilde{\mu}_\theta - \delta_\theta\|_\rho^2 \leq \varepsilon$.

Here we assume that the estimator $\widetilde{\mu}_\theta$ for δ_θ is bounded by $2Q_{\max}$ since δ_θ is bounded by $2R_{\max}/(1-\gamma)$ under Assumption 3.1. Moreover, in light of Algorithm 1, ε can be viewed as the estimation error of the TD-error using $\{\mu_\omega, \omega \in \Omega\}$, i.e., $\sup_{\theta\in\Theta} \inf_{\omega\in\Omega} \|\delta_\theta - \mu_\omega\|_\rho$. Now we update $\{\theta_t\}_{t\geq 0}$ by

$$\theta_{t+1} \leftarrow \theta_t - \alpha_t \cdot \widetilde{\mu}_{\theta_t}(s_t, a_t) \cdot \big[\nabla_\theta Q_{\theta_t}(s_t, a_t) - \gamma \cdot \nabla_\theta Q_{\theta_t}(s_t', a_t')\big], \tag{4.5}$$

where s_t' is the next state given $(s_t, a_t) \sim \rho$, and $a_t' = \operatorname{argmax}_{a\in\mathcal{A}} Q_{\theta_t}(s_t', a)$. The following theorem shows that any limit point of $\{\theta_t\}_{t\geq 0}$ is in the vicinity of a stationary point of $J(\cdot)$ with error proportional to ε.

Theorem 4.5. Assume that the learning rates $\{\alpha_t\}_{t\geq 0}$ in (4.5) satisfy $\sum_{t\geq 0} \alpha_t = \infty$ and $\sum_{t\geq 0} \alpha_t^2 < \infty$. Under Assumptions 3.1, 4.1, and 4.4, with probability one, we have $\limsup_{t\geq 0} \|\nabla_\theta J(\theta_t)\|_2 \leq 8G_{\max} \cdot \varepsilon$.

To understand this theorem, first notice that the update direction in (4.5) is a noisy version of $\mathbb{E}_{(s,a)\sim\rho}[\widetilde{\mu}_{\theta_t}(s,a) \cdot \nabla_\theta \delta_{\theta_t}(s,a)]$, which is a biased estimate of $\nabla_\theta J(\theta_t)$. Moreover, under Assumptions 4.1 and 4.4, such a bias can be bounded by $8G_{\max} \cdot \varepsilon$. Due to this bias in gradient estimation, $\{\theta_t\}_{t\geq 1}$ can only enter a neighborhood of a stationary point. In addition, for Algorithm 1 with $\mathcal{G}$ being the class of linear functions, the optimization oracle outputs $\mu_{\omega(\theta)} = \phi^\top \omega(\theta)$ for any $\theta \in \Theta$. Applying Cauchy-Schwarz inequality to (4.4), we have $\|\nabla J(\theta^*)\|_2 \leq 2G_{\max} \cdot \|\delta_{\theta^*} - \mu_{\omega(\theta^*)}\|_\rho$. Thus, Theorems 4.3 and 4.5 yield consistent results, which implies that the error incurred by estimating the TD-error using function class $\mathcal{G}$ leads to unavoidable error of our approach.

5 Experiments

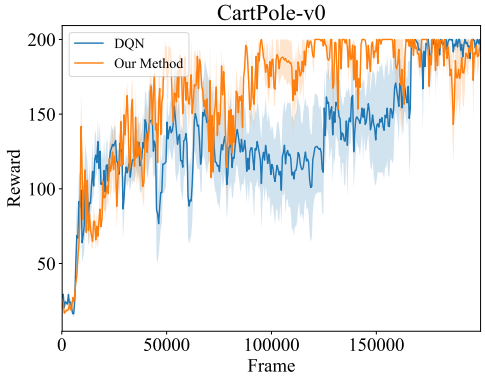
To justify our proposed method for estimating value function, we compare it with the classical deep Q-network (DQN) on several control tasks from the OpenAI Gym (Brockman et al., 2016). For a fair comparison, we use the codes from OpenAI Baselines (Dhariwal et al., 2017) for DQN, and use the same parametrization for both our method and DQN. The parameters are fine-tuned for each task to achieve the state-of-art performance. We run the algorithms with 5 random seeds and report the average rewards with 50% confidence intervals. Figure 1 (a)-(c) illustrates the empirical comparison results for the environments CartPole-v0, MountainCar-v0 and Pendulum-v0. We can see that in all these tasks, our method achieves the equivalent performance to DQN. The experiment settings are as follows.

- **CartPole-v0.** For both methods, we set the learning rate for Q-network, i.e. α_t, as 10^{-3} with batch size 32. The Q-network is a one-layer network with 30 hidden nodes. The μ-network's learning rate and structure are the same as the Q-network.
- **Pendulum-v0.** For both methods, we set α_t as 10^{-4} with batch size 1000. The Q-network is a two-layer network with 40 hidden nodes for each layer. The learning rate and structure of the μ-network are the same as Q-network.
- **MountainCar-v0.** For both methods, we set α_t as 10^{-4} with batch size 1000. The Q-network is a two-layer network with 20 hidden nodes for each layer. The learning rate for the μ-network, i.e. β_t, is also 10^{-4} with the same structure as Q-network.

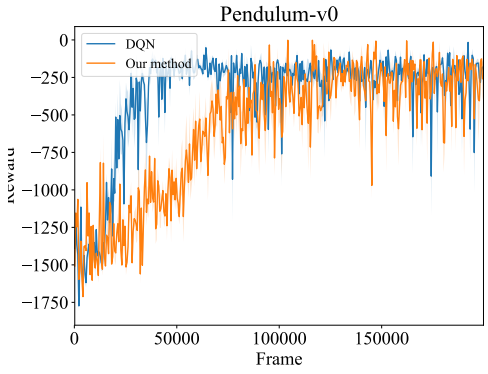
As is shown in Theorem 4.3, we require $\lim_{t\to\infty} \alpha_t/\beta_t = 0$ to guarantee the solution of inner maximization problem $\max_\omega L(\theta, \omega)$ asymptotically. To justify this theoretical result, we test our method on CartPole-v0 with different setting of α_t and β_t. The results are illustrated in Figure 1-(d). We set α_t/β_t as 100, 1, and 0.01, respectively. As the learning rate ratio α_t/β_t deceases, the reward increases faster and become more stable around the solution. From Case I, when α_t/β_t is too big, our method cannot guarantee the solution of control task, which is in accordance with Theorem 4.3.

6 Conclusion

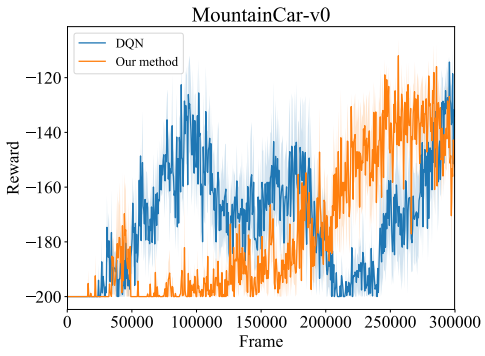
In this work, we propose an online first-order algorithm for the problem of Bellman residual minimization with nonlinear function approximation in general. Our algorithm is motivated by a duality formulation. Moreover, we establish the convergence of our problem by via ODE approximation, utilizing tools from stochastic approximation. In addition, we also establish a finite-time convergence result under the assumption of a computational oracle. Finally, numerical experiments are provided to back up our theory.



