Neural Temporal-Difference Learning Converges to Global Optima

Qi Cai ∗ Zhuoran Yang † Jason D. Lee ‡ Zhaoran Wang ∗

Abstract

Temporal-difference learning (TD), coupled with neural networks, is among the most fundamental building blocks of deep reinforcement learning. However, due to the nonlinearity in value function approximation, such a coupling leads to nonconvexity and even divergence in optimization. As a result, the global convergence of neural TD remains unclear. In this paper, we prove for the first time that neural TD converges at a sublinear rate to the global optimum of the mean-squared projected Bellman error for policy evaluation. In particular, we show how such global convergence is enabled by the overparametrization of neural networks, which also plays a vital role in the empirical success of neural TD. Beyond policy evaluation, we establish the global convergence of neural (soft) Q-learning, which is further connected to that of policy gradient algorithms.

1 Introduction

Given a policy, temporal-difference learning (TD) (Sutton, 1988) aims to learn the corresponding (action-)value function by following the semigradients of the mean-squared Bellman error in an online manner. As the most-used policy evaluation algorithm, TD serves as the “critic” component of many reinforcement learning algorithms, such as the actor-critic algorithm (Konda and Tsitsiklis, 2000) and trust-region policy optimization (Schulman et al., 2015). In particular, in deep reinforcement learning, TD is often applied to learn value functions parametrized by neural networks (Lillicrap et al., 2015; Mnih et al., 2016; Haarnoja et al., 2018), which gives rise to neural TD. As policy improvement relies crucially on policy evaluation, the optimization efficiency and statistical accuracy of neural TD are critical to the performance of deep reinforcement learning. Towards theoretically understanding deep reinforcement learning, the goal of this paper is to characterize the convergence of neural TD.

Despite the broad applications of neural TD, its convergence remains rarely understood. Even with linear value function approximation, the nonasymptotic convergence of TD remains open until recently (Bhandari et al., 2018; Lakshminarayanan and Szepesvari, 2018; Dalal et al., 2018; Srikant and Ying, 2019), although its asymptotic convergence is well understood (Jaakkola et al., 1994; Tsitsiklis and Van Roy, 1997; Borkar and Meyn, 2000; Kushner and Yin, 2003; Borkar, 2009). Meanwhile, with nonlinear value function approximation, TD is known to diverge in general (Baird, 1995; Boyan and Moore, 1995; Tsitsiklis and Van Roy, 1997). To remedy this issue, Bhatnagar et al. (2009) propose nonlinear (gradient) TD, which uses the tangent

*Northwestern University
†Princeton University
‡University of Southern California
vectors of nonlinear value functions in place of the feature vectors in linear TD. Unlike linear TD, which converges to the global optimum of the mean-squared projected Bellman error (MSPBE), nonlinear TD is only guaranteed to converge to a local optimum asymptotically. As a result, the statistical accuracy of the value function learned by nonlinear TD remains unclear. In contrast to such conservative theory, neural TD, which straightforwardly combines TD with neural networks without the explicit local linearization in nonlinear TD, often learns a desired value function that generalizes well to unseen states in practice (Duan et al., 2016; Amiranashvili et al., 2018; Henderson et al., 2018). Hence, a gap separates theory from practice.

There exist three obstacles towards closing such a theory-practice gap: (i) MSPBE has an expectation over the transition dynamics within the squared loss, which forbids the construction of unbiased stochastic gradients (Sutton and Barto, 2018). As a result, even with linear value function approximation, TD largely eludes the classical optimization framework, as it follows biased stochastic semigradients. (ii) When the value function is parametrized by a neural network, MSPBE is nonconvex in the weights of the neural network, which may introduce undesired stationary points such as local optima and saddle points (Jain and Kar, 2017). As a result, even an ideal algorithm that follows the population gradients of MSPBE may get trapped. (iii) Due to the interplay between the bias in stochastic semigradients and the nonlinearity in value function approximation, neural TD may even diverge (Baird, 1995; Boyan and Moore, 1995; Tsitsiklis and Van Roy, 1997), instead of converging to an undesired stationary point, as it lacks the explicit local linearization in nonlinear TD (Bhatnagar et al., 2009). Such divergence is also not captured by the classical optimization framework.

**Contribution.** Towards bridging theory and practice, we establish the first nonasymptotic global rate of convergence of neural TD. In detail, we prove that randomly initialized neural TD converges to the global optimum of MSPBE at the rate of $1/T$ with population semigradients and at the rate of $1/\sqrt{T}$ with stochastic semigradients. Here $T$ is the number of iterations and the (action-)value function is parametrized by a sufficiently wide two-layer neural network. Moreover, we prove that the projection in MSPBE allows for a sufficiently rich class of functions, which has the same representation power of a reproducing kernel Hilbert space associated with the random initialization. As a result, for a broad class of reinforcement learning problems, neural TD attains zero MSPBE. Beyond policy evaluation, we further establish the global convergence of neural (soft) Q-learning, which allows for policy improvement. In particular, we prove that, under stronger regularity conditions, neural (soft) Q-learning converges at the same rate of neural TD to the global optimum of MSPBE for policy optimization. Also, by exploiting the connection between (soft) Q-learning and policy gradient algorithms (Schulman et al., 2017; Haarnoja et al., 2018), we establish the global convergence of a variant of the policy gradient algorithm (Williams, 1992; Szepesvári, 2010; Sutton and Barto, 2018).

At the core of our analysis is the overparametrization of the two-layer neural network for value function approximation (Zhang et al., 2016; Neyshabur et al., 2018; Allen-Zhu et al., 2018; Arora et al., 2019), which enables us to circumvent the three obstacles above. In particular, overparametrization leads to an implicit local linearization that varies smoothly along the solution path, which mirrors the explicit one in nonlinear TD (Bhatnagar et al., 2009). Such an implicit local linearization enables us to circumvent the third obstacle of possible divergence. Moreover, overparametrization allows us to establish a notion of one-point monotonicity (Harker and Pang, 1990; Facchinei and Pang, 2007) for the semigradients followed by neural TD, which ensures its evolution towards the global optimum of MSPBE along the solution path. Such a notion of monotonicity enables us to circumvent the first and second obstacles of bias and nonconvexity. Broadly speaking, our theory backs the empirical success of overparametrized neural networks in deep reinforcement learning. In particular, we show that instead of being a curse, overparametrization is indeed a blessing for minimizing MSPBE in the presence of bias, nonconvexity, and even divergence.
More Related Work. There is a large body of literature on the convergence of linear TD under both asymptotic (Jaakkola et al., 1994; Tsitsiklis and Van Roy, 1997; Borkar and Meyn, 2000; Kushner and Yin, 2003; Borkar, 2009) and nonasymptotic (Bhandari et al., 2018; Lakshminarayanan and Szepesvari, 2018; Dalal et al., 2018; Srikant and Ying, 2019) regimes. See Dann et al. (2014) for a detailed survey. In particular, our analysis is based on the recent breakthrough in the nonasymptotic analysis of linear TD (Bhandari et al., 2018) and its extension to linear Q-learning (Zou et al., 2019). An essential step of our analysis is bridging the evolution of linear TD and neural TD through the implicit local linearization induced by overparametrization.

To incorporate nonlinear value function approximation into TD, Bhatnagar et al. (2009) propose the first convergent nonlinear TD based on explicit local linearization, which however only converges to a local optimum of MSPBE. See Geist and Pietquin (2013); Bertsekas (2019) for a detailed survey. In contrast, we prove that, with the implicit local linearization induced by overparametrization, neural TD, which is simpler to implement and more widely used in deep reinforcement learning than nonlinear TD, provably converges to the global optimum of MSPBE.

There exist various extensions of TD, including least-squares TD (Bhatnagar et al., 2009) and gradient TD (Sutton et al., 2009a, b; Bhatnagar et al., 2009; Liu et al., 2015; Du et al., 2017; Wang et al., 2017; Touati et al., 2017). In detail, least-squares TD is based on batch update, which loses the computational and statistical efficiency of the online update in TD. Meanwhile, gradient TD follows unbiased stochastic gradients, but at the cost of introducing another optimization variable. Such a reformulation leads to bilevel optimization, which is less stable in practice when combined with neural networks (Pfau and Vinyals, 2016). As a result, both extensions of TD are less widely used in deep reinforcement learning (Duan et al., 2016; Amiranashvili et al., 2018; Henderson et al., 2018). Moreover, when using neural networks for value function approximation, the convergence to the global optimum of MSPBE remains unclear for both extensions of TD.

Our work is also related to the recent breakthrough in understanding overparametrized neural networks, especially their generalization error (Zhang et al., 2016; Neyshabur et al., 2018; Allen-Zhu et al., 2018; Arora et al., 2019). See Fan et al. (2019) for a detailed survey. In particular, Daniely (2017); Allen-Zhu et al. (2018); Arora et al. (2019); Chizat and Bach (2018); Jacot et al. (2018); Lee et al. (2019) characterize the implicit local linearization in the context of supervised learning, where we train an overparametrized neural network by following the stochastic gradients of the mean-squared error. In contrast, neural TD does not follow the stochastic gradients of any objective function, hence leading to possible divergence, which makes the convergence analysis more challenging.

2 Background

In Section 2.1, we briefly review policy evaluation in reinforcement learning. In Section 2.2, we introduce the corresponding optimization formulations.

2.1 Policy Evaluation

We consider a Markov decision process $\langle S, A, P, r, \gamma \rangle$, in which an agent interacts with the environment to learn the optimal policy that maximizes the expected total reward. At the $t$-th time step, the agent has a state $s_t \in S$ and takes an action $a_t \in A$. Upon taking the action, the agent enters the next state $s_{t+1} \in S$ according to the transition probability $P(\cdot | s_t, a_t)$ and receives a random reward $r_t = r(s_t, a_t)$ from the environment. The action that the agent takes at each state is decided by a policy $\pi: S \rightarrow \Delta$, where $\Delta$ is the
set of all probability distributions over \( \mathcal{A} \). The performance of policy \( \pi \) is measured by the expected total reward, \( J(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t \mid a_t \sim \pi(s_t)] \), where \( \gamma < 1 \) is the discount factor.

Given policy \( \pi \), policy evaluation aims to learn the following two functions, the value function \( V^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a, a_t \sim \pi(s_t)] \) and the action-value function (Q-function) \( Q^\pi(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a, a_t \sim \pi(s_t)] \). Both functions form the basis for policy improvement. Without loss of generality, we focus on learning the Q-function in this paper. We define the Bellman evaluation operator,

\[
\mathcal{T}^\pi Q(s, a) = \mathbb{E}[r(s, a) + \gamma Q(s', a') \mid s' \sim \mathcal{P}(\cdot \mid s, a), a' \sim \pi(s')],
\]

for which \( Q^\pi \) is the fixed point, that is, the solution to the Bellman equation \( Q = \mathcal{T}^\pi Q \).

### 2.2 Optimization Formulation

Corresponding to (2.1), we aim to learn \( Q^\pi \) by minimizing the mean-squared Bellman error (MSBE),

\[
\min_{\theta} \text{MSBE}(\theta) = \mathbb{E}_{(s, a) \sim \mu} \left[ (\hat{Q}_\theta(s, a) - \mathcal{T}^\pi \hat{Q}_\theta(s, a))^2 \right],
\]

where the Q-function is parametrized as \( \hat{Q}_\theta \) with parameter \( \theta \). Here \( \mu \) is the stationary distribution of \((s, a)\) corresponding to policy \( \pi \). Due to Q-function approximation, we focus on minimizing the following surrogate of MSBE, namely the projected mean-squared Bellman error (MSPBE),

\[
\min_{\theta} \text{MSPBE}(\theta) = \mathbb{E}_{(s, a) \sim \mu} \left[ (\hat{Q}_\theta(s, a) - \Pi_{\mathcal{F}} \mathcal{T}^\pi \hat{Q}_\theta(s, a))^2 \right].
\]

Here \( \Pi_{\mathcal{F}} \) is the projection onto a function class \( \mathcal{F} \). For example, for linear Q-function approximation (Sutton, 1988), \( \mathcal{F} \) takes the form \( \{ \hat{Q}_{\theta'} : \theta' \in \Theta \} \), where \( \hat{Q}_{\theta'} \) is linear in \( \theta' \) and \( \Theta \) is the set of feasible parameters. As another example, for nonlinear Q-function approximation (Bhatnagar et al., 2009), \( \mathcal{F} \) takes the form \( \{ \hat{Q}_\theta + \nabla_\theta \hat{Q}_\theta^\top (\theta' - \theta) : \theta' \in \Theta \} \), which consists of the local linearization of \( \hat{Q}_{\theta'} \) at \( \theta \).