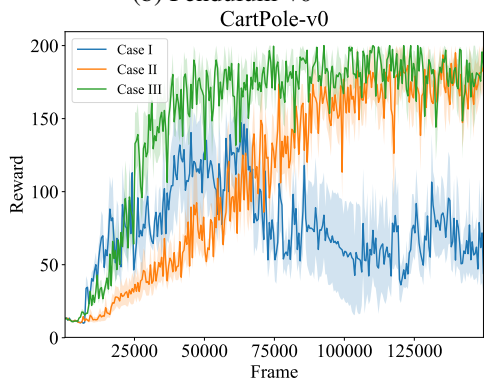
(a) CartPole-v0



(b) Pendulum-v0



(a) MountainCar-v0



(b) CartPole-v0 with various α_t/β_t

Figure 1: In (a)-(c) we compare our method and DQN on three classical control tasks. Each plot shows the average reward during training across 5 random runs, with 50% confidence interval. As shown in these plots, our method achieves the equivalent performance to the classical DQN. In addition, in (d) we show the performances of our method for CartPole-v0 under different learning rates. For Case I, we set $\alpha_t = 10^{-2}, \beta_t = 1e-4$ with $\alpha_t/\beta_t = 100$. For Case II, we set $\alpha_t = 10^{-4}, \beta_t = 10^{-4}$ with $\alpha_t/\beta_t = 1$. For Case III, we set $\alpha_t = 10^{-4}, \beta_t = 10^{-2}$ with $\alpha_t/\beta_t = 0.01$. The figure shows that $\alpha_t/\beta_t \to 0$ is critical for our method to work, which agrees with the theory.

REFERENCES

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted Q-iteration in continuous action-space MDPs. In *Advances in neural information processing systems*, pp. 9–16, 2008a.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008b.

Leemon Baird et al. Residual algorithms: Reinforcement learning with function approximation. In *International Conference on Machine Learning*, pp. 30–37, 1995.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, pp. 1, 2018.

Shalabh Bhatnagar, Doina Precup, David Silver, Richard S Sutton, Hamid R Maei, and Csaba Szepesvári. Convergent temporal-difference learning with arbitrary smooth function approximation. In *Advances in Neural Information Processing Systems*, pp. 1204–1212, 2009.

Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Yichen Chen and Mengdi Wang. Stochastic primal-dual methods and sample complexity of reinforcement learning. *arXiv preprint arXiv:1612.02516*, 2016.

Woon Sang Cho and Mengdi Wang. Deep primal-dual reinforcement learning: Accelerating actor-critic using bellman duality. *arXiv preprint arXiv:1712.02467*, 2017.

Bo Dai, Albert Shaw, Niao He, Lihong Li, and Le Song. Boosting the actor with dual critic. *arXiv preprint arXiv:1712.10282*, 2017a.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Jianshu Chen, and Le Song. Smoothed dual embedding control. *arXiv preprint arXiv:1712.10285*, 2017b.

Gal Dalal, Balazs Szorenyi, Gugan Thoppe, and Shie Mannor. Concentration bounds for two timescale stochastic approximation with applications to reinforcement learning. *arXiv preprint arXiv:1703.05376*, 2017a.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analysis for td (0) with linear function approximation. *arXiv preprint arXiv:1704.01161*, 2017b.

Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, and Yuhuai Wu. Openai baselines. `https://github.com/openai/baselines`, 2017.

Joseph L Doob. *Stochastic processes*, volume 7. Wiley New York, 1953.

Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pp. 1049–1058, 2017.

Amir M Farahmand, Mohammad Ghavamzadeh, Shie Mannor, and Csaba Szepesvári. Regularized policy iteration. In *Advances in Neural Information Processing Systems*, pp. 441–448, 2009.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361, 2017.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. *arXiv preprint arXiv:1702.07121*, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6629–6640, 2017.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pp. 1008–1014, 2000.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, New York, NY, 2003.

Harold Joseph Kushner and Dean S Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer Science & Business Media, 1978.

Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of machine learning research*, 4(Dec):1107–1149, 2003.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms. In *Conference on Uncertainty in Artificial Intelligence*, pp. 504–513. AUAI Press, 2015.

Hamid Reza Maei, Csaba Szepesvári, Shalabh Bhatnagar, and Richard S Sutton. Toward off-policy learning control with function approximation. In *nternational Conference on International Conference on Machine Learning*, pp. 719–726, 2010.

Odalric-Ambrym Maillard, Rémi Munos, Alessandro Lazaric, and Mohammad Ghavamzadeh. Finite-sample analysis of bellman residual minimization. In *Proceedings of 2nd Asian Conference on Machine Learning*, pp. 299–314, 2010.

Michel Metivier and Pierre Priouret. Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *IEEE Transactions on Information Theory*, 30(2):140–151, 1984.

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

Jacques Neveu. *Discrete-parameter martingales*. Elsevier, 1975.

HL Prasad, LA Prashanth, and Shalabh Bhatnagar. Actor-critic algorithms for learning Nash equilibria in n-player general-sum games. *arXiv preprint arXiv:1401.2086*, 2014.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Herbert Robbins, Sutton Monro, et al. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

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

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

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. Cambridge: MIT press, 1998.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *International Conference on Machine Learning*, pp. 993–1000. ACM, 2009.

Richard S Sutton, A Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17(1): 2603–2631, 2016.

John N Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In *Advances in Neural Information Processing Systems*, pp. 1075–1081, 1997.

Mengdi Wang. Primal-dual pi-learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Huizhen Yu. On convergence of emphatic temporal-difference learning. In *Conference on Learning Theory*, pp. 1724–1751, 2015.

A ALGORITHMS FOR POLICY EVALUATION AND SOFT Q-LEARNING

For policy evaluation, we replace $\mathcal{T}^*$ in (3.4) by the Bellman evaluation operator $\mathcal{T}^\pi$, which yields a saddle point problem

$$\min_{\theta\in\mathbb{R}^d}\max_{\omega\in\mathbb{R}^p} L^\pi(\theta,\omega)=\mathbb{E}_{s,a,s',a'}\big\{-1/2\cdot[\mu_\omega(s,a)]^2+\big[Q_\theta(s,a)-r(s,a)-\gamma\cdot Q_\theta(s',a')\big]\cdot\mu_\omega(s,a)\big\}, \tag{A.1}$$

where $(s,a)\sim\rho$, $s'\sim P(\cdot\,|\,s,a)$ is the next state given (s,a), and $a\sim\pi(\cdot\,|\,s')$. The gradients of $L^\pi(\theta,\omega)$ with respect to θ and ω are given by

$$\nabla_\theta L^\pi(\theta,\omega)=\mathbb{E}_{s,a,s',a'}\big\{\mu_\omega(s,a)\cdot\big[\nabla_\theta Q_\theta(s,a)-\gamma\cdot\nabla_\theta Q_\theta(s',a')\big]\big\},$$
$$\nabla_\omega L^\pi(\theta,\omega)=\mathbb{E}_{s,a,s',a'}\big\{\nabla_\omega\mu_\omega(s,a)\cdot\big[Q_\theta(s,a)-R(s,a)-\gamma\cdot Q_\theta(s',a')-\mu_\omega(s,a)\big]\big\}.$$

Therefore, replacing the primal and dual gradients by their unbiased estimates, we obtain Algorithm 2.

Algorithm 2 A Primal-Dual Algorithm for Policy Evaluation

Input: Initial parameter estimates $\theta_0\in\mathbb{R}^d$ and $\omega_0\in\mathbb{R}^p$, primal and dual stepsizes $\{\alpha_t,\beta_t\}_{t\geq 0}$.
for $t=0,1,2,\ldots$ until convergence **do**
Sample $(s_t,a_t)\sim d^\pi$, observe $r(s_t,a_t)$ and the next state s_t'. Sample action $a'\sim\pi(\cdot\,|\,s')$.
Update the parameters by

$$\omega_{t+1}\leftarrow\Pi_\Omega\big\{\omega_t+\beta_t\cdot\nabla_\omega\mu_{\omega_t}(s_t,a_t)\cdot\big[Q_{\theta_t}(s_t,a_t)-r(s_t,a_t)-\gamma\cdot Q_{\theta_t}(s_t',a')-\mu_{\omega_t}(s_t,a_t)\big]\big\},$$
$$\theta_{t+1}\leftarrow\Pi_\Theta\big\{\theta_t-\alpha_t\cdot\mu_{\omega_t}(s_t,a_t)\cdot\big[\nabla_\theta Q_{\theta_t}(s_t,a_t)-\gamma\cdot\nabla_\theta Q_{\theta_t}(s_t',a')\big]\big\}.$$

end for

Furthermore, soft Q-learning is proposed Haarnoja et al. (2017) based on the maximum entropy principle. This problem aims to find the fixed point of the soft Bellman operator $\mathcal{T}^\sharp$ defined by