Throughout this paper, we assume that we are able to sample tuples in the form of \((s, a, r, s', a')\) from the stationary distribution of policy \( \pi \) in an independent and identically distributed manner, although our analysis can be extended to handle temporal dependence using the proof techniques of Bhandari et al. (2018). With a slight abuse of notation, we use \( \mu \) to denote the stationary distribution of \((s, a, r, s', a')\) corresponding to policy \( \pi \) and any of its marginal distributions.

### 3 Neural Temporal-Difference Learning

TD updates the parameter \( \theta \) of the Q-function by taking the stochastic semigradient descent step (Sutton, 1988; Szepesvári, 2010; Sutton and Barto, 2018),

\[
\theta' \leftarrow \theta - \eta \cdot (\hat{Q}_\theta(s, a) - r(s, a) - \gamma \hat{Q}_\theta(s', a')) \cdot \nabla_\theta \hat{Q}_\theta(s, a),
\]

which corresponds to the MSBE in (2.2). Here \((s, a, r, s', a') \sim \mu \) and \( \eta > 0 \) is the stepsize. In a more general context, (3.1) is referred to as TD(0). In this paper, we focus on TD(0), which is abbreviated as TD, and leave the extension to TD(\( \lambda \)) to future work.

In the sequel, we denote the state-action pair \((s, a) \in \mathcal{S} \times \mathcal{A} \) by a vector \( x \in \mathcal{X} \subseteq \mathbb{R}^d \) with \( d > 2 \). We consider \( \mathcal{S} \) to be continuous and \( \mathcal{A} \) to be finite. Without loss of generality, we assume that \( \|x\|_2 = 1 \) and \( |r(x)| \) is upper bounded by a constant \( \tau > 0 \) for any \( x \in \mathcal{X} \). We use a two-layer neural network

\[
\hat{Q}(x; W) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} b_r \sigma(W_r^\top x)
\]
to parametrize the Q-function. Here $\sigma$ is the rectified linear unit (ReLU) activation function $\sigma(y) = \max\{0, y\}$ and the parameter $\theta = (b_1, \ldots, b_m, W_1, \ldots, W_m)$ are initialized as $b_r \sim \text{Unif}\{-1, 1\}$ and $W_r \sim N(0, I_d/d)$ for any $r \in [m]$ independently. During training, we only update $W = (W_1, \ldots, W_m) \in \mathbb{R}^{md}$, while keeping $b = (b_1, \ldots, b_m) \in \mathbb{R}^m$ fixed as the random initialization. To ensure global convergence, we incorporate an additional projection step with respect to $W$. See Algorithm 1 for a detailed description.

**Algorithm 1** Neural TD

1: **Initialization:** $b_r \sim \text{Unif}\{-1, 1\}$, $W_r(0) \sim N(0, I_d/d)$ ($r \in [m]$), $W = W(0)$, $S_B = \{W \in \mathbb{R}^{md} : \|W - W(0)\|_2 \leq B\}$ ($B > 0$)
2: **For** $t = 0$ to $T - 2$:
3: Sample a tuple $(s, a, r, s', a')$ from the stationary distribution $\mu$ of policy $\pi$
4: Let $x = (s, a)$, $x' = (s', a')$
5: Bellman residual calculation: $\delta \leftarrow \hat{Q}(x; W(t)) - r - \gamma \hat{Q}(x'; W(t))$
6: TD update: $\hat{W}(t + 1) \leftarrow W(t) - \eta \delta \cdot \nabla_W \hat{Q}(x; W(t))$
7: Projection: $W(t + 1) \leftarrow \text{argmin}_{W \in S_B} \|W - \hat{W}(t + 1)\|_2$
8: Averaging: $\overline{W} \leftarrow \frac{t + 1}{t + 2} \cdot \overline{W} + \frac{1}{t + 2} \cdot W(t + 1)$
9: **End For**
10: **Output:** $\hat{Q}_{\text{out}}(\cdot) \leftarrow \hat{Q}(\cdot; \overline{W})$

To understand the intuition behind the global convergence of neural TD, note that for the TD update in (3.1), we have from (2.1) that

$$
E_{(s, a, r, s', a') \sim \mu} [(\hat{Q}_\theta(s, a) - r(s, a) - \gamma \hat{Q}_\theta(s', a')) \cdot \nabla_\theta \hat{Q}_\theta(s, a)]
= E_{(s, a) \sim \mu} [(\hat{Q}_\theta(s, a) - E[r(s, a) + \gamma Q(s', a') | s' \sim P(\cdot | s, a), a' \sim \pi(s')] \cdot \nabla_\theta \hat{Q}_\theta(s, a)]
= E_{(s, a) \sim \mu} [(\hat{Q}_\theta(s, a) - T^\pi \hat{Q}_\theta(s, a)) \cdot \nabla_\theta \hat{Q}_\theta(s, a)].
$$

Here (i) is the Bellman residual at $(s, a)$, while (ii) is the gradient of the first term in (i). Although the TD update in (3.1) resembles the stochastic gradient descent step for minimizing a mean-squared error, it is not an unbiased stochastic gradient of any objective function. However, we show that the TD update yields a descent direction towards the global optimum of the MSPBE in (2.3). Moreover, as the neural network becomes wider, the function class $\mathcal{F}$ that $\Pi_F$ projects onto in (2.3) becomes richer. Correspondingly, the MSPBE reduces to the MSBE in (2.2) as the projection becomes closer to identity, which implies the recovery of the desired Q-function $Q^\pi$ such that $Q^\pi = T^\pi \hat{Q}^\pi$. See Section 4 for a more rigorous characterization.

## 4 Main Results

In Section 4.1, we characterize the global optimality of the stationary point attained by Algorithm 1 in terms of minimizing the MSPBE in (2.3) and its other properties. In Section 4.2, we establish the nonasymptotic global rates of convergence of neural TD to the global optimum of the MSPBE when following the population semigradients in (3.3) and the stochastic semigradients in (3.1), respectively.

We use the subscript $E_\mu[\cdot]$ to denote the expectation over the randomness of the tuple $(s, a, r, s, a')$ (or its concise form $(x, r, x')$) conditional on all other randomness, e.g., the random initialization and the
random current iterate. Meanwhile, we use the subscript $\mathbb{E}_{\text{init}, \mu} [\cdot]$ when we are taking the expectation over all randomness, including the random initialization.

### 4.1 Properties of Stationary Point

We consider the population version of the TD update in Line 6 of Algorithm 1.

$$
\hat{W}(t + 1) \leftarrow W(t) - \eta \cdot \mathbb{E}_\mu [\delta(x, r, x'; W(t)) \cdot \nabla_W \hat{Q}(x; W(t))],
$$

(4.1)

where $\mu$ is the stationary distribution and $\delta(x, r, x'; W(t)) = \hat{Q}(x; W(t)) - r - \gamma \hat{Q}(x'; W(t))$ is the Bellman residual at $(x, r, x')$. The stationary point $W^\dagger$ of (4.1) satisfies the following stationarity condition,

$$
\mathbb{E}_\mu [\delta(x, r, x'; W^\dagger) \cdot \nabla_W \hat{Q}(x; W^\dagger)]^\top (W - W^\dagger) \geq 0, \text{ for any } W \in S_B.
$$

(4.2)

Also, note that

$$
\hat{Q}(x; W) = \frac{1}{\sqrt{m}} \sum_{r=1}^m b_r \sigma(W_r^\top x) = \frac{1}{\sqrt{m}} \sum_{r=1}^m b_r \mathbb{I} \{W_r^\top x > 0\} W_r^\top x
$$

and $\nabla_W \hat{Q}(x; W) = b_r \mathbb{I} \{W_r^\top x > 0\} x$ almost everywhere in $\mathbb{R}^{md}$. Meanwhile, recall that $S_B = \{W \in \mathbb{R}^{md} : \|W - W(0)\|_2 \leq B\}$. We define the function class

$$
\mathcal{F}_{B,m}^\dagger = \left\{ \frac{1}{\sqrt{m}} \sum_{r=1}^m b_r \mathbb{I} \{W_r^\top x > 0\} W_r^\top x : W \in S_B \right\},
$$

(4.3)

which consists of the local linearization of $\hat{Q}(x; W)$ at $W = W^\dagger$. Then (4.2) takes the following equivalent form

$$
\langle \hat{Q}(\cdot ; W^\dagger) - \mathcal{T}^\pi \hat{Q}(\cdot ; W^\dagger), f(\cdot) - \hat{Q}(\cdot ; W^\dagger) \rangle_\mu \geq 0, \text{ for any } f \in \mathcal{F}_{B,m}^\dagger,
$$

(4.4)

which implies $\hat{Q}(\cdot ; W^\dagger) = \Pi_{\mathcal{F}_{B,m}^\dagger} \mathcal{T}^\pi \hat{Q}(\cdot ; W^\dagger)$ by the definition of the projection induced by $\langle \cdot, \cdot \rangle_\mu$. By (2.3), $\hat{Q}(\cdot ; W^\dagger)$ is the global optimum of the MSPBE that corresponds to the projection onto $\mathcal{F}_{B,m}^\dagger$.

Intuitively, when using an overparametrized neural network with width $m \to \infty$, the average variation in each $W_r$ diminishes to zero. Hence, roughly speaking, we have $\mathbb{I} \{W_r(t)^\top x > 0\} = \mathbb{I} \{W_r(0)^\top x > 0\}$ with high probability for any $t \in [T]$. As a result, the function class $\mathcal{F}_{B,m}^\dagger$ defined in (4.3) approximates

$$
\mathcal{F}_{B,m} = \left\{ \frac{1}{\sqrt{m}} \sum_{r=1}^m b_r \mathbb{I} \{W_r(0)^\top x > 0\} W_r^\top x : W \in S_B \right\}.
$$

(4.5)

In the sequel, we show that, to characterize the global convergence of Algorithm 1 with a sufficiently large $m$, it suffices to consider $\mathcal{F}_{B,m}$ in place of $\mathcal{F}_{B,m}^\dagger$, which simplifies the analysis, since the distribution of $W(0)$ is given. To this end, we define the approximate stationary point $W^*$ with respect to the function class $\mathcal{F}_{B,m}$ defined in (4.5).

**Definition 4.1 (Approximate Stationary Point $W^*$).** If $W^* = (W^*_1, \ldots, W^*_m) \in \mathbb{R}^{md}$ satisfies

$$
\mathbb{E}_\mu [\delta_0(x, r, x'; W^*) \cdot \nabla_W \hat{Q}_0(x; W^*)]^\top (W - W^*) \geq 0, \text{ for any } W \in S_B,
$$

(4.6)
where we define

\[
\hat{Q}_0(x; W) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} b_r \{ W_r(0)^\top x > 0 \} W_r^\top x, \tag{4.7}
\]

\[
\delta_0(x, r, x'; W) = \hat{Q}_0(x; W) - r - \gamma \hat{Q}_0(x'; W), \tag{4.8}
\]

then we say that $W^*$ is an approximate stationary point of the population update in (4.1). Here $W^*$ depends on the random initialization $b = (b_1, \ldots, b_m)$ and $W(0) = (W_1(0), \ldots, W_m(0))$.

The next lemma proves that such an approximate stationary point uniquely exists, since it is the fixed point of the operator $\Pi_{\mathcal{F}_{B,m}^T \pi}$, which is a contraction in the $\ell_2$-norm associated with the stationary distribution $\mu$.

**Lemma 4.2 (Existence, Uniqueness, and Optimality of $W^*$).** There exists a unique approximate stationary point $W^*$ for any $b \in \mathbb{R}^m$ and $W(0) \in \mathbb{R}^{md}$. Also, $\hat{Q}_0(\cdot; W^*)$ is the global optimum of the MSPBE that corresponds to the projection onto $\mathcal{F}_{B,m}$ in (4.5).

**Proof.** See Appendix B.1 for a detailed proof.

### 4.2 Global Convergence

In this section, we establish the main results on the global convergence of neural TD in Algorithm 1. We first lay out the following regularity condition on the stationary distribution $\mu$.

**Assumption 4.3 (Regularity of Stationary Distribution $\mu$).** There exists a constant $c_0 > 0$ such that for any $\tau \geq 0$ and $w \sim N(0, I_{d/d})$, it holds almost surely that

\[
\mathbb{E}_\mu[\mathbb{I}\{|w^\top x| \leq \tau\} \mid w] \leq c_0 \cdot \tau / \|w\|_2. \tag{4.9}
\]

Assumption 4.3 regularizes the density of $\mu$ in terms of the marginal distribution of $x$. In particular, it is straightforwardly implied when the density of $\mu$ in terms of state $s$ is upper bounded.

**Population Update:** The next theorem establishes the nonasymptotic global rate of convergence of neural TD when it follows population semigradients. Recall that the approximate stationary point $W^*$ and $\hat{Q}_0(\cdot; W^*)$ are defined in Definition 4.1. Also, $B$ is the radius of the set of feasible $W$, which is defined in Algorithm 1, $T$ is the number of iterations, $\gamma$ is the discount factor, and $m$ is the width of the neural network in (3.2).

**Theorem 4.4 (Convergence of Population Update).** We set $\eta = (1 - \gamma)/8$ in Algorithm 1 and replace the TD update in Line 6 by the population update in (4.1). Under Assumption 4.3, the output $\hat{Q}_{\text{out}}$ of Algorithm 1 satisfies

\[
\mathbb{E}_{\text{init}, \mu}\left[\left(\hat{Q}_{\text{out}}(x) - \hat{Q}_0(x; W^*)\right)^2\right] \leq \frac{16B^2}{(1 - \gamma)^2 T} + O(B^3 m^{-1/2} + B^{5/2} m^{-1/4}),
\]

where the expectation is taken with respect to all randomness, including the random initialization and the stationary distribution $\mu$.

**Proof.** The key to the proof of Theorem 4.4 is the one-point monotonicity of the population semigradient $\bar{g}(t)$, which is established through the local linearization $\hat{Q}_0(x; W)$ of $\hat{Q}(x; W)$. See Appendix C.5 for a detailed proof.
**Stochastic Update:** To further prove the global convergence of neural TD when it follows stochastic semigradients, we first establish an upper bound of their variance, which affects the choice of the stepsize $\eta$. For notational simplicity, we define the stochastic and population semigradients as
\[
g(t) = \delta(x, r; x', W(t)) \cdot \nabla_W Q(x; W(t)), \quad \overline{g}(t) = \mathbb{E}_\mu [g(t)].
\]

**Lemma 4.5 (Variance Bound).** There exists $\sigma_\overline{g}^2 = O(B^2)$ such that the variance of the stochastic semigradient is upper bounded as $\mathbb{E}_{\text{init, } \mu} [\|g(t) - \overline{g}(t)\|_2^2] \leq \sigma_\overline{g}^2$ for any $t \in [T]$.

**Proof.** See Appendix B.2 for a detailed proof.

Based on Theorem 4.4 and Lemma 4.5, we establish the global convergence of neural TD in Algorithm 1.

**Theorem 4.6 (Convergence of Stochastic Update).** We set $\eta = \min\{(1 - \gamma)/8, 1/\sqrt{T}\}$ in Algorithm 1. Under Assumption 4.3, the output $\hat{Q}_{\text{out}}$ of Algorithm 1 satisfies
\[
\mathbb{E}_{\text{init, } \mu} \left[ \left( \hat{Q}_{\text{out}}(x) - \hat{Q}_0(x; W^*) \right)^2 \right] \leq \frac{16(B^2 + \sigma_\overline{g}^2)}{(1 - \gamma)^2 \sqrt{T}} + O(B^3 m^{-1/2} + B^{5/2} m^{-1/4}).
\]

**Proof.** See Appendix C.6 for a detailed proof.

As the width of the neural network $m \to \infty$, Lemma 4.2 implies that $\hat{Q}_0(\cdot; W^*)$ is the global optimum of the MSPBE in (2.3) with a richer function class $\mathcal{F}_{B, \infty}$ to project onto. In fact, the function class $\mathcal{F}_{B, \infty} - \hat{Q}(\cdot; W(0))$ is a subset of an RKHS with $\mathcal{H}$-norm upper bounded by $B$. Here $\hat{Q}(\cdot; W(0))$ is defined in (3.2). See Appendix A.2 for a more detailed discussion on the representation power of $\mathcal{F}_{B, \infty}$. Therefore, if the desired Q-function $Q^\pi(\cdot)$ falls into $\mathcal{F}_{B, \infty}$, it is the global optimum of the MSPBE. By Lemma 4.2 and Theorem 4.6, we approximately obtain $Q^\pi(\cdot) = \hat{Q}_0(\cdot; W^*)$ through $\hat{Q}_{\text{out}}(\cdot)$.

More generally, the following proposition quantifies the distance between $\hat{Q}_0(\cdot; W^*)$ and $Q^\pi(\cdot)$ in the case that $Q^\pi(\cdot)$ does not fall into the function class $\mathcal{F}_{B, m}$. In particular, it states that the $\ell_2$-norm distance $\|\hat{Q}_0(\cdot; W^*) - Q^\pi(\cdot)\|_\mu$ is upper bounded by the distance between $Q^\pi(\cdot)$ and $\mathcal{F}_{B, m}$.

**Proposition 4.7 (Convergence of Stochastic Update to $Q^\pi$).** It holds that $\|\hat{Q}_0(\cdot; W^*) - Q^\pi(\cdot)\|_\mu \leq (1 - \gamma)^{-1} \cdot \|\Pi_{\mathcal{F}_{B, m}} Q^\pi(\cdot) - Q^\pi(\cdot)\|_\mu$, which by Theorem 4.6 implies
\[
\mathbb{E}_{\text{init, } \mu} \left[ \left( \hat{Q}_{\text{out}}(x) - Q^\pi(x) \right)^2 \right] \leq \frac{32(B^2 + \sigma_\overline{g}^2)}{(1 - \gamma)^2 \sqrt{T}} + 2\mathbb{E}_{\text{init, } \mu} \left[ \left( \Pi_{\mathcal{F}_{B, m}} Q^\pi(x) - Q^\pi(x) \right)^2 \right] + O(B^3 m^{-1/2} + B^{5/2} m^{-1/4}).
\]

**Proof.** See Appendix B.3 for a detailed proof.

Proposition 4.7 implies that if $Q^\pi(\cdot) \in \mathcal{F}_{B, \infty}$, then $\hat{Q}_{\text{out}}(\cdot) \to Q^\pi(\cdot)$ as $T, m \to \infty$. In other words, neural TD converges to the global optimum of the MSPBE in (2.3), or equivalently, the MSBE in (2.2), both of which have objective value zero.

## 5 Proof Sketch

In the sequel, we sketch the proofs of Theorems 4.4 and 4.6 in Section 4.
5.1 Implicit Local Linearization via Overparametrization

Recall that as defined in (4.7), \( \hat{Q}_0(x; W) \) takes the form

\[
\hat{Q}_0(x; W) = \Phi(x)^\top W,
\]

where \( \Phi(x) = \frac{1}{\sqrt{m}} \cdot (1 \{W_1(0)^\top x > 0\} x, \ldots, 1 \{W_m(0)^\top x > 0\} x) \in \mathbb{R}^{md}, \)

which is linear in the feature map \( \Phi(x) \). In other words, with respect to \( W \), \( \hat{Q}_0(x; W) \) linearizes the neural network \( \hat{Q}(x; W) \) defined in (3.2) locally at \( W(0) \). The following lemma characterizes the difference between \( \hat{Q}(x; W(t)) \), which is along the solution path of neural TD in Algorithm 1, and its local linearization \( \hat{Q}_0(x; W(t)) \). In particular, we show that the error of such a local linearization diminishes to zero as \( m \to \infty \). For notational simplicity, we use \( \hat{Q}_t(x) \) to denote \( \hat{Q}(x; W(t)) \) in the sequel. Note that by (4.7) we have \( Q_0(x) = Q(x; W(0)) = \hat{Q}_0(x; W(0)) \). Recall that \( B \) is the radius of the set of feasible \( W \) in (4.5).

Lemma 5.1 (Local Linearization of Q-Function). There exists a constant \( c_1 > 0 \) such that for any \( t \in [T] \), it holds that

\[
\mathbb{E}_{\text{init}, \mu} \left[ \left| \hat{Q}_t(x) - \hat{Q}_0(x; W(t)) \right|^2 \right] \leq 4c_1 B^3 \cdot m^{-1/2}.
\]

Proof. See Appendix C.1 for a detailed proof.

As a direct consequence of Lemma 5.1, the next lemma characterizes the effect of local linearization on population semigradients. Recall that \( \overline{\delta}(t) \) is defined in (4.10). We denote by \( \overline{\delta}_0(t) \) the locally linearized population semigradient, which is defined by replacing \( \hat{Q}_t(x) \) in \( \overline{\delta}(t) \) with its local linearization \( \hat{Q}_0(x; W(t)) \). In other words, by (4.10), (4.7), and (4.8), we have

\[
\overline{\delta}(t) = \mathbb{E}_{\mu} \left[ \delta(x, r, x'; W(t)) \cdot \nabla_W \hat{Q}(x; W(t)) \right],
\]

\[
\overline{\delta}_0(t) = \mathbb{E}_{\mu} \left[ \delta_0(x, r, x'; W(t)) \cdot \nabla_W \hat{Q}_0(x; W(t)) \right].
\]

Lemma 5.2 (Local Linearization of Semigradient). Let \( \overline{\delta} \) be the upper bound of the reward \( r(x) \) for any \( x \in \mathcal{X} \). There exists a constant \( c_2 > 0 \) such that for any \( t \in [T] \), it holds that

\[
\mathbb{E}_{\text{init}} \left[ \left\| \overline{\delta}(t) - \overline{\delta}_0(t) \right\|_2^2 \right] \leq (56c_1 B^3 + 24c_2 B + 6c_1 B^2 \sigma^2) \cdot m^{-1/2}.
\]

Proof. See Appendix C.2 for a detailed proof.

Lemmas 5.1 and 5.2 show that the error of local linearization diminishes as the degree of overparametrization increases along \( m \). As a result, we do not require the explicit local linearization in nonlinear TD (Bhatnagar et al., 2009). Instead, we show that such an implicit local linearization suffices to ensure the global convergence of neural TD.

5.2 Proofs for Population Update

The characterization of the locally linearized Q-function in Lemma 5.1 and the locally linearized population semigradients in Lemma 5.2 allows us to establish the following descent lemma, which extends Lemma 3 of Bhandari et al. (2018) for characterizing linear TD.
Lemma 5.3 (Population Descent Lemma). For \( \{W(t)\}_{t \in [T]} \) in Algorithm 1 with the TD update in Line 6 replaced by the population update in (4.1), it holds that

\[
\|W(t + 1) - W^*\|_2^2 \leq \|W(t) - W^*\|_2^2 - (2\eta(1 - \gamma) - 8\eta^2) \cdot \mathbb{E}_\mu \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right] + 2\eta^2 \cdot \|\overline{g}(t) - \overline{g}_0(t)\|_2^2 + 2\eta B \cdot \|\overline{g}(t) - \overline{g}_0(t)\|_2.
\]

**Proof.** See Appendix C.3 for a detailed proof. \( \square \)

Lemma 5.3 shows that, with a sufficiently small stepsize \( \eta \), \( \|W(t) - W^*\|_2 \) decays at each iteration up to the error of local linearization, which is characterized by Lemma 5.2. By combining Lemmas 5.2 and 5.3 and further plugging them into a telescoping sum, we establish the convergence of \( \hat{Q}_0(\cdot; W^*) \) to the global optimum \( \hat{Q}_0(\cdot; W^*) \) of the MSPBE. See Appendix C.5 for a detailed proof.