$$(\mathcal{T}^\sharp Q)(s,a)=r(s,a)+\gamma\cdot\mathbb{E}\bigg(\tau\cdot\log\bigg\{\sum_{a\in\mathcal{A}}\exp[Q(s_{t+1},a)/\tau]\bigg\}\,\bigg|\,s_t=s,a_t=a\bigg), \tag{A.2}$$

where $\tau>0$ is the temperature parameter. By definition, $\mathcal{T}^\sharp$ can be seen as a smooth approximation of the Bellman optimality operator, where the $\max$ function in (2.2) is replaced by the softmax function, where parameter τ controls the approximation error. It is known that $\mathcal{T}^\sharp$ also admits a unique fixed point, denoted by $Q^\sharp\colon\mathcal{S}\times\mathcal{A}\to\mathbb{R}$. To estimate $Q^\sharp$ with function approximation, similar to (3.5) and (A.1), we consider the saddle point problem $\min_{\theta\in\mathbb{R}^d}\max_{\omega\in\mathbb{R}^p}L^\sharp(\theta,\omega)$, where $L^\sharp(\theta,\omega)$ is given by

$$L^\sharp(\theta,\omega)=\mathbb{E}_{s,a,s'}\bigg[-1/2\cdot[\mu_\omega(s,a)]^2+\bigg(Q_\theta(s,a)-r(s,a)-\gamma\cdot\tau\cdot\log\bigg\{\sum_{a\in\mathcal{A}}\exp[Q_\theta(s',a)/\tau]\bigg\}\bigg)\cdot\mu_\omega(s,a)\bigg],$$

where $(s,a)\sim\rho$ and $s'\sim P(\cdot\,|\,s,a)$. Here ρ is the stationary distribution on $\mathcal{S}\times\mathcal{A}$ induced by the behavioral policy π_b. By direct computation, the gradients of $L^\sharp(\theta,\omega)$ with respect to θ and ω are given by

$$\nabla_\theta L^\sharp(\theta,\omega)=\mathbb{E}_{s,a,s',a'}\bigg\{\mu_\omega(s,a)\cdot\bigg[\nabla_\theta Q_\theta(s,a)-\gamma\cdot\sum_{a\in\mathcal{A}}\nu_\theta(s',a)\cdot\nabla Q_\theta(s',a)\bigg]\bigg\},$$
$$\nabla_\omega L^\sharp(\theta,\omega)=\mathbb{E}_{s,a,s',a'}\bigg[\nabla_\omega\mu_\omega(s,a)\cdot\bigg(Q_\theta(s,a)-R(s,a)-\gamma\cdot\tau\cdot\log\bigg\{\sum_{a\in\mathcal{A}}\exp[Q_\theta(s',a)/\tau]\bigg\}-\mu_\omega(s,a)\bigg)\bigg],$$

where we define $\nu_\theta\colon\mathcal{S}\times\mathcal{A}\to[0,1]$ by

$$\nu_\theta(s,a)=\frac{\exp[Q_\theta(s,a)/\tau]}{\sum_{a'\in\mathcal{A}}\exp[Q_\theta(s,a)/\tau]},\qquad\forall(s,a)\in\mathcal{S}\times\mathcal{A},\forall\theta\in\mathbb{R}^d.$$

Thus, we obtain a stochastic gradient algorithm for estimating $Q^\sharp$, which is stated in Algorithm 3.

Algorithm 3 A Primal-Dual Algorithm for Soft Q-Learning

Input: Initial parameter estimates $\theta_0 \in \mathbb{R}^d$ and $\omega_0 \in \mathbb{R}^p$, primal and dual stepsizes $\{\alpha_t, \beta_t\}_{t\geq 0}$.
for $t = 0, 1, 2, \ldots$ until convergence **do**
Sample $(s_t, a_t) \sim \rho$, observe reward $r(s_t, a_t)$ and the next state s_t'.
Sample action $a_t' = b$ with proabrbility $\nu_{\theta_t}(s_t', b)$ for any $b \in \mathcal{A}$.
Compute the TD error $\delta_t = Q_{\theta_t}(s_t, a_t) - r(s_t, a_t) - \gamma \cdot \tau \cdot \log\{\sum_{a\in\mathcal{A}} \exp[Q_{\theta_t}(s_t', a)/\tau]\}$.
Update the parameters by

$$\omega_{t+1} \leftarrow \Pi_\Omega\big\{\omega_t + \beta_t \cdot \nabla_\omega \mu_{\omega_t}(s_t, a_t) \cdot \big[\delta_t - \mu_{\omega_t}(s_t, a_t)\big]\big\},$$
$$\theta_{t+1} \leftarrow \Pi_\Theta\big\{\theta_t - \alpha_t \cdot \mu_{\omega_t}(s_t, a_t) \cdot \big[\nabla_\theta Q_{\theta_t}(s_t, a_t) - \gamma \cdot \nabla_\theta Q_{\theta_t}(s_t', a_t')\big]\big\}.$$

end for

B Proofs of the Main Results

B.1 Proof of Theorem 4.3

Proof. The proof of this theorem consists of two steps, where we consider the faster and slower timescales seperately.

Step 1. Faster timescale. We first consider the convergence of $\{\theta_t, \omega_t\}_{t\geq 1}$ in the faster timescale. Using ODE approximation, we will show that the sequence of dual variables $\{\omega_t\}_{t\geq 0}$ generated from (3.8) tracks the solution of the inner maximization problem, i.e., $\mathrm{argmax}_\omega\, L(\theta_t, \omega)$. To begin with, for notational simplicity, we define $\phi_t = \phi(s_t, a_t)$,

$$\delta_t = Q_{\theta_t}(s_t, a_t) - r(s_t, a_t) - \gamma \cdot Q_{\theta_t}(s_t', a_t'), \quad A_t = \nabla_\theta Q_{\theta_t}(s_t, a_t) - \gamma \cdot \nabla_\theta Q_{\theta_t}(s_t', a_t'), \tag{B.1}$$

for all $t \geq 0$, where $a_t' = \mathrm{argmax}_{a\in\mathcal{A}}\, Q_{\theta_t}(s_t', a)$. Then the updating rules in Algorithms 1 reduce to

$$\omega_{t+1} \leftarrow \Pi_\Omega\big[\omega_t + \beta_t \cdot (\delta_t - \phi_t^\top \omega_t) \cdot \phi_t\big], \qquad \theta_{t+1} \leftarrow \Pi_\Theta\big[\theta_t - \alpha_t \cdot \phi_t^\top \omega_t \cdot A_t\big], \tag{B.2}$$

where $\{\alpha_t, \beta_t\}_{t\geq 0}$ are the learning rates, and Π_Θ is the projection operator. Moreover, we define $\mathcal{F}_t = \sigma(\{s_\tau, a_\tau, r(s_\tau, a_\tau), s_\tau'\}_{\tau\leq t})$ as the σ-algebra generated by the history until time t. Thus, θ_t and ω_t are $\mathcal{F}_{t-1}$-measurable for any $t \geq 1$. Furthermore, for any $\theta \in \mathbb{R}^d$ and $\omega \in \mathbb{R}^p$, we define

$$h(\theta, \omega) = \mathbb{E}_{(s,a)\sim\rho}\big\{\phi(s, a) \cdot [Q_\theta(s, a) - (\mathcal{T}^* Q_\theta)(s, a) - \phi(s, a)^\top \omega]\big\}, \tag{B.3}$$
$$g(\theta, \omega) = \mathbb{E}_{(s,a)\sim\rho}\big(\phi(s, a)^\top \omega \cdot \big\{\nabla_\theta Q_\theta(s, a) - [\nabla_\theta(\mathcal{T}^* Q_\theta)](s, a)\big\}\big). \tag{B.4}$$

Then by definition, we have $\mathbb{E}[\delta_t - \phi_t^\top \omega_t) \cdot \phi_t \,|\, \mathcal{F}_{t-1}] = h(\theta_t, \omega_t)$ and $\mathbb{E}[\phi_t^\top \omega_t \cdot A_t \,|\, \mathcal{F}_{t-1}] = g(\theta_t, \omega_t)$ for any $t \geq 1$, which implies that $\{\zeta_t\}_{t\geq 1}$ and $\{\eta_t\}_{t\geq 1}$ defined by

$$\zeta_t = (\delta_t - \phi_t^\top \omega_t) \cdot \phi_t - h(\theta_t, \omega_t), \qquad \eta_t = \phi_t^\top \omega_t \cdot A_t - g(\theta_t, \omega_t) \tag{B.5}$$

are two martingale difference sequence with respect to filtration $\{\mathcal{F}_t\}_{t\geq 1}$. Moreover, we define $\mathcal{C}_\Theta(\theta)$ as the outer normal cone of Θ at $\theta \in \Theta$, and define $\mathcal{C}_\Omega(\omega)$ for Ω similarly.

Thus, (B.2) can be written as

$$\begin{aligned}\begin{pmatrix}\theta_{t+1}\\ \omega_{t+1}\end{pmatrix} &= \Pi_{\Theta\times\Omega}\left[\begin{pmatrix}\theta_t\\ \omega_t\end{pmatrix} - \beta_t \cdot \begin{pmatrix}\alpha_t/\beta_t \cdot A_t \cdot \phi_t^\top \omega_t\\ \phi_t \phi_t^\top \omega_t - \delta_t \cdot \phi_t\end{pmatrix}\right]\\ &= \begin{pmatrix}\theta_t\\ \omega_t\end{pmatrix} - \beta_t \cdot \begin{pmatrix}\alpha_t/\beta_t \cdot A_t \cdot \phi_t^\top \omega_t\\ \phi_t \phi_t^\top \omega_t - \delta_t \cdot \phi_t\end{pmatrix} + \begin{pmatrix}\xi_t^\theta\\ \xi_t^\omega\end{pmatrix},\end{aligned} \tag{B.6}$$

where $\Pi_{\Theta\times\Omega}$ is the projection onto product set $\Theta \times \Omega = \{(u, v) \colon u \in \Theta, v \in \Omega\}$, and ζ_t^θ and ζ_t^ω in (B.5) are the correction terms induced by projection. Thus, by definition, we have $\xi_t^\theta \in -\mathcal{C}_\Theta(\theta_{t+1})$ and $\xi_t^\omega \in -\mathcal{C}_\Omega(\omega_{t+1})$, respectively.

Furthermore, under Assumption 4.1, both $Q_\theta(s, a)$, $\nabla_\theta Q_\theta(s, a)$, and $\phi(s, a)$ are bounded, which implies that there exist a constant $C > 0$ such that $\mathbb{E}[\|\zeta_t\|_2^2 + \|\eta_t\|_2^2 \,|\, \mathcal{F}_{t-1}] \leq C$ for all $t \geq 1$.

Moreover, let $M_T = \sum_{t=1}^{T} \beta_t \cdot ({\xi_t^\theta}^\top, {\xi_t^\omega}^\top)^\top \in \mathbb{R}^{d+p}$ for any $T \geq 1$. Since $\sum_{t\geq 1} \beta_t^2 < \infty$, $\{M_T\}_{T\geq 1}$ is a square-integrable martingale sequence. Moreover, by the Martingale convergence theorem (Proposition VII-2-3(c) on page 149 of Neveu (1975)), $\{M_T\}_{T\geq 1}$ converges almost surely. Thus, for any $\epsilon > 0$, by Doob's martingale inequality Doob (1953), we have

$$\mathbb{P}\Big(\sup_{N\geq T} \|M_N - M_T\| \geq \epsilon \Big) \leq \frac{\sup_{N\geq T} \mathbb{E}[\|M_N - M_T\|_2^2]}{\epsilon^2} \leq \frac{2C^2 \sum_{t\geq T} \beta_t^2}{\epsilon^2},$$

which converges to zero as T goes to infinity. Furthermore, recall that, under the two-timescale assumption of the learning rates, we have $\alpha_t/\beta_t \to 0$ as t goes to infinity, which implies that $\lim_{t\to\infty} \alpha_t/\beta_t \cdot A_t \cdot \phi_t^\top \omega_t = 0$.

Now we apply the Kushner-Clark lemma (Kushner & Clark, 1978, see also Theorem D.2 in §D for details) to sequence $\{(\theta_t^\top, \omega_t^\top)^\top\}_{t\geq 1}$, which implies that the asymptotic behavior of $\{(\theta_t, \omega_t)\}_{t\geq 1}$ is characterized by the projected ODE

$$\dot{\theta} = 0 + \xi^\theta, \qquad \dot{\omega} = h(\theta, \omega) + \xi^\omega, \tag{B.7}$$

where h is defined in (B.3), and ξ^θ and ξ^ω satisfy $\xi^\theta(t) \in -\mathcal{C}_\Theta(\theta(t))$ and $\xi^\omega(t) \in -\mathcal{C}_\Theta(\omega(t))$ for any $t \geq 0$. Specifically, sequence $\{(\theta_t^\top, \omega_t^\top)^\top\}_{t\geq 1}$ converges almost surely to the set of asymptotically stable equilibria of the projected ODE in (B.7), which is given by

$$\big\{(\theta^*, \omega^*) \colon \theta^* \in \Theta, h(\theta^*, \omega^*) \in \mathcal{C}_\Omega(\omega^*)\big\}. \tag{B.8}$$

Recall that Ω is the Euclidean ball in $\mathbb{R}^p$ with radius R_ω. Thus, the boundary of Ω, $\partial\Omega$, is $\{\omega \colon \|\omega\|_2 = R_\omega\}$. For any $\omega \in \partial\Omega$, the outer normal cone is $\mathcal{C}_\Omega(\omega) = \{\lambda \cdot \omega \colon \lambda \geq 0\}$.

In the sequel, we show that, for any (θ^*, ω^*) in the equilibria in (B.8), ω^* is in the interior of Ω, i.e., $\|\omega^*\|_2 < R_\omega$. This implies that $\{(\theta_t, \omega_t)\}_{t\geq 1}$ converges almost surely to $\{(\theta^*, \omega^*) \colon \theta^* \in \Theta, h(\theta^*, \omega^*) = 0\}$.

Then, by definition, we have $h[\theta, \omega(\theta)] = 0$ for any $\theta \in \mathbb{R}^d$, which implies that $\omega(\theta)$ is the unique maximizer of $\max_{\omega\in\mathbb{R}^p} L(\theta, \omega)$ since $L(\theta, \omega)$ is a strictly concave quadratic function of ω and $h(\theta, \omega) = \nabla_\omega L(\theta, \omega)$. Moreover, since $\|\phi(s, a)\|_2$ is bounded for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and the eigenvalues of $\mathbb{E}_{(s,a)\sim\rho}[\phi(s, a)\phi(s, a)^\top]$ are at least $\sigma_{\min} > 0$, by norm inequality, we have

$$\|\omega(\theta)\|_2 \leq 2Q_{\max}/\sigma_{\min} \cdot \sup_{(s,a)} \|\phi(s, a)\|_2. \tag{B.9}$$

Thus, if R_ω is larger than the right-hand side of (B.9), $\omega(\theta)$ is in the interior of Ω for any $\theta \in \Theta$.