### 5.3 Proofs for Stochastic Update

Recall that the stochastic semigradient \( g(t) \) is defined in (4.10). In parallel with Lemma 5.3, the following lemma additionally characterizes the effect of the variance of \( g(t) \), which is induced by the randomness of the current tuple \( (x, r, x') \). We use the subscript \( \mathbb{E}_{W}[\cdot] \) to denote the expectation over the randomness of the current iterate \( W(t) \) conditional on the random initialization \( b \) and \( W(0) \). Correspondingly, \( \mathbb{E}_{W,\mu}[\cdot] \) is over the randomness of both the current tuple \( (x, r, x') \) and the current iterate \( W(t) \) conditional on the random initialization.

**Lemma 5.4 (Stochastic Descent Lemma).** For \( \{W(t)\}_{t \in [T]} \) in Algorithm 1, it holds that

\[
\mathbb{E}_{W,\mu}[\|W(t + 1) - W^*\|_2^2] \\
\leq \mathbb{E}_W[\|W(t) - W^*\|_2^2] - (2\eta(1 - \gamma) - 8\eta^2) \cdot \mathbb{E}_{W,\mu} \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right] \\
+ \mathbb{E}_W[2\eta^2 \cdot \|\overline{g}(t) - \overline{g}_0(t)\|_2^2] + 2\eta B \cdot \|\overline{g}(t) - \overline{g}_0(t)\|_2.
\]

**Proof.** See Appendix C.4 for a detailed proof. \( \square \)

To ensure the global convergence of neural TD in the presence of the variance of \( g(t) \), we rescale the stepsize to be of order \( T^{-1/2} \). The rest proof of Theorem 4.6 mirrors that of Theorem 4.4. See Appendix C.6 for a detailed proof.

### 6 Extension to Policy Optimization

With the Q-function learned by TD, policy iteration may be applied to learn the optimal policy. Alternatively, Q-learning more directly learns the optimal policy and its Q-function using temporal-difference update. Compared with TD, Q-learning aims to solve the projected Bellman optimality equation

\[
Q = \Pi_{\mathcal{F}} T Q, \quad \text{with} \quad T Q(s, a) = \mathbb{E} \left[ r(s, a) + \gamma \max_{a' \in A} Q(s', a') \mid s' \sim \mathcal{P}(\cdot \mid s, a) \right],
\]

which replaces the Bellman evaluation operator \( T^\pi \) in (2.3) with the Bellman optimality operator \( T \). When \( \Pi_{\mathcal{F}} \) is identity, the fixed-point solution to (6.1) is the Q-function \( Q^\pi(\pi^*; s, a) \) of the optimal policy \( \pi^* \), which
maximizes the expected total reward (Szepesvári, 2010; Sutton and Barto, 2018). Compared with TD, the max operator in $\mathcal{T}$ makes the analysis more challenging and hence requires stronger regularity conditions. In the following, we first introduce neural Q-learning and then establish its global convergence. Finally, we discuss the corresponding implication for policy gradient algorithms.

### 6.1 Neural Q-Learning

In parallel with (3.1), we update the parameter $\theta$ of the optimal Q-function by

$$
\theta' \leftarrow \theta - \eta \cdot (\hat{Q}_\theta(s, a) - r(s, a) - \gamma \max_{a' \in A} \hat{Q}_\theta(s', a')) \cdot \nabla \hat{Q}_\theta(s, a),
$$

(6.2)

where the tuple $(s, a, r, s')$ is sampled from the stationary distribution $\mu_{\text{exp}}$ of an exploration policy $\pi_{\text{exp}}$ in an independent and identically distributed manner. We present the detailed neural Q-learning algorithm in Algorithm 2. Similar to Definition 4.1, we define the approximate stationary point $W^*$ of Algorithm 2 by

$$
E_{\mu_{\text{exp}}} [\delta_0(x, r, x'; W^*) \cdot \nabla_W \hat{Q}_0(x; W^*)]^\top (W - W^*) \geq 0, \text{ for any } W \in S_B,
$$

(6.3)

where the Bellman residual is now $\delta_0(x, r, x'; W) = \hat{Q}_0(x; W) - r - \gamma \max_{a' \in A} \hat{Q}_0(s', a'; W)$. Following the same analysis of neural TD in Lemma 4.2, we have that $\hat{Q}_0(\cdot; W^*)$ is the unique fixed-point solution to the projected Bellman optimality equation $Q = \Pi_{\mathcal{F}_{B,m}} \mathcal{T}Q$, where the function class $\mathcal{F}_{B,m}$ is define in (4.5).

**Algorithm 2 Neural Q-Learning**

1: **Initialization:** $b_r \sim \text{Unif}([-1, 1]), W_r(0) \sim N(0, I_d/d)$ ($r \in [m]$), $\bar{W} = W(0)$, $S_B = \{W \in \mathbb{R}^{nd} : \|W - W(0)\|_2 \leq B\}$ ($B > 0$) 

   Exploration policy $\pi_{\text{exp}}$ such that $\pi_{\text{exp}}(s) \in \Delta$ has positive density for any $s \in \mathcal{S}$

2: **For** $t = 0$ to $T - 2$:

3: Sample a tuple $(s, a, r, s')$ from the stationary distribution $\mu_{\text{exp}}$ of the exploration policy $\pi_{\text{exp}}$

4: Let $x = (s, a)$, $x' = (s', \arg\max_{a' \in A} \hat{Q}(s', a'; W(t)))$

5: Bellman residual calculation: $\delta \leftarrow \hat{Q}(x; W(t)) - r - \gamma \hat{Q}(x'; W(t))$

6: TD update: $\tilde{W}(t + 1) \leftarrow W(t) - \eta \delta \cdot \nabla_W \hat{Q}(x; W(t))$

7: Projection: $W(t + 1) \leftarrow \arg\min_{W \in S_B} \|W - \tilde{W}(t + 1)\|_2$

8: Averaging: $\bar{W} \leftarrow \frac{t + 2}{t + 1} \cdot \bar{W} + \frac{1}{t + 1} \cdot W(t + 1)$

9: **End For**

10: **Output:** $\hat{Q}_{\text{out}}(\cdot) \leftarrow \hat{Q}(\cdot; \bar{W})$

### 6.2 Global Convergence

To establish the global convergence of neural Q-learning, we lay out an extra regularity condition on the exploration policy $\pi_{\text{exp}}$, which is not required by neural TD. Such a regularity condition ensures that $x' = (s', a')$ with the greedy action $a'$ in Line 4 of Algorithm 2 follows a similar distribution to that of $x = (s, a)$, which is the stationary distribution $\mu_{\text{exp}}$ of the exploration policy $\pi_{\text{exp}}$. Recall that $\hat{Q}_0(x; W)$ is defined in (4.7) and $\gamma$ is the discount factor.
Assumption 6.1 (Regularity of Exploration Policy \( \pi_{\text{exp}} \)). There exists a constant \( \nu > 0 \) such that for any \( W_1, W_2 \in S_B \), it holds that

\[
E_{x \sim \mu_{\text{exp}}} \left[ (\hat{Q}_0(x; W_1) - \hat{Q}_0(x; W_2))^2 \right] \geq (\gamma + \nu)^2 \cdot E_{s \sim \mu_{\text{exp}}} \left[ (\hat{Q}_0^\sharp(s; W_1) - \hat{Q}_0^\sharp(s; W_2))^2 \right],
\]

where \( \hat{Q}_0^\sharp(s; W) = \max_{a \in A} \hat{Q}_0(s, a; W) \).

We remark that Melo et al. (2008); Zou et al. (2019) establish the global convergence of linear Q-learning based on an assumption that implies (6.4). Although Assumption 6.1 is strong, we are not aware of any weaker regularity condition in the literature, even for linear Q-learning. As our focus is to go beyond linear Q-learning to analyze neural Q-learning, we do not attempt to weaken such a regularity condition in this paper.

The following regularity condition on \( \mu_{\text{exp}} \) mirrors Assumption 4.3, but additionally accounts for the max operator in the Bellman optimality operator.

Assumption 6.2 (Regularity of Stationary Distribution \( \mu_{\text{exp}} \)). There exists a constant \( c_3 > 0 \) such that for any \( \tau \geq 0 \) and \( w \sim N(0, I_d/d) \), it holds almost surely that

\[
E_{s \sim \mu_{\text{exp}}} \left[ \max_{a \in A} 1 \left\{ |w^T(s, a)| \leq \tau \right\} \right] \leq c_3 \cdot \tau / \|w\|_2.
\]

In parallel with Theorem 4.6, the following theorem establishes the global convergence of neural Q-learning in Algorithm 2.

Theorem 6.3 (Convergence of Stochastic Update). We set \( \eta \) to be of order \( T^{-1/2} \) in Algorithm 2. Under Assumptions 6.1 and 6.2, the output \( \hat{Q}_{\text{out}} \) of Algorithm 2 satisfies

\[
E_{\text{init}, \mu_{\text{exp}}} \left[ (\hat{Q}_{\text{out}}(x) - \hat{Q}_0(x; W^*))^2 \right] = O(B^2 T^{-1/2} + B^3 m^{-1/2} + B^{5/2} m^{-1/4}).
\]

**Proof.** See Appendix D.1 for a detailed proof.

Corresponding to Proposition 4.7, Theorem 6.3 also implies the convergence to \( Q^{\pi^*}(s, a) \), which is omitted due to space limitations.

### 6.3 Implication for Policy Gradient

Theorem 6.3 can be further extended to handle neural soft Q-learning, where the max operator in the Bellman optimality operator is replaced by a more general softmax operator (Haarnoja et al., 2017; Neu et al., 2017). By exploiting the equivalence between soft Q-learning and policy gradient algorithms (Schulman et al., 2017; Haarnoja et al., 2018), we establish the global convergence of a variant of the policy gradient algorithm. Due to space limitations, we defer the discussion to Appendix E.

### 7 Conclusions

In this paper we prove that neural TD converges at a sublinear rate to the global optimum of the MSPBE for policy evaluation. In particular, we show how such global convergence is enabled by the overparametrization of neural networks. Moreover, we extend the convergence result to policy optimization, including (soft) Q-learning and policy gradient. Our results shed new light on the theoretical understanding of RL with neural networks, which is widely employed in practice.
References

Allen-Zhu, Z., Li, Y. and Liang, Y. (2018). Learning and generalization in overparameterized neural networks, going beyond two layers. arXiv preprint arXiv:1811.04918.

Amiranashvili, A., Dosovitskiy, A., Koltun, V. and Brox, T. (2018). TD or not TD: Analyzing the role of temporal differencing in deep reinforcement learning. arXiv preprint arXiv:1806.01175.

Arora, S., Du, S. S., Hu, W., Li, Z. and Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. arXiv preprint arXiv:1901.08584.

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In International Conference on Machine Learning.

Bertsekas, D. P. (2019). Feature-based aggregation and deep reinforcement learning: A survey and some new implementations. IEEE/CAA Journal of Automatica Sinica, 6 1–31.

Bhandari, J., Russo, D. and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. arXiv preprint arXiv:1806.02450.

Bhatnagar, S., Precup, D., Silver, D., Sutton, R. S., Maei, H. R. and Szepesvári, C. (2009). Convergent temporal-difference learning with arbitrary smooth function approximation. In Advances in Neural Information Processing Systems.

Borkar, V. S. (2009). Stochastic Approximation: A Dynamical Systems Viewpoint, vol. 48. Springer.

Borkar, V. S. and Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. SIAM Journal on Control and Optimization, 38 447–469.

Boyan, J. A. (1999). Least-squares temporal difference learning. In International Conference on Machine Learning.

Boyan, J. A. and Moore, A. W. (1995). Generalization in reinforcement learning: Safely approximating the value function. In Advances in Neural Information Processing Systems.

Bradtke, S. J. and Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. Machine Learning, 22 33–57.

Chizat, L. and Bach, F. (2018). A note on lazy training in supervised differentiable programming. arXiv preprint arXiv:1812.07956.

Dalal, G., Szörsényi, B., Thoppe, G. and Mannor, S. (2018). Finite sample analyses for TD(0) with function approximation. In AAAI Conference on Artificial Intelligence.

Daniely, A. (2017). SGD learns the conjugate kernel class of the network. In Advances in Neural Information Processing Systems.

Dann, C., Neumann, G. and Peters, J. (2014). Policy evaluation with temporal differences: A survey and comparison. Journal of Machine Learning Research, 15 809–883.

Du, S. S., Chen, J., Li, L., Xiao, L. and Zhou, D. (2017). Stochastic variance reduction methods for policy evaluation. In International Conference on Machine Learning.
Duan, Y., Chen, X., Houthooft, R., Schulman, J. and Abbeel, P. (2016). Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*.

Facchinei, F. and Pang, J.-S. (2007). *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Science & Business Media.

Fan, J., Ma, C. and Zhong, Y. (2019). A selective overview of deep learning. *arXiv preprint arXiv:1904.05526*.

Geist, M. and Pietquin, O. (2013). Algorithmic survey of parametric value function approximation. *IEEE Transactions on Neural Networks and Learning Systems*, **24** 845–867.

Ghavamzadeh, M., Lazaric, A., Maillard, O. and Munos, R. (2010). LSTD with random projections. In *Advances in Neural Information Processing Systems*.

Haarnoja, T., Tang, H., Abbeel, P. and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*.

Haarnoja, T., Zhou, A., Abbeel, P. and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*.

Harker, P. T. and Pang, J.-S. (1990). Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical Programming*, **48** 161–220.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D. and Meger, D. (2018). Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*.

Hofmann, T., Schölkopf, B. and Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics* **1171–1220**.

Jaakkola, T., Jordan, M. I. and Singh, S. P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*.

Jacot, A., Gabriel, F. and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*.

Jain, P. and Kar, P. (2017). Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, **10** 142–336.

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in Neural Information Processing Systems*.

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

Lakshminarayanan, C. and Szepesvari, C. (2018). Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*.

Lazaric, A., Ghavamzadeh, M. and Munos, R. (2010). Finite-sample analysis of LSTD. In *International Conference on Machine Learning*.

14
Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J. and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720.*

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

Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S. and Petrik, M. (2015). Finite-sample analysis of proximal gradient TD algorithms. In *Conference on Uncertainty in Artificial Intelligence.*

Melo, F. S., Meyn, S. P. and Ribeiro, M. I. (2015). An analysis of reinforcement learning with function approximation. In *International Conference on Machine Learning.*

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

Neu, G., Jonsson, A. and Gómez, V. (2017). A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798.*

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y. and Srebro, N. (2018). Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076.*

Pfau, D. and Vinyals, O. (2016). Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945.*

Rahimi, A. and Recht, B. (2008a). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems.*

Rahimi, A. and Recht, B. (2008b). Uniform approximation of functions with random bases. In *Annual Allerton Conference on Communication, Control, and Computing.*

Schulman, J., Chen, X. and Abbeel, P. (2017). Equivalence between policy gradients and soft Q-learning. *arXiv preprint arXiv:1704.06440.*

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

Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and TD learning. *arXiv preprint arXiv:1902.00923.*

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning,* 3 9–44.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction.* MIT press.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C. and Wiewiora, E. (2009a). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *International Conference on Machine Learning.*

Sutton, R. S., Maei, H. R. and Szepesvári, C. (2009b). A convergent $o(\sqrt{n})$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems.*
Szepesvári, C. (2010). Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, **4** 1–103.

Touati, A., Bacon, P.-L., Precup, D. and Vincent, P. (2017). Convergent tree-backup and retrace with function approximation. *arXiv preprint arXiv:1705.09322*.

Tsitsiklis, J. N. and Van Roy, B. (1997). Analysis of temporal-difference learning with function approximation. In *Advances in Neural Information Processing Systems*.

Tu, S. and Recht, B. (2017). Least-squares temporal difference learning for the linear quadratic regulator. *arXiv preprint arXiv:1712.08642*.

Wang, Y., Chen, W., Liu, Y., Ma, Z.-M. and Liu, T.-Y. (2017). Finite sample analysis of the GTD policy evaluation algorithms in Markov setting. In *Advances in Neural Information Processing Systems*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, **8** 229–256.

Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Zou, S., Xu, T. and Liang, Y. (2019). Finite-sample analysis for SARSA and Q-learning with linear function approximation. *arXiv preprint arXiv:1902.02234*.
A Representation Power of $\mathcal{F}_{B,m}$

A.1 Background on RKHS

We consider the following kernel function

$$K(x, y) = \int_{\mathcal{W}} \phi(x; w)\phi(y; w)p(w)dw.$$  \hfill (A.1)

Here $\phi$ is a random feature map parametrized by $w$, which follows a distribution with density $p(\cdot)$ (Rahimi and Recht, 2008a). Any function in the RKHS induced by $K(\cdot, \cdot)$ takes the form

$$f_c(x) = \int_{\mathcal{W}} c(w)\phi(x; w)p(w)dw,$$  \hfill (A.2)

such that each $c(\cdot)$ corresponds to a function $f_c(\cdot)$. The following lemma connects the $\mathcal{H}$-norm of $f_c(\cdot)$ to the $\ell_2$-norm of $c(\cdot)$ associated with the density $p(\cdot)$, denoted as $\|c\|_p$.

**Lemma A.1.** It holds that $\|f_c\|^2_{\mathcal{H}} = \|c\|^2_p = \int c(w)^2 p(w)dw$.

**Proof.** Recall if $f(x) = \int_{\mathcal{X}} a(y)K(x, y)dy$, then by the reproducing property (Hofmann et al., 2008), we have

$$\|f\|^2_{\mathcal{H}} = \int_{\mathcal{X} \times \mathcal{X}} a(x)a(y)K(x, y)dxdy.$$  

Now we write $f(\cdot)$ in the form of (A.2). By (A.1), we have

$$f(x) = \int_{\mathcal{X}} a(y)K(x, y)dy$$

$$= \int_{\mathcal{X}} a(y) \int_{\mathcal{W}} \phi(x; w)\phi(y; w)p(w)dwdy$$

$$= \int_{\mathcal{W}} \left( \int_{\mathcal{X}} a(y)\phi(y; w)dy \right) \phi(x; w)p(w)dw.$$  

Thus, for $c(w) = \int_{\mathcal{X}} a(y)\phi(y; w)dy$, we have

$$\|f\|^2_{\mathcal{H}} = \int_{\mathcal{X} \times \mathcal{X}} a(y)a(x)K(x, y)dxdy$$

$$= \int_{\mathcal{X} \times \mathcal{X}} a(y)a(x) \left( \int_{\mathcal{W}} \phi(x; w)\phi(y; w)p(w)dw \right) dxdy$$

$$= \int_{\mathcal{W}} \left( \int_{\mathcal{X}} a(y)\phi(y; w)dy \right) \left( \int_{\mathcal{X}} a(x)\phi(x; w)dx \right) p(w)dw$$

$$= \int_{\mathcal{W}} c(w)^2 p(w)dw = \|c\|^2_p,$$

which completes the proof of Lemma A.1. \hfill \Box
A.2 \( \mathcal{F}_{B,\infty} \) as RKHS

We characterize the approximate stationary point \( \hat{Q}_0(x; W^*) \) defined in Definition 4.1, which is attained by Algorithm 1 according to Theorems 4.4 and 4.6. We focus on its representation power when \( m \to \infty \). We first write \( \mathcal{F}_{B,m} \) in (4.5) as

\[
\mathcal{F}_{B,m} = \left\{ f(x) = \hat{Q}(x; W(0)) + \sum_{r=1}^{m} \phi_r(x)\top (W_r - W_r(0)) : W \in \mathcal{S}_B \right\},
\]

(A.3)

where the feature map \( \{\phi_r(x)\}_{r=1}^{m} \) is defined as

\[
\phi_r(x) = \phi(x; W_r(0)) = \frac{1}{\sqrt{m}} \cdot \mathbb{1}\{W_r(0)\top x > 0\} x \quad \text{for any} \ r \in [m].
\]

As \( m \to \infty \), the empirical distribution supported on \( \{\phi_r(x)\}_{r=1}^{m} \), which has sample size \( m \), converges to the corresponding population distribution. Therefore, from (A.3) we obtain

\[
\mathcal{F}_{B,\infty} = \left\{ f(x) = f_0(x) + \int \phi(x; w)\top \alpha(w) \cdot p(w)dw : \int \|\alpha(w)\|_2^2 \cdot p(w)dw \leq B^2 \right\}.
\]

Here \( p(w) \) is the density of \( N(0, I_d/d) \) and \( f_0(x) = \lim_{m \to \infty} \hat{Q}(x; W(0)) \), which by the central limit theorem is a Gaussian process indexed by \( x \). Furthermore, as discussed in Appendix A.1, \( \phi(x; W) \) induces an RKHS, namely \( \mathcal{H} \), which is the completion of the set of all functions that take the form

\[
f(x) = \sum_{i=1}^{N} a_i K(x, x_i), \quad x_i \in \mathcal{X}, \quad a_i \in \mathbb{R}, \quad N \in \mathbb{N},
\]

where \( K(x, y) = \mathbb{E}_{w \sim N(0, I_d/d)} [\mathbb{1}\{w\top x > 0, w\top y > 0\} x\top y] \).

In particular, \( \mathcal{H} \) is equipped with the inner product induced by \( \langle K(\cdot, x_i), K(\cdot, x_j)\rangle_{\mathcal{H}} = K(x_i, x_j) \). Rahimi and Recht (2008b) prove that, similar to Lemma A.1, for any \( f_1(\cdot) = \int \phi(\cdot; w)\top \alpha_1(w) \cdot p(w)dw \) and \( f_2(\cdot) = \int \phi(\cdot; w)\top \alpha_2(w) \cdot p(w)dw \), we have \( f_1, f_2 \in \mathcal{H} \), and moreover, their inner product has the following equivalence

\[
\langle f_1, f_2 \rangle_{\mathcal{H}} = \int \alpha_1(w)\top \alpha_2(w) \cdot p(w)dw.
\]

As a result, we have

\[
\mathcal{F}_{B,\infty} = \left\{ f = f_0 + h : \|h\|_{\mathcal{H}} \leq B \right\}, \quad \text{(A.4)}
\]

which is known to be a rich function class (Hofmann et al., 2008). As \( m \to \infty \), \( \hat{Q}_0(\cdot; W^*) \) becomes the fixed-point solution to the projected Bellman equation

\[
Q = \Pi_{\mathcal{F}_{B,\infty}} \mathcal{T}^\pi Q,
\]

which also implies that \( \hat{Q}_0(\cdot; W^*) \) is the global optimum of the MSPBE

\[
\mathbb{E}_{\mu}[\{(Q(x) - \Pi_{\mathcal{F}_{B,\infty}} \mathcal{T}^\pi Q(x))^2\}].
\]