Now we assume to the contrary that $\omega^* \in \partial\Omega$. By (B.8), there exists $\lambda > 0$ such that

$$\lambda \cdot \omega^* = h(\theta^*, \omega^*) = \mathbb{E}_{(s,a)\sim\rho}[\phi(s, a)\phi(s, a)^\top] \cdot [\omega(\theta^*) - \omega^*], \tag{B.10}$$

where the second equality follows from (B.3) and (4.1). For the right-hand side of (B.10), we have

$$[\omega(\theta^*) - \omega^*]^\top \cdot \mathbb{E}_{(s,a)\sim\rho}[\phi(s, a)\phi(s, a)^\top] \cdot [\omega(\theta^*) - \omega^*] \geq c_{\min} \cdot \|\omega^* - \omega(\theta^*)\|_2^2 > 0. \tag{B.11}$$

However, since $\omega(\theta^*)$ is in the interior of Ω, we have

$$\lambda \cdot \langle \omega(\theta^*) - \omega^*, \omega^* \rangle = \lambda \cdot \|\omega^*\|_2 \cdot \|\omega(\theta^*)\|_2 - \lambda \cdot \|\omega^*\|_2^2 < 0. \tag{B.12}$$

Thus, the contradiction between (B.11) and (B.12) implies that our assumption that $\omega^* \in \partial\Omega$ is not true, i.e., ω^* is in the interior of Ω. In this case, $\mathcal{C}_\Omega(\omega^*)$ is an empty set. Thus the asymptotically stable equilibria of the ODE in (B.7) are given by

$$\big\{(\theta^*, \omega^*) \colon \theta^* \in \Theta, h(\theta^*, \omega^*) = 0\big\} = \big\{[\theta^*, \omega(\theta^*)] \colon \theta^* \in \Theta\big\}. \tag{B.13}$$

This implies that, in the faster timescale, we could fix the primal parameter at $\theta \in \Theta$, and the dual variable converges to $\omega(\theta)$, which is the unique solution of the dual optimization problem $\max_\omega L(\theta, \omega)$. In other words, using two timescale updates, we essentially solve the dual problem for each θ.

Step 2. Slower timescale. Note that (B.13) cannot characterize the asymptotic behavior of the primal variable. Now we proceed to establish a finer characterization of the asymptotic behavior of $\{\theta_t, \omega_t\}_{t\geq 1}$ by looking into the slower timescale. For ease of presentation, let $\mathcal{E}_t = \sigma(\theta_\tau, \tau \leq t)$ be the σ-field generated by $\{\theta_\tau, \tau \leq t\}$. In addition, we define

$$\psi_t^{(1)} = -\phi_t^\top \omega_t \cdot A_t + \mathbb{E}[\phi_t^\top \omega_t \cdot A_t \,|\, \mathcal{E}_t] \qquad \psi_t^{(2)} = -\mathbb{E}[\phi_t^\top \omega_t \cdot A_t \,|\, \mathcal{E}_t] + g[\theta_t, \omega(\theta_t)], \quad \text{(B.14)}$$

where function g is defined in (B.4), A_t and ϕ_t are defined in (B.1). It holds that $\{\psi_t^{(1)}\}_{t\geq 1}$ is a martingale difference sequence. Note that by the definition of g, we have $\mathbb{E}[\phi_t^\top \omega(\theta_t) \cdot A_t \,|\, \mathcal{E}_t] = g[\theta_t, \omega(\theta_t)]$. Using the notation in (B.14), the primal update in (B.2) can be written as

$$\theta_{t+1} = \Pi_\Theta\big[\theta_t - \alpha_t \cdot g[\theta_t, \omega(\theta_t)] + \alpha_t \cdot \psi_t^{(1)} + \alpha_t \cdot \psi_t^{(2)}\big]. \quad \text{(B.15)}$$

As shown in the first step of the proof, $\omega(\theta_t) - \theta_t$ converges to zero as t goes to infinity. Moreover, under Assumption 4.1, both $\nabla_\theta Q_\theta(s,a)$ and $\phi(s,a)$ are bounded. By Cauchy-Schwarz inequality,

$$\begin{aligned}\|\psi_t^{(2)}\|_2 &= \big\|\mathbb{E}\{\phi_t^\top [\omega(\theta_t) - \omega_t] \cdot A_t \,|\, \mathcal{E}_t\}\big\|_2 \\ &\leq \sup_{(s,a)} \|\phi(s,a)\|_2 \cdot \sup_{(s,a)} \|\nabla_\theta Q_\theta(s,a)\|_2 \cdot \mathbb{E}[\|\omega(\theta_t) - \omega_t\|_2 \,|\, \mathcal{E}_t],\end{aligned}$$

which converges to zero almost surely. Besides, the boundedness of $\nabla_\theta Q_\theta(s,a)$ and $\phi(s,a)$ also implies that there exist a constant $C > 0$ such that $\mathbb{E}[\|\psi_t^{(1)}\|_2^2 \,|\, \mathcal{E}_t] \leq C$ for all $t \geq 1$.

Furthermore, we define $W_T = \sum_{t=1}^T \alpha_t \cdot \psi_t^{(1)} \in \mathbb{R}^d$ for any $T \geq 1$. Since $\sum_{t\geq 1} \alpha_t^2 < \infty$, $\{W_T\}_{T\geq 1}$ is a square-integrable martingale sequence, which converges almost surely by the Martingale convergence theorem (Neveu, 1975). Moreover, Doob's martingale inequality (Doob, 1953) implies that

$$\lim_{T\to\infty} \mathbb{P}\Big(\sup_{N\geq T} \|W_N - W_T\| \geq \epsilon\Big) \leq \lim_{T\to\infty} \frac{\sup_{N\geq T} \mathbb{E}[\|W_N - W_T\|_2^2]}{\epsilon^2} \leq \lim_{T\to\infty} \frac{2C^2 \sum_{t\geq T} \alpha_t^2}{\epsilon^2} = 0.$$

To apply the Kushner-Clark lemma, we additionally need to verify that function $\bar{g}(\theta) = g[\theta, \omega(\theta)]$ is a continuous. To see this, note that $\omega(\theta)$ defined in (4.1) is a continuous function, and $g(\theta, \omega)$ in (B.4) is continuous in both θ and ω.

Finally, applying by the Kushner-Clark lemma (Kushner & Clark, 1978) to sequence $\{\theta_t\}_{t\geq 1}$, it holds that $\{\theta_t\}_{t\geq 1}$ converges almost surely to the set of asymptotically stable equilibria of the ODE

$$\dot{\theta} = \bar{g}(\theta) + \xi^\theta, \qquad \xi^\theta(t) \in -\mathcal{C}_\Theta(\theta(t)), \quad \text{(B.16)}$$

where $\mathcal{C}_\Theta(\theta)$ is the outer normal cone of Θ. Since Θ is a Euclidean ball with radius R_θ, if θ is on the boundary of Θ, i.e., $\|\theta\|_2 = R_\theta$, we have $\mathcal{C}_\Theta(\theta) = \{\lambda \cdot \theta, \lambda \geq 0\}$. Thus, for any asymptotically stable equilibrium θ^* of (B.16), if θ^* is in the interior of Θ, i.e., $\|\theta\| < R_\theta$, we have $\bar{g}(\theta^*) = 0$. Additionally, if $\|\theta^*\|_2 = R_{\max}$, we have $\bar{g}(\theta^*) \in \mathcal{C}_\Theta(\theta^*)$, which implies that there exists $\lambda > 0$ such that $\bar{g}(\theta^*) = \lambda \cdot \theta^*$. Therefore, we conclude the proof of Theorem 4.3.

□

B.2 Proof of Theorem 4.5

Proof. In the sequel, to simplify the notation, we use C to denote absolute constant, whose value might change from line to line. In addition, for any $\theta \in \mathbb{R}^d$, we define

$$\begin{aligned}\widetilde{\nabla} J(\theta) &= \mathbb{E}_{s,a,s'}\big\{\widetilde{\mu}_\theta(s,a) \cdot \big[\nabla_\theta Q_\theta(s,a) - \gamma \cdot \sum_{a\in\mathcal{A}} \mathbb{1}\big\{a' = \operatorname*{argmax}_{b\in\mathcal{A}} Q_\theta(s',b)\big\} \cdot \nabla_\theta Q_\theta(s',a')\big]\big\} \\ &= \mathbb{E}_{(s,a)\sim\rho}\big[\widetilde{\mu}_\theta(s,a) \cdot \nabla_\theta \delta_\theta(s,a)\big], \quad \text{(B.17)}\end{aligned}$$

where $\delta_\theta = Q_\theta - \mathcal{T}^* Q_\theta$ is the TD-error, and $\widetilde{\mu}_\theta$ is the output of the optimization oracle for query θ.