18
If we further assume that the Bellman evaluation operator $T^\pi$ satisfies $T^\pi \hat{Q}_0(\cdot; W^*) - f_0(\cdot) \in \mathcal{H}$ and $B$ is sufficiently large such that $\|T^\pi \hat{Q}_0(\cdot; W^*) - f_0(\cdot)\|_\mathcal{H} \leq B$, then the projection $\Pi_{\mathcal{F}_{B,\infty}}$ reduces to identity at $T^\pi \hat{Q}_0(\cdot; W^*)$, which implies $\hat{Q}_0(\cdot; W^*) = Q^\pi(\cdot)$ as they both solve the Bellman equation $Q = T^\pi Q$. In other words, if the Bellman evaluation operator is closed with respect to $\mathcal{F}_{B,\infty}$, which up to the intercept of $f_0(\cdot)$ is a ball with radius $B$ in $\mathcal{H}$, the approximate stationary point $\hat{Q}_0(\cdot; W^*)$ is the unique fixed-point solution to the Bellman equation or equivalently the global optimum of the MSBE

$$
\mathbb{E}_\mu [(Q(x) - T^\pi Q(x))^2].
$$

## B Proofs for Section 4

### B.1 Proof of Lemma 4.2

**Proof.** Following the same argument for $W^\dagger$ in (4.4) and the definition of $W^*$ in (4.6), we know that $\hat{Q}_0(\cdot; W^*)$ is a fixed-point solution to the projected Bellman equation

$$Q = \Pi_{\mathcal{F}_{B,m}} T^\pi Q. \tag{B.1}$$

Meanwhile, the Bellman evaluation operator $T^\pi$ is a $\gamma$-contraction in the $\ell_2$-norm $\|\cdot\|_\mu$ with $\gamma < 1$, since

$$\mathbb{E}_{x \sim \mu} [(T^\pi Q_1(x) - T^\pi Q_2(x))^2] = \gamma^2 \mathbb{E}_{x \sim \mu} [\mathbb{E}(Q_1(x') - Q_2(x') \mid s' \sim \mathcal{P}(\cdot \mid s, a), a' \sim \pi(s'))]^2 \leq \gamma^2 \mathbb{E}_{x \sim \mu} [(Q_1(x) - Q_2(x))^2],$$

where the second equality follows from Hölder’s inequality and the fact that marginally $x'$ and $x$ have the same stationary distribution. Since the projection onto a convex set is nonexpansive, $\Pi_{\mathcal{F}_{B,m}} T^\pi$ is also a $\gamma$-contraction. Thus, the projected Bellman equation in (B.1) has a unique fixed-point solution $\hat{Q}_0(\cdot; W^*)$ in $\mathcal{F}_{B,m}$, which corresponds to the unique approximate stationary point $W^*$. \qed

### B.2 Proof of Lemma 4.5

**Proof.** It suffices to show that $\mathbb{E}_{init,\mu} [\|g(t)\|_2^2]$ is both upper bounded. By (4.10) we have

$$\mathbb{E}_{init,\mu} [\|g(t)\|_2^2] = \mathbb{E}_{init,\mu} [\|\delta(x, r, x'; W(t)) \cdot \nabla_W \hat{Q}_t(x)\|_2^2] \leq \mathbb{E}_{init,\mu} [\|\delta(x, r, x'; W(t))\|_2^2], \tag{B.2}$$

where the inequality follows from the fact that, for any $W \in S_B$,

$$\|\nabla_W \hat{Q}(x; W)\|_2 = \frac{1}{m} \sum_{r=1}^m \mathbb{1}\{W^\top x > 0\} \cdot \|x\|_2 \leq 1 \tag{B.3}$$

almost everywhere. Using the fact that $x$ and $x'$ have the same marginal distribution we obtain

$$\mathbb{E}_{init,\mu} [\|\delta(x, r, x'; W(t))\|_2^2] \leq \mathbb{E}_{init,\mu} [3(\hat{Q}_t(x)^2 + \pi^2 + \hat{Q}_t(x')^2)] = \mathbb{E}_{init,\mu} [(\tilde{Q}_t(x)^2 + 3\pi^2)]. \tag{B.4}$$

By (B.3), we know that $\hat{Q}(x; W)$ is 1-Lipschitz continuous with respect to $W$. Therefore, we have

$$|\hat{Q}_t(x) - \hat{Q}_0(x)| \leq \|W(t) - W(0)\|_2 \leq B, \tag{B.5}$$
Plugging (B.5) into (B.4) and using the Cauchy-Schwarz inequality we obtain
\[ \mathbb{E}_{\text{init}, \mu} \left[ \left| \delta (x, r, x'; W(t)) \right|^2 \right] \leq \mathbb{E}_{\text{init}, \mu} \left[ 12 \hat{Q}_0(x)^2 + 12 B^2 + 3 \pi^2 \right]. \] (B.6)

Note that by the initialization of \( \hat{Q}_0(x) \) as defined in (3.2), we have
\[ \mathbb{E}_{\text{init}, \mu} \left[ \left| \hat{Q}_0(x) \right|^2 \right] = \frac{1}{m} \sum_{r=1}^{m} \mathbb{E}_{\text{init}} \left[ \left| \sigma (W_r(0)^\top x) \right|^2 \right] \leq \mathbb{E}_{w \sim N(0, I_d/d)} \| w \|_2^2 = 1. \] (B.7)

Combining (B.2), (B.6), and (B.7) we obtain \( \mathbb{E}_{\text{init}, \mu} \left[ \| g(t) \|_2^2 \right] = O(B^2) \). Since
\[ \mathbb{E}_{\text{init}, \mu} \left[ \| g(t) - \bar{g}(t) \|_2^2 \right] = \mathbb{E}_{\text{init}} \left[ \mathbb{E}_\mu \left[ \| g(t) - \bar{g}(t) \|_2^2 \right] \right] \leq \mathbb{E}_{\text{init}} \left[ \mathbb{E}_\mu \left[ \| g(t) \|_2^2 \right] \right] = \mathbb{E}_{\text{init}, \mu} \left[ \| g(t) \|_2^2 \right], \]
we conclude the proof of Proposition 4.5. \( \square \)

### B.3 Proof of Proposition 4.7

**Proof.** By the triangle inequality, we have
\[ \| \hat{Q}_0(\cdot; W^*) - Q^\pi(\cdot) \|_\mu \leq \| \hat{Q}_0(\cdot; W^*) - \Pi_{\mathcal{F}_{B,m}} Q^\pi(\cdot) \|_\mu + \| \Pi_{\mathcal{F}_{B,m}} Q^\pi(\cdot) - Q^\pi(\cdot) \|_\mu. \] (B.8)

Since \( Q^\pi(\cdot) \) is the fixed-point solution to the Bellman equation, we replace \( Q^\pi(\cdot) \) by \( T^\pi Q^\pi(\cdot) \) and obtain
\[ \Pi_{\mathcal{F}_{B,m}} Q^\pi(\cdot) = \Pi_{\mathcal{F}_{B,m}} T^\pi Q^\pi(\cdot). \] (B.9)

Meanwhile, by Lemma 4.2, \( \hat{Q}_0(\cdot; W^*) \) is the solution to the projected Bellman equation, that is,
\[ \hat{Q}_0(\cdot; W^*) = \Pi_{\mathcal{F}_{B,m}} T^\pi \hat{Q}_0(\cdot; W^*). \] (B.10)

Combining (B.9) and (B.10), we obtain
\[ \| \hat{Q}_0(\cdot; W^*) - \Pi_{\mathcal{F}_{B,m}} Q^\pi(\cdot) \|_\mu = \| \Pi_{\mathcal{F}_{B,m}} T^\pi \hat{Q}_0(\cdot; W^*) - \Pi_{\mathcal{F}_{B,m}} T^\pi Q^\pi(\cdot) \|_\mu \]
\[ \leq \gamma \cdot \| \hat{Q}_0(\cdot; W^*) - Q^\pi(\cdot) \|_\mu, \] (B.11)
where the inequality follows from the fact that \( \Pi_{\mathcal{F}_{B,m}} T^\pi \) is a \( \gamma \)-contraction, as discussed in the proof of Lemma 4.2. Plugging (B.11) into (B.8), we obtain
\[ (1 - \gamma) \cdot \| \hat{Q}_0(\cdot; W^*) - Q^\pi(\cdot) \|_\mu \leq \| \Pi_{\mathcal{F}_{B,m}} Q^\pi(\cdot) - Q^\pi(\cdot) \|_\mu, \]
which completes the proof of Proposition 4.7. \( \square \)

### C Proofs for Section 5

#### C.1 Proof of Lemma 5.1

**Proof.** By the definition that \( \hat{Q}_t(x) = \hat{Q}(x; W(t)) \) and the definition of \( \hat{Q}_0(x; W(t)) \) in (4.7), we have
\[
| \hat{Q}_t(x) - \hat{Q}_0(x; W(t)) | \\
= \frac{1}{\sqrt{m}} \left| \sum_{r=1}^{m} ( \mathbb{1} \{ W_r(t)^\top x > 0 \} - \mathbb{1} \{ W_r(0)^\top x > 0 \} ) : b_r W_r(t)^\top x \right| \\
\leq \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \left| \mathbb{1} \{ W_r(t)^\top x > 0 \} - \mathbb{1} \{ W_r(0)^\top x > 0 \} \right| \left( \| W_r(0)^\top x \| + \| W_r(t) - W_r(0) \|_2 \right),
\] (C.1)
Thus, we finish the proof of Lemma

\[ |W_r(t)^\top x| \leq |W_r(t)^\top x - W_r(0)^\top x| \leq \|W_r(t) - W_r(0)\|_2. \]

Thus, we obtain

\[ |\mathbb{I}\{W_r(t)^\top x > 0\} - \mathbb{I}\{W_r(0)^\top x > 0\}| \leq \mathbb{I}\{W_r(0)^\top x \leq \|W_r(t) - W_r(0)\|_2\}. \]  

(C.2)

Plugging (C.2) into (C.1), we obtain the following upper bound,

\[
\|\hat{Q}_t(x) - \hat{Q}_0(x; W(t))\| \\
\leq \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{I}\{|W_r(0)^\top x| \leq \|W_r(t) - W_r(0)\|_2\} \cdot \left( |W_r(0)^\top x| + \|W_r(t) - W_r(0)\|_2 \right) \\
\leq \frac{2}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{I}\{|W_r(0)^\top x| \leq \|W_r(t) - W_r(0)\|_2\} \cdot \|W_r(t) - W_r(0)\|_2.
\]

Here the second inequality follows from the fact that

\[ \mathbb{I}\{|x| \leq y\} |x| \leq \mathbb{I}\{|x| \leq y\} y \]

for any \( x \) and \( y > 0 \). To characterize \( \mathbb{E}_{\text{init}, \mu}[|\hat{Q}_t(x) - \hat{Q}_0(x; W(t))|^2] \), we first invoke the Cauchy-Schwarz inequality and the fact that \( \|W(t) - W(0)\|_2 \leq B \), which gives

\[
|\hat{Q}_t(x) - \hat{Q}_0(x; W(t))|^2 \leq \frac{4B^2}{m} \sum_{r=1}^{m} \mathbb{I}\{|W_r(0)^\top x| \leq \|W_r(t) - W_r(0)\|_2\}.
\]

Taking expectation on both sides, by Lemma F.1 we obtain

\[
\mathbb{E}_{\text{init}, \mu}\left[|\hat{Q}_t(x) - \hat{Q}_0(x; W(t))|^2\right] \leq 4c_1 B^3 \cdot m^{-1/2}.
\]

Thus, we finish the proof of Lemma 5.1.

\[ \square \]

C.2 Proof of Lemma 5.2

Proof. By the definition of \( \overline{\eta}(t) \) and \( \overline{\eta}_0(t) \) in (5.1) and (5.2), respectively, we have

\[
\|\overline{\eta}(t) - \overline{\eta}_0(t)\|_2 = \left\| \mathbb{E}_{\mu}\left[ \delta(x, r; x' W(t)) \cdot \nabla_W \hat{Q}_t(x) - \delta_0(x, r; x' W(t)) \cdot \nabla_W \hat{Q}_0(x; W(t)) \right]\right\|_2 \\
\leq \left\| \mathbb{E}_{\mu}\left[ \left( \delta(x, r; x' W(t)) - \delta_0(x, r; x' W(t)) \right) \cdot \nabla_W \hat{Q}_t(x) \right. \right. \\
\left. \left. + \delta_0(x, r; x' W(t)) \cdot \left( \nabla_W \hat{Q}_t(x) - \nabla_W \hat{Q}_0(x; W(t)) \right) \right]\right\|_2 \\
\leq \mathbb{E}_{\mu}\left[ |\delta(x, r; x' W(t)) - \delta_0(x, r; x' W(t))| + |\delta_0(x, r; x' W(t))| \cdot \left\| \nabla_W \hat{Q}_t(x) - \nabla_W \hat{Q}_0(x; W(t)) \right\|_2 \right]. \]  

(C.3)

Here to obtain the second inequality, we use the fact that, for any \( t \in [T] \),

\[ \|\nabla_W \hat{Q}_t(x)\|_2 \leq \|x\|_2 = 1. \]
Taking expectation with respect to the random initialization on the both sides of (C.3), we obtain
\[
\mathbb{E}_{\text{init}} \left[ \| \mathbf{p}(t) - \mathbf{p}_0(t) \|^2 \right] \\
\leq 2 \mathbb{E}_{\text{init}, \mu} \left[ | \delta(x, r, x'; W(t)) - \delta_0(x, r, x'; W(t)) |^2 \right] \\
\quad + 2 \mathbb{E}_{\text{init}} \mathbb{E}_\mu \left[ | \delta_0(x, r, x'; W(t)) |^2 \right] \\
\quad + \mathbb{E}_\mu \left[ \| \nabla_W \hat{Q}_t(x) - \nabla_W \hat{Q}_0(x; W(t)) \|^2 \right].
\] (C.4)

In the following, we characterize the three terms on the right-hand side of (C.4).

For (i) in (C.4), note that
\[
\left| \delta(x, r, x'; W(t)) - \delta_0(x, r, x'; W(t)) \right|^2 \\
= \left| (\hat{Q}_t(x) - r - \gamma \hat{Q}_t(x')) - (\hat{Q}_0(x; W(t)) - r - \gamma \hat{Q}_0(x'; W(t))) \right|^2 \\
= \left| (\hat{Q}_t(x) - \hat{Q}_0(x; W(t))) - \gamma (\hat{Q}_t(x') - \hat{Q}_0(x'; W(t))) \right|^2 \\
\leq 2 \left( \hat{Q}_t(x) - \hat{Q}_0(x; W(t)) \right)^2 + 2 \left( \hat{Q}_t(x') - \hat{Q}_0(x'; W(t)) \right)^2.
\] (C.5)

Since \( x \) and \( x' \) follow the same stationary distribution \( \mu \) on the right-hand side of (C.5), by Lemma 5.1 we have
\[
\mathbb{E}_{\text{init}, \mu} \left[ | \delta(x, r, x'; W(t)) - \delta_0(x, r, x'; W(t)) |^2 \right] \\
\leq 4 \mathbb{E}_{\text{init}, \mu} \left[ | \hat{Q}_t(x) - \hat{Q}_0(x; W(t)) |^2 \right] \leq 16c_1B^3 \cdot m^{-1/2}.
\] (C.6)

For (ii) in (C.4), we have
\[
\| \nabla_W \hat{Q}_t(x) - \nabla_W \hat{Q}_0(x; W(t)) \|^2 \\
\leq \frac{1}{m} \sum_{r=1}^{m} \left( \mathbb{1} \{ W_r(t)^\top > 0 \} - \mathbb{1} \{ W_r(0)^\top > 0 \} \right)^2 \| x \|^2 \\
\leq \frac{1}{m} \sum_{r=1}^{m} \mathbb{1} \{ | W_r(0)^\top x | \leq \| W_r(t) - W_r(0) \|_2 \},
\] (C.7)

where the inequality follows from (C.2) and the fact that \( \| x \|_2 = 1 \).

For (iii) in (C.4), we have
\[
| \delta_0(x, r, x'; W(t)) |^2 \leq 3 \left( \hat{Q}_0(x; W(t))^2 + \gamma^2 \hat{Q}_0(x'; W(t))^2 \right).
\] (C.8)

To obtain an upper bound of the right-hand side of (C.8), we use the fact that
\[
| \hat{Q}_0(x; W(t)) - \hat{Q}_0(x) | \leq \| W(t) - W(0) \|_2 \cdot \| x \|_2 \leq B,
\]

which follows from (4.7), and obtain
\[
\mathbb{E}_\mu \left[ \hat{Q}_0(x; W(t))^2 \right] = \mathbb{E}_\mu \left[ \left( \hat{Q}_0(x) + \hat{Q}_0(x; W(t)) - \hat{Q}_0(x) \right)^2 \right] \leq 2 \mathbb{E}_\mu \left[ \hat{Q}_0(x)^2 \right] + 2B^2.
\]

22
Since $x$ and $x'$ follow the same stationary distribution $\mu$ on the right-hand side of (C.8) and $|\gamma| < 1$, we have

$$\begin{align*}
\mathbb{E}_\mu \left[ \left| \delta_0(x, r, r'; W(t)) \right|^2 \right] &\leq 12 \mathbb{E}_\mu \left[ \hat{Q}_0(x)^2 \right] + 12B^2 + 3\sigma^2. \quad \text{(C.9)}
\end{align*}$$

Plugging (C.6), (C.7), and (C.9) into (C.4), we obtain

$$\mathbb{E}_{\text{init}} \left[ \| \mathbf{g}(t) - \mathbf{g}_0(t) \|^2_2 \right] \leq 32c_1B^3 \cdot m^{-1/2} + 2\mathbb{E}_{\text{init}} \left[ \left( 12 \mathbb{E}_\mu \left[ \hat{Q}_0(x)^2 \right] + 12B^2 + 3\sigma^2 \right) \cdot \left( \frac{1}{m} \sum_{r=1}^m 1\{|W_r(0)^\top x| \leq \|W_r(t) - W_r(0)\|_2\} \right) \right].$$

Invoking Lemmas F.1 and F.2, we obtain

$$\mathbb{E}_{\text{init}} \left[ \| \mathbf{g}(t) - \mathbf{g}_0(t) \|^2_2 \right] \leq (56c_1B^3 + 24c_2B + 6c_1B\sigma^2) \cdot m^{-1/2},$$

which finishes the proof of Lemma 5.2. \hfill \Box

### C.3 Proof of Lemma 5.3

**Proof.** Recall that

$$\begin{align*}
\mathbf{g}(t) &= \mathbb{E}_\mu \left[ \delta(x, r, r'; W(t)) \cdot \nabla_W \hat{Q}(x; W(t)) \right], \quad \text{(C.10)}
\mathbf{g}_0(t) &= \mathbb{E}_\mu \left[ \delta_0(x, r, r'; W(t)) \cdot \nabla_W \hat{Q}_0(x; W(t)) \right]. \quad \text{(C.11)}
\end{align*}$$

We denote the locally linearized population semigradient $\mathbf{g}_0(t)$ evaluated at the approximate stationary point $W^*$ by

$$\mathbf{g}_0^* = \mathbb{E}_\mu \left[ \delta_0(x, r, r'; W^*) \cdot \nabla_W \hat{Q}_0(x; W^*) \right]. \quad \text{(C.12)}$$

For any $W(t)$ ($t \in [T]$), by the convexity of $S_B$, we have

$$\begin{align*}
\|W(t + 1) - W^*\|^2_2 &= \|\Pi_{S_B}(W(t) - \eta \cdot \mathbf{g}(t)) - \Pi_{S_B}(W^* - \eta \cdot \mathbf{g}_0^*)\|^2_2 \\
&\leq \|\Pi_{S_B}(W(t) - W^*) - \eta \cdot \mathbf{g}_0^*\|^2_2 \\
&= \|W(t) - W^*\|^2_2 - 2\eta \cdot (\mathbf{g}(t) - \mathbf{g}_0^\top) (W(t) - W^*) + \eta^2 \cdot \|\mathbf{g}(t) - \mathbf{g}_0^*\|^2_2. \quad \text{(C.13)}
\end{align*}$$

We decompose the inner product $(\mathbf{g}(t) - \mathbf{g}_0^*)^\top (W(t) - W^*)$ on the right-hand side of (C.13) into two terms,

$$\begin{align*}
(\mathbf{g}(t) - \mathbf{g}_0^*)^\top (W(t) - W^*) &= (\mathbf{g}_0(t) - \mathbf{g}_0^*)^\top (W(t) - W^*) + (\mathbf{g}(t) - \mathbf{g}_0(t))^\top (W(t) - W^*) \\
&\geq (\mathbf{g}_0(t) - \mathbf{g}_0^*)^\top (W(t) - W^*) - B \cdot \|\mathbf{g}(t) - \mathbf{g}_0(t)\|_2. \quad \text{(C.14)}
\end{align*}$$

It remains to characterize the first term $(\mathbf{g}_0(t) - \mathbf{g}_0^*)^\top (W(t) - W^*)$ on the right-hand side of (C.14), since the second term $\|\mathbf{g}(t) - \mathbf{g}_0(t)\|_2$ is characterized by Lemma 5.2. Note that by (C.11) and (C.12), we have

$$\mathbf{g}_0(t) - \mathbf{g}_0^* = \mathbb{E}_\mu \left[ \left( \delta_0(x, r, r'; W(t)) - \delta_0(x, r, r'; W^*) \right) \cdot \nabla_W \hat{Q}_0(x; W(0)) \right]. \quad \text{(C.15)}$$

23
where we use the following consequence of (4.7),
\[ \nabla_W \hat{Q}_0(x; W(0)) = \nabla_W \hat{Q}_0(x; W^*). \]

Moreover, by (4.8) it holds that
\[ \delta_0(x, r, x'; W(t)) - \delta_0(x, r, x'; W^*) \]
\[ = \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right) - \gamma \left( \hat{Q}_0(x'; W(t)) - \hat{Q}_0(x'; W^*) \right). \]  

Combining (4.7), (C.15), and (C.16), we have
\[ \left( \overline{\gamma}_0(t) - \overline{\gamma}_0(t) \right)^\top (W(t) - W^*) \]
\[ = \mathbb{E}_\mu \left[ \left( \delta_0(x, r, x'; W(t)) - \delta_0(x, r, x'; W^*) \right) \cdot \left( \nabla_W \hat{Q}_0(x; W(0))^\top (W(t) - W^*) \right) \right] \]
\[ = \mathbb{E}_\mu \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right. \]
\[ - \gamma \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right) \cdot \left( \hat{Q}_0(x'; W(t)) - \hat{Q}_0(x'; W^*) \right) \]
\[ \geq (1 - \gamma) \cdot \mathbb{E}_\mu \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right]. \]

where the last inequality is from the fact that x and x' have the same marginal distribution under μ and therefore by the Cauchy-Schwarz inequality,
\[ \mathbb{E}_\mu \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right) \cdot \left( \hat{Q}_0(x'; W(t)) - \hat{Q}_0(x'; W^*) \right) \right] \]
\[ \leq \mathbb{E}_\mu \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right] \cdot \mathbb{E}_\mu \left[ \left( \hat{Q}_0(x'; W(t)) - \hat{Q}_0(x'; W^*) \right)^2 \right]^{1/2} \]
\[ = \mathbb{E}_\mu \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right]. \]

Inequality (C.17) is the key to our convergence result. It shows that the locally linearized population semi-gradient update \( \overline{\gamma}_0(t) \) is one-point monotonic to the approximate stationary point \( W^* \).