Note that θ_t in (4.5) is updated in the direction of an unbiased estimate of $\widetilde{\nabla} J(\theta_t)$. To simplify the notation, we let

$$\zeta_t = \widetilde{\mu}_{\theta_t}(s_t, a_t) \cdot \big[\nabla_\theta Q_{\theta_t}(s_t, a_t) - \nabla_\theta Q_{\theta_t}(s_t', a_t')\big] - \widetilde{\nabla}_\theta J(\theta_t), \tag{B.18}$$

where $\widetilde{\nabla}_\theta J(\theta_t)$ is defined in (B.17), $a_t' = \mathrm{argmax}_{a \in \mathcal{A}} Q_{\theta_t}(s_t', a)$. Then the update rule in (4.5) can be written as $\theta_{t+1} = \theta_t - \alpha_t \cdot [\widetilde{\nabla} J(\theta_t) + \zeta_t]$. By Assumptions 4.1 and 4.2, $\{\zeta_t\}_{t \geq 0}$ is a sequence of bounded and centered random vectors. Moreover, since both $\widetilde{\mu}_\theta$ and $\nabla_\theta Q_\theta$ are bounded by $Q_{\max}$ and $G_{\max}$ on $\mathcal{S} \times \mathcal{A}$, respectively, ζ_t defined in (B.18) is a bounded random variable satisfying $\|\zeta_t\|_2 \leq 4Q_{\max} \cdot G_{\max}$. Let $\mathcal{F}_t$ be the σ-field generated by $\{\theta_j, j \leq t\}$. Then we have $\mathbb{E}(\xi_t \,|\, \mathcal{F}_t) = 0$ and $\mathbb{E}(\|\xi_t\|^2 \,|\, \mathcal{F}_t) \leq C$ for some constant $C > 0$.

Moreover, note that the gradient of $J(\theta)$ can be written as $\nabla_\theta J(\theta) = \mathbb{E}_{(s,a)\sim\rho}[\delta_\theta(s,a) \cdot \nabla_\theta \delta_\theta(s,a)]$. By the definition of the TD-error, under Assumption 4.1, δ_θ and $\nabla_\theta \delta_\theta$ are bounded by $2Q_{\max}$ and $2G_{\max}$ on $\mathcal{S} \times \mathcal{A}$ respectively. Moreover, since $\nabla_\theta \delta_\theta$ is Lipschitz in θ, there exists a constant $L > 0$ such that $\nabla_\theta J(\theta)$ is L-Lipschitz. Thus, we have

$$\begin{aligned} J(\theta) &\leq J(\theta_t) - \langle \nabla_\theta J(\theta_t), \theta_{t+1} - \theta_t \rangle + L/2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 \\ &\leq J(\pi_{\theta_t}) - \alpha_t \cdot \big\langle \nabla_\theta J(\theta_t), \widetilde{\nabla} J(\theta_t) + \zeta_t \big\rangle + \alpha_t^2 \cdot L/2 \cdot \|\widetilde{\nabla} J(\theta_t) + \zeta_t\|_2^2. \end{aligned} \tag{B.19}$$

Taking conditional expectation on both sides of (B.19) given $\mathcal{F}_t$, since ζ_t is centered with finite variance, we obtain that

$$\mathbb{E}[J(\theta_{t+1}) \,|\, \mathcal{G}_t] \leq J(\theta_t) - \alpha_t \cdot \big\langle \nabla_\theta J(\theta_t), \widetilde{\nabla} J(\theta_t) \big\rangle - \alpha_t^2 \cdot L/2 \cdot \|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2 - C \cdot \alpha_t^2, \tag{B.20}$$

where $C > 0$ is an absolute constant. Note that $\widetilde{\nabla} J(\theta)$ is a biased estimate of $\nabla_\theta J(\theta)$. Under Assumptions 4.1 and 4.4, the bias can be bounded via Cauchy-Schwarz inequality:

$$\begin{aligned} \|\nabla_\theta J(\theta) - \widetilde{\nabla} J(\theta)\|_2 &= \big\|\mathbb{E}_{(s,a)\sim\rho}\big\{\nabla_\theta \delta_\theta(s,a) \cdot [\delta_\theta(s,a) - \widetilde{\mu}_\theta(s,a)]\big\}\big\|_2 \\ &\leq \mathbb{E}_{(s,a)\sim\rho}\big[\|\nabla_\theta \delta_\theta(s,a)\|_2\big] \cdot \|\delta_\theta - \widetilde{\mu}_\theta\|_\rho \leq 2G_{\max} \cdot \varepsilon, \end{aligned} \tag{B.21}$$

where ε is the error of the optimization oracle in Assumption 4.4. We denote $\widetilde{\nabla} J(\theta_t) - \nabla_\theta J(\theta_t)$ by ψ_t hereafter to simplify the notation. Moreover, by Cauchy-Schwarz inequality, we have

$$\langle \nabla_\theta J(\theta_t), \widetilde{\nabla} J(\theta_t) \rangle = \langle \nabla_\theta J(\theta_t), \nabla_\theta J(\theta_t) - \psi_t \rangle \geq \|\nabla_\theta J(\theta_t)\|_2^2 - \|\nabla_\theta J(\theta_t)\|_2 \cdot \|\psi_t\|_2. \tag{B.22}$$

Similarly, for $\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$, we obtain that

$$\|\widetilde{\nabla} J(\theta_t)\|_2^2 \leq \|\nabla_\theta J(\theta_t)\|_2^2 + \|\psi_t\|_2^2 + 2 \cdot \|\nabla_\theta J(\theta_t)\|_2 \cdot \|\psi_t\|_2. \tag{B.23}$$

Combining (B.20), (B.21), (B.22), and (B.23), we have

$$\begin{aligned} &\mathbb{E}[J(\theta_{t+1}) \,|\, \mathcal{F}_t] \\ &\quad\leq J(\theta_t) + C \cdot \alpha_t^2 - \alpha_t \cdot (1 - \alpha_t \cdot L/2) \cdot \|\nabla_\theta J(\theta_t)\|_2^2 \\ &\qquad\quad + \alpha_t \cdot (1 + \alpha_t \cdot L) \cdot \|\nabla_\theta J(\theta_t)\|_2 \cdot \|\psi_t\|_2 \\ &\quad\leq J(\theta_t) + C \cdot \alpha_t^2 - \alpha_t \cdot (1 - \alpha_t \cdot L/2) \cdot \|\nabla_\theta J(\theta_t)\|_2^2 \\ &\qquad\quad + \alpha_t \cdot 2(1 + \alpha_t \cdot L) \cdot G_{\max} \cdot \varepsilon \cdot \|\nabla_\theta J(\theta_t)\|_2. \end{aligned} \tag{B.24}$$

Since $\sum_{t \geq 0} \alpha_t^2 < \infty$, α_t converges to zero as t goes to infinity. When t is sufficiently large such that $4\alpha_t \cdot L < 1$, (B.24) implies that

$$\begin{aligned} \mathbb{E}[J(\theta_{t+1}) \,|\, \mathcal{F}_t] &\leq J(\theta_t) - \alpha_t/2 \cdot \|\nabla_\theta J(\theta_t)\|_2^2 + 4\alpha_t \cdot G_{\max} \cdot \varepsilon \cdot \|\nabla_\theta J(\theta_t)\|_2 + C \cdot \alpha_t^2 \\ &= J(\theta_t) - \alpha_t/2 \cdot \|\nabla_\theta J(\theta_t)\|_2 \cdot \big[\|\nabla_\theta J(\theta_t)\|_2 - 8G_{\max} \cdot \varepsilon\big] + C \cdot \alpha_t^2. \end{aligned} \tag{B.25}$$

Furthermore, since $J(\theta)$ is a nonnegative function and $\sum_{t \geq 1} \alpha_t^2 < \infty$, by (B.25) we have

$$\sum_{t \geq 0} \alpha_t \cdot \|\nabla_\theta J(\theta_t)\|_2 \cdot \big[\|\nabla_\theta J(\theta_t)\|_2 - 8G_{\max} \cdot \varepsilon\big] < \infty. \tag{B.26}$$

In the sequel, we show by contradiction that every limit point θ^* of $\{\theta_t\}_{t\geq 1}$ satisfies that $\|\nabla_\theta J(\theta^*)\|_2 \leq 8G_{\max}\cdot\varepsilon$. Let $\nu > 0$ be an arbitrary number. We first show that $\{t\colon \|\nabla_\theta J(\theta_t)\|_2 < 8G_{\max}\cdot\varepsilon + \nu\}$ is an infinite set. Suppose this is false, then there exists an integer t_0 such that $\|\nabla_\theta J(\theta_t)\|_2 \geq 8G_{\max}\cdot\varepsilon + \nu$ for all $t \geq t_0$. Since $\sum_{t\geq 0}\alpha_t = \infty$, we have

$$\sum_{t\geq t_0}\alpha_t\cdot\|\nabla_\theta J(\theta_t)\|_2\cdot\big[\|\nabla_\theta J(\theta_t)\|_2 - 8G_{\max}\cdot\varepsilon\big] \geq \sum_{t\geq t_0}\alpha_t\cdot\nu\cdot(8G_{\max}\cdot\varepsilon+\nu) = \infty,$$

which contradicts (B.26). Thus, there are infinite θ_t's satisfying $\|\nabla_\theta J(\theta_t)\|_2 \leq 8G_{\max}\epsilon + \nu$. Since ν can be arbitrarily small, this implies

$$\liminf_{t\geq 0}\|\nabla_\theta J(\theta_t)\|_2 \leq 8G_{\max}\cdot\varepsilon. \tag{B.27}$$

Let $\varepsilon_b = 8G_{\max}\cdot\varepsilon$. It remains to show that $\limsup_{t\geq 0}\|\nabla_\theta J(\theta_t)\|_2 \leq \varepsilon_b$, which, combined with (B.27), establishes the theorem.

Suppose this argument does not hold, then for any $\nu > 0$, there exists $\varepsilon > 0$ such that $\|\nabla_\theta J(\theta_t)\|_2 \geq \delta_b + 2\nu$ for infinitely many t. Moreover, for this particular ν, $\|\nabla_\theta J(\theta_t)\|_2 \leq \delta_b + \nu$ also holds for infinitely many t. We define index sets $\mathcal{N}_1$ and $\mathcal{N}_2$ by

$$\mathcal{N}_1 = \{\theta_t\colon \|\nabla_\theta J(\theta_t)\|_2 \geq \delta_b + 2\nu\}, \qquad \mathcal{N}_2 = \{\theta_t\colon \|\nabla_\theta J(\theta_t)\|_2 \leq \delta_b + \nu\},$$

which are two disjoint and closed sets by the continuity of $\|\nabla_\theta J(\theta)\|_2$. Moreover, we define

$$D(\mathcal{N}_1,\mathcal{N}_2) = \inf_{\theta\in\mathcal{N}_1}\inf_{\theta'\in\mathcal{N}_2}\|\theta-\theta'\|_2, \tag{B.28}$$

which is a positive number since $\mathcal{N}_1$ and $\mathcal{N}_2$ are disjoint. In addition, since both $\mathcal{N}_1$ and $\mathcal{N}_2$ are infinite sets, there exists an index set $\mathcal{I}\subseteq\mathbb{N}$ such that the subsequence $\{\theta_t\}_{t\in\mathcal{I}}$ of $\{\theta_t\}_{t\geq 0}$ crosses $\mathcal{N}_1$ and $\mathcal{N}_2$ infinitely often. That is, there exists two sequences $\{s_i\}_{i\geq 1}$ and $\{t_i\}_{i\geq 1}\subseteq\mathbb{N}$ such that $\{\theta_t\}_{t\in\mathcal{I}} = \bigcup_{i\geq 1}\{\theta_{s_i},\theta_{s_{i+1}},\ldots,\theta_{t_i-1}\}$. Furthermore, we have $\{\theta_{s_i}\}_{i\geq 1}\subseteq\mathcal{N}_1$, $\{\theta_{t_i}\}_{i\geq 1}\subseteq\mathcal{N}_2$, and

$$\delta_b + \varepsilon \leq \|\nabla_\theta J(\theta_\ell)\|_2 \leq \delta_b + 2\varepsilon \qquad \text{for all} \quad \ell\in\cup_{i\geq 1}\{s_i+1,\ldots,t_i-1\}.$$

Hence, by triangle inequality, we have

$$\sum_{t\in\mathcal{I}}\|\theta_{t+1}-\theta_t\|_2 = \sum_{i=1}^{\infty}\sum_{\ell=s_i}^{t_i-1}\|\theta_{\ell+1}-\theta_\ell\|_2 \geq \sum_{i=1}^{\infty}\|\theta_{t_i}-\theta_{s_i}\|_2 \geq \sum_{i=1}^{\infty}D(\mathcal{N}_1,\mathcal{N}_2) = \infty, \tag{B.29}$$

where $D(\mathcal{N}_1,\mathcal{N}_2) > 0$ is defined in (B.28). Moreover, by (B.26), we have

$$\infty > \sum_{t\in\mathcal{I}}\alpha_t\cdot\|\nabla_\theta J(\theta_t)\|_2\cdot\big[\|\nabla_\theta J(\theta_t)\|_2 - \delta_b\big] \geq \sum_{t\in\mathcal{I}}\alpha_t\cdot(\delta_b+\nu)\cdot\nu,$$

which further implies that $\sum_{t\in\mathcal{I}}\alpha_t < \infty$. However, by the update rule in (4.5), under Assumption 4.1, it holds that $\|\theta_{t+1}-\theta_t\|_2 \leq \alpha_t\cdot 2Q_{\max}\cdot G_{\max}$, where we use the boundedness of $\widetilde{\mu}_\theta(s,a)$ and $\nabla_\theta Q_\theta(s,a)$. Thus, it holds that

$$\sum_{t\in\mathcal{I}}\|\theta_{t+1}-\theta_t\|_2 \leq 2Q_{\max}\cdot G_{\max}\sum_{t\in\mathcal{I}}\alpha_t < \infty,$$

which contradicts (B.29). This contradiction implies that $\limsup_{t\geq 0}\|\nabla_\theta J(\theta_t)\|_2 \leq \varepsilon_b$. Therefore, we conclude the proof of Theorem 4.5. □

C STATISTICAL ERROR

In this section, using tools from nonparametric regression, we establish the statistical error of our saddle point formulation in (3.5). To this end, we assume that the state space $\mathcal{S}$ is a closed compact

subset of the Euclidean space $\mathbb{R}^m$ and let $\mathcal{H}$ be the Sobolev space $\mathcal{W}^k(\mathbb{R}^m)$ restricted on $\mathcal{S}$. In addition, we define

$$\mathcal{H}^{\mathcal{A}} = \big\{Q \in \mathcal{B}(\mathcal{S} \times \mathcal{A}) \colon Q(\cdot, a) \in \mathcal{H}, \forall a \in \mathcal{A}\big\}, \qquad \|Q\|_{\mathcal{H}} = \bigg[\sum_{a \in \mathcal{A}} \|Q(\cdot, a)\|_{\mathcal{W}^k}^2\bigg]^{1/2}$$

for any $Q \in \mathcal{H}^{\mathcal{A}}$, where $\|\cdot\|_{\mathcal{W}^k}$ is the Sobolev norm.

We consider the Bellman residual minimization problem with function class $\mathcal{H}^{\mathcal{A}}$ based on n i.i.d. observations $\{(s_i, a_i, s_i')\}$, where $(s_i, a_i) \sim \rho$ and s_i' is the next state. Replacing the loss function in (3.5) by its sample-based counterpart, we define

$$L_n(Q, \mu) = \sum_{i=1}^{n} \big\{-1/2 \cdot [\mu(s_i, a_i)]^2 + \big[Q(s_i, a_i) - r(s_i, a_i) - \gamma \cdot \max_{a_i' \in \mathcal{A}} Q(s_i', a_i')\big] \cdot \mu(s_i, a_i)\big\} \tag{C.1}$$

$$= \frac{1}{2} \cdot \sum_{i=1}^{n} \big\{ \big[Q(s_i, a_i) - r(s_i, a_i) - \gamma \cdot \max_{a_i' \in \mathcal{A}} Q(s_i', a_i')\big]^2 - \big[Q(s_i, a_i) - \mu(s_i, a_i) - r(s_i, a_i) - \gamma \cdot \max_{a_i' \in \mathcal{A}} Q(s_i', a_i')\big]^2 \big\}.$$

Hence, if we denote $Q - \mu$ in (C.1) by $h \in \mathcal{H}^{\mathcal{A}}$, $L_n(Q, \mu)$ becomes

$$\widetilde{L}_n(Q, h) = \frac{1}{2} \cdot \sum_{i=1}^{n} \big\{ \big[Q(s_i, a_i) - r(s_i, a_i) - \gamma \cdot \max_{a_i' \in \mathcal{A}} Q(s_i', a_i')\big]^2 - \big[h(s_i, a_i) - r(s_i, a_i) - \gamma \cdot \max_{a_i' \in \mathcal{A}} Q(s_i', a_i')\big]^2 \big\}. \tag{C.2}$$

Therefore, the sample-based sample-based functional optimization problem

$$\min_{Q \in \mathcal{H}^{\mathcal{A}}} \max_{h \in \mathcal{H}^{\mathcal{A}}} \widetilde{L}_n(Q, h) - \lambda_\mu \cdot \|h\|_{\mathcal{H}} + \lambda_Q \cdot \|Q\|_{\mathcal{H}} \tag{C.3}$$

reduces to the batch Bellman Residual Minimization algorithm studied in Antos et al. (2008b); Farahmand et al. (2009; 2016). Here λ_μ and λ_Q in (C.3) are two positive regularization parameters, and we adopt regularization to avoid overfitting. Moreover, since $\mu = Q - h$, the optimization problem in (C.3) is equivalent to

$$\min_{Q \in \mathcal{H}^{\mathcal{A}}} \max_{\mu \in \mathcal{H}^{\mathcal{A}}} \widetilde{L}_n(Q, h) - \lambda_\mu \cdot \|Q - \mu\|_{\mathcal{H}} + \lambda_Q \cdot \|Q\|_{\mathcal{H}}. \tag{C.4}$$

Let $(\widehat{Q}, \widehat{\mu})$ be the solution of (C.4). Using the theoretical result in Farahmand et al. (2009; 2016), we obtain the statistical rate of $\widehat{Q}$, which is stated in the following theorem.

Theorem C.1. Besides Assumption 3.1, we further make the following assumptions.

(a). The Sobolev space $W^k(\mathbb{R}^m)$ satisfies that $2k > m$.

(b). We assume that any $Q \in \mathcal{H}^{\mathcal{A}}$ satisfies that $|Q(s, a)| \le Q_{\max}$ with $Q_{\max} \ge R_{\max}/(1 - \gamma)$. Moreover, $\mathcal{H}^{\mathcal{A}}$ contains the optimal value function Q^*.

(c). For any $Q \in \mathcal{H}^{\mathcal{A}}$, there exists positive constants K_1 and K_2 such that $\|\mathcal{T}^* Q\|_{\mathcal{H}} \le K_1 + K_2 \cdot \|Q\|_{\mathcal{H}}$.

Moreover, let $\alpha = m/(2k)$. We set the regularization parameters in (C.4) to be

$$\lambda_\mu = \lambda_Q = \big(n \cdot \|Q^*\|_{\mathcal{H}}^2\big)^{-1/(1+\alpha)}.$$

Let $(\widehat{Q}, \widehat{\mu})$ be the solution of (C.4). Then for any $\eta \in (0, 1)$, we have

$$\|\widehat{Q} - Q^*\|_\rho^2 \le \frac{n^{-1/(1+\alpha)}}{(1-\gamma)^2} \cdot \big[C_1(Q^*, K_1, K_2) \cdot \log(1/\eta) + C_2(Q^*, K_1, K_2)\big]$$

with probability at least $1 - \eta$, where we define $C_1(Q^*, K_1, K_2) = C \cdot (K_1 + K_2^2) \cdot \|Q^*\|_{\mathcal{H}}^{2\alpha/(1+\alpha)}$ and $C_2(Q^*, K_1, K_2) = C \cdot K_1^2 \cdot (1 + \|Q^*\|_{\mathcal{H}}^{-2/(1+\alpha)})$ for some absolute constant $C > 0$.

Proof. The proof follows from the results in Farahmand et al. (2009; 2016). Note that the functional optimization problems in (C.4) and (C.3) are equivalent. Applying Theorem 11 in Farahmand et al. (2016) to the problem in (C.3), we obtain that

$$\|\widehat{Q} - T^*\widehat{Q}\|_\rho^2 \leq n^{-1/(1+\alpha)} \cdot \big[C_1(Q^*, K_1, K_2) \cdot \log(1/\eta) + C_2(Q^*, K_1, K_2)\big] \tag{C.5}$$

with probability at least $1 - \eta$. In addition, since Q^* is the unique fixed point of $\mathcal{T}^*$, by triangle inequality, we have

$$\|\widehat{Q} - Q^*\|_\rho \leq \|\widehat{Q} - \mathcal{T}^*\widehat{Q}\|_\rho + \|\mathcal{T}^*\widehat{Q} - \mathcal{T}^*Q^*\|_\rho \leq \|\widehat{Q} - \mathcal{T}^*\widehat{Q}\|_\rho + \gamma \cdot \|\widehat{Q} - Q^*\|_\rho,$$

where the last inequality holds since $\mathcal{T}^*$ is γ-contractive. Thus it holds that $\|\widehat{Q} - Q^*\|_\rho \leq (1 - \gamma)^{-1} \cdot \|\widehat{Q} - \mathcal{T}^*\widehat{Q}\|_\rho$. Therefore, by (C.5), we obtain that

$$\begin{aligned}\|\widehat{Q} - Q^*\|_\rho^2 &\leq (1-\gamma)^{-2} \cdot \|\widehat{Q} - \mathcal{T}^*\widehat{Q}\|_\rho^2 \\ &\leq n^{-1/(1+\alpha)} \cdot (1-\gamma)^{-2} \cdot \big[C_1(Q^*, K_1, K_2) \cdot \log(1/\eta) + C_2(Q^*, K_1, K_2)\big],\end{aligned}$$

which concludes the proof of Theorem C.1. □

D KUSHNER-CLARK LEMMA

We state here the well-known Kushner-Clark Lemma (Kushner & Clark, 1978; Metivier & Priouret, 1984; Prasad et al., 2014) in the sequel.

Let Γ be an operator that projects a vector onto a compact set $\mathcal{X} \subseteq \mathbb{R}^N$. Define a vector $\widehat{\Gamma}(\cdot)$ as

$$\widehat{\Gamma}[h(x)] = \lim_{0<\eta\to 0} \left\{ \frac{\Gamma[x + \eta h(x)] - x}{\eta} \right\},$$

for any $x \in \mathcal{X}$ and with $h : \mathcal{X} \to \mathbb{R}^N$ continuous. Consider the following recursion in N dimensions

$$x_{t+1} = \Gamma\big\{x_t + \gamma_t[h(x_t) + \xi_t + \beta_t]\big\}. \tag{D.1}$$

The ODE associated with (D.1) is given by

$$\dot{x} = \widehat{\Gamma}[h(x)]. \tag{D.2}$$

Assumption D.1. We make the following assumptions:

- $h(\cdot)$ is a continuous $\mathbb{R}^N$-valued function.
- The sequence $\{\beta_t\}$, $t \geq 0$ is a bounded random sequence with $\beta_t \to 0$ almost surely as $t \to \infty$.
- The stepsizes γ_t, $t \geq 0$ satisfy $\gamma_t \to 0$ as $t \to \infty$ and $\sum_t \gamma_t = \infty$.
- The sequence ξ_t, $t \geq 0$ satisfies for any $\epsilon > 0$

$$\lim_t \mathbb{P}\left(\sup_{n \geq t} \left\| \sum_{\tau=t}^{n} \gamma_\tau \xi_\tau \right\|_2 \geq \epsilon \right) = 0.$$

Then the Kushner-Clark Lemma says the following.

Theorem D.2. Under Assumption D.1, suppose that the ODE (D.2) has a compact set $\mathcal{K}^*$ as its set of asymptotically stable equilibria. Then x_t in (D.1) converges almost surely to $\mathcal{K}^*$ as $t \to \infty$.