Also, for \( \| \overline{\gamma}(t) - \overline{\gamma}_0 \|_2^2 \) on the right-hand side of (C.13), we have
\[ \| \overline{\gamma}(t) - \overline{\gamma}_0 \|_2^2 \leq 2\| \overline{\gamma}_0(t) - \overline{\gamma}_0 \|_2^2 + 2\| \overline{\gamma}(t) - \overline{\gamma}_0(t) \|_2^2. \]  

For the first term on the right-hand side of (C.19), by (C.15), (C.16), and the Cauchy-Schwarz inequality, we have
\[ \| \overline{\gamma}_0(t) - \overline{\gamma}_0 \|_2^2 = \mathbb{E}_\mu \left[ \left( \delta_0(x, r, x'; W(t)) - \delta_0(x, r, x'; W^*) \right) \cdot \nabla_W \hat{Q}_0(x; W(0)) \right]^2 \]
\[ \leq \mathbb{E}_\mu \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) - \gamma \hat{Q}_0(x'; W(t)) + \gamma \hat{Q}_0(x'; W^*) \right)^2 \right] \]
\[ \leq 4\mathbb{E}_\mu \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right]. \]

where the first inequality follows from the fact
\[ \| \nabla_W \hat{Q}_0(x; W(0)) \|_2 \leq \| x \|_2 = 1. \]

Plugging (C.17), (C.19), and (C.20) into (C.13), we finish the proof of Lemma 5.3.
C.4 Proof of Lemma 5.4

Proof. For any $W(t)$ ($t \in [T]$), by the convexity of $S_B$, (4.10), and (C.12), we have
\[ \|W(t + 1) - W^*\|^2 \leq \|\Pi_{S_B} (W(t) - \eta \cdot g(t)) - \Pi_{S_B} (W^* - \eta \cdot \overline{g}_0)\|^2 \]
\[ \leq \|(W(t) - \eta \cdot g(t)) - (W^* - \eta \cdot \overline{g}_0)\|^2 \]
\[ = \|W(t) - W^*\|^2 - 2\eta \cdot (g(t) - \overline{g}_0)^	op (W(t) - W^*) + \eta^2 \cdot \|g(t) - \overline{g}_0\|^2. \]

Taking expectation on both sides conditional on $W(t)$, we obtain
\[ \mathbb{E}_\mu [\|W(t + 1) - W^*\|^2 | W(t)] \]
\[ \leq \|W(t) - W^*\|^2 - 2\eta \cdot (\overline{g}(t) - \overline{g}_0)^	op (W(t) - W^*) + \eta^2 \cdot \mathbb{E}_\mu [\|g(t) - \overline{g}_0\|^2 | W(t)]. \] (C.21)

For the inner product $(\overline{g}(t) - \overline{g}_0)^	op (W(t) - W^*)$ on the right-hand side of (C.21), it follows from (C.14) and (C.17) that
\[ (\overline{g}(t) - \overline{g}_0)^	op (W(t) - W^*) \geq (1 - \gamma) \cdot \mathbb{E}_\mu \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right] - B \cdot \|\overline{g}(t) - \overline{g}_0(t)\|. \]

Meanwhile, for $\mathbb{E}_\mu [\|g(t) - \overline{g}_0\|^2 | W(t)]$ on the right-hand side of (C.21), we have the decomposition
\[ \mathbb{E}_\mu [\|g(t) - \overline{g}_0\|^2 | W(t)] = \|\overline{g}(t) - \overline{g}_0\|^2 + \mathbb{E}_\mu [\|g(t) - \overline{g}(t)\|^2 | W(t)] \]
\[ \leq 8\mathbb{E}_\mu \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right] \|W(t)\| + 2\|\overline{g}(t) - \overline{g}_0(t)\|^2 + \mathbb{E}_\mu [\|g(t) - \overline{g}(t)\|^2 | W(t)], \]

where the inequality follows from (C.19) and (C.20). Taking expectation on the both sides of (C.21) with respect to $W(t)$, we complete the proof of Lemma 5.4. \qed

C.5 Proof of Theorem 4.4

Proof. By Lemma 5.2 we have
\[ \mathbb{E}_{\text{init}} [\|g(t) - \overline{g}_0(t)\|^2] = O(B^3 m^{-1/2}), \] (C.22)
\[ \mathbb{E}_{\text{init}} [B \cdot \|\overline{g}(t) - \overline{g}_0(t)\|] = O(B^5/2 m^{-1/4}). \] (C.23)

Setting $\eta = (1 - \gamma)/8$ in Algorithm 1, by (C.22), (C.23), and Lemma 5.3, we have
\[ \mathbb{E}_{\text{init}, \mu} \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right] = \mathbb{E}_{\text{init}} [\|W(t) - W^*\|^2 - \|W(t + 1) - W^*\|^2] \]
\[ = \frac{\mathbb{E}_{\text{init}} [\|W(0) - W^*\|^2 - \|W(T) - W^*\|^2]}{(1 - \gamma)^2/8} + O(B^3 m^{-1/2} + B^{5/2} m^{-1/4}) \]
\[ + O(B^3 m^{-1/2} + B^{5/2} m^{-1/4}). \] (C.24)

Telescoping (C.24) for $t = 0, \ldots, T - 1$, we obtain
\[ \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_{\text{init}, \mu} \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right] \]
\[ = \frac{\mathbb{E}_{\text{init}} [\|W(0) - W^*\|^2 - \|W(T) - W^*\|^2]}{T(1 - \gamma)^2/8} + O(B^3 m^{-1/2} + B^{5/2} m^{-1/4}) \]
\[ \leq \frac{8B^2}{T(1 - \gamma)^2} + O(B^3 m^{-1/2} + B^{5/2} m^{-1/4}). \]
Recall that as defined in (4.7), \( \hat{Q}_0(\cdot; W) \) is linear in \( W \). By Jensen’s inequality, we have
\[
\mathbb{E}_{\text{init}, \mu} \left[ (\hat{Q}_0(x; W) - \hat{Q}_0(x; W^*))^2 \right] \leq \frac{8B^2}{T(1-\gamma)^2} + O(B^3m^{-1/2} + B^{5/2}m^{-1/4}). \tag{C.25}
\]
Next we characterize the output \( \hat{Q}_{\text{out}}(\cdot) = \hat{Q}(\cdot; W) \) of Algorithm 1. Since \( S_B \) is convex and \( W \in S_B \), by Lemma 5.1 we have
\[
\mathbb{E}_{\text{init}, \mu} \left[ (\hat{Q}_0(x; W) - \hat{Q}_0(x; W^*))^2 \right] = O(B^3m^{-1/2}). \tag{C.26}
\]
Using the Cauchy-Schwarz inequality we have
\[
\mathbb{E}_{\text{init}, \mu} \left[ (\hat{Q}_{\text{out}}(x) - \hat{Q}_0(x; W^*))^2 \right] \leq \mathbb{E}_{\text{init}, \mu} \left[ 2(\hat{Q}(x; W) - \hat{Q}_0(x; W))^2 + 2(\hat{Q}_0(x; W) - \hat{Q}_0(x; W^*))^2 \right],
\]
into which we plugging (C.25) and (C.26) and obtain
\[
\mathbb{E}_{\text{init}, \mu} \left[ (\hat{Q}_{\text{out}}(x) - \hat{Q}_0(x; W^*))^2 \right] \leq \frac{16B^2}{T(1-\gamma)^2} + O(B^3m^{-1/2} + B^{5/2}m^{-1/4}), \tag{C.27}
\]
which completes the proof of Theorem 4.4. \( \square \)

### C.6 Proof of Theorem 4.6

**Proof.** Similar to (C.24), by Lemmas 4.5, 5.2, and 5.4 we have
\[
\mathbb{E}_{\text{init}, \mu} \left[ (\hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*))^2 \right] \leq \mathbb{E}_{\text{init}} \left[ \|W(t) - W^*\|_2^2 \right] - \mathbb{E}_{\text{init}} \left[ \|W(t + 1) - W^*\|_2^2 \right] + \eta^2 \cdot \sigma_g^2 + O(B^3m^{-1/2} + B^{5/2}m^{-1/4}) \tag{C.28}
\]
Telescoping (C.28) for \( t = 0, \ldots, T - 1 \), by \( \eta^2 \leq 1/T \) we have
\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_{\text{init}, \mu} \left[ (\hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*))^2 \right] \leq \frac{\mathbb{E}_{\text{init}} \left[ \|W(t) - W^*\|_2^2 \right] + \sigma_g^2}{T \cdot (2\eta(1-\gamma) - 8\eta^2)} + O(B^3m^{-1/2} + B^{5/2}m^{-1/4}) \leq \frac{B^2 + \sigma_g^2}{\sqrt{T}} \cdot \frac{1}{\sqrt{T} \cdot (2\eta(1-\gamma) - 8\eta^2)} + O(B^3m^{-1/2} + B^{5/2}m^{-1/4}), \tag{C.29}
\]
where \( \eta = \min\{1/\sqrt{T}, (1-\gamma)/8\} \). Note that when \( T \geq (8/(1-\gamma))^2 \), we have \( \eta = 1/\sqrt{T} \) and
\[
\sqrt{T} \cdot (2\eta(1-\gamma) - 8\eta^2) = 2(1-\gamma) - 8/\sqrt{T} \geq 1 - \gamma.
\]
Meanwhile, when \( T < (8/(1-\gamma))^2 \), we have \( \eta = (1-\gamma)/8 \) and
\[
\sqrt{T} \cdot (2\eta(1-\gamma) - 8\eta^2) = \sqrt{T} \cdot (1-\gamma)^2/8 \geq (1-\gamma)^2/8.
\]

26
Since $|1 - \gamma| < 1$, we obtain that for any $T \in \mathbb{N}$,
\[
\frac{1}{\sqrt{T} \cdot (2\eta(1 - \gamma) - 8\eta^2)} \leq \frac{8}{(1 - \gamma)^2}.
\] (C.30)

Similar to (C.25) and (C.27), by combining (C.29) and (C.30) with Lemma 5.1, we obtain
\[
\mathbb{E}_{\text{init},\mu} \left[ \left( \hat{Q}_{\text{out}}(x) - \hat{Q}_0(x; W^*) \right)^2 \right] \leq \frac{16(B^2 + \sigma^2)}{\sqrt{T} \cdot (1 - \gamma)^2} + O(B^3m^{-1/2} + B^{5/2}m^{-1/4}),
\]
which completes the proof of Theorem 4.6.

\section{Proofs for Section 6}

Similar to the population semigradient $\overline{\nabla}(t)$ in policy evaluation, we define
\[
z(t) = \left( \hat{Q}(x; W(t)) - T\hat{Q}(x; W(t)) \right) \cdot \nabla_W \hat{Q}(x; W(t)),
\] (D.1)
\[
\overline{\nabla}(t) = \mathbb{E}_{\mu_{\exp}} \left[ \left( \hat{Q}(x; W(t)) - T\hat{Q}(x; W(t)) \right) \cdot \nabla_W \hat{Q}(x; W(t)) \right],
\] (D.2)
\[
\overline{\nabla}_0(t) = \mathbb{E}_{\mu_{\exp}} \left[ \left( \hat{Q}_0(x; W(t)) - T\hat{Q}_0(x; W(t)) \right) \cdot \nabla_W \hat{Q}_0(x; W(t)) \right],
\] (D.3)
\[
\overline{\nabla}_0(t) = \mathbb{E}_{\mu_{\exp}} \left[ \left( \hat{Q}_0(x; W^*) - T\hat{Q}_0(x; W^*) \right) \cdot \nabla_W \hat{Q}_0(x; W^*) \right].
\] (D.4)

Our proof extends that of Theorem 2 in Zou et al. (2019) for characterizing linear Q-learning. We additionally incorporate the error of local linearization and also handle soft Q-learning in the next section. The following lemma is analogous to Lemma 4.2.

\textbf{Lemma D.1.} Under Assumption 6.1, there exists a unique approximate stationary point $W^*$ that satisfies (6.3).

\textbf{Proof.} We prove the lemma by showing that $T$ is a contraction in $\| \cdot \|_{\mu_{\exp}}$ for any $Q_1, Q_2 \in \mathcal{F}_{B,m}$. By the definition of the Bellman optimality operator $T$, we have
\[
\| TQ_1 - TQ_2 \|_2^2 = \gamma^2 \mathbb{E}_{s \sim \mu_{\exp}} \left[ \left( \max_{a \in \mathcal{A}} Q_1(s, a) - \max_{a \in \mathcal{A}} Q_2(s, a) \right)^2 \right].
\]

Under Assumption 6.1, for any $Q_1, Q_2 \in \mathcal{F}_{B,m}$, we have
\[
\mathbb{E}_{s \sim \mu_{\exp}} \left[ \left( \max_{a \in \mathcal{A}} Q_1(s, a) - \max_{a \in \mathcal{A}} Q_2(s, a) \right)^2 \right] \leq (\gamma + \nu)^{-2} \mathbb{E}_{\mu_{\exp}} \left[ (Q_1(x) - Q_2(x))^2 \right].
\]

Therefore, $\Pi_{\mathcal{F}_{B,m}} T$ is a $\gamma/(\gamma + \nu)$-contraction in $\| \cdot \|_{\mu_{\exp}}$, since $\Pi_{\mathcal{F}_{B,m}}$ is nonexpansive. Since the set $S_B$ of feasible $W$ is closed and bounded, $\mathcal{F}_{B,m}$ is complete under $\| \cdot \|_{\mu_{\exp}}$. Thus, $\Pi_{\mathcal{F}_{B,m}} T$ has a unique fixed point in $\mathcal{F}_{B,m}$, which corresponds to $W^*$.

The following lemma is analogous to Lemma 5.2 with a similar proof.

\textbf{Lemma D.2.} For any $t \in [T]$, we have
\[
\mathbb{E}_{\text{init}} \left[ \| \overline{\nabla}(t) - \overline{\nabla}_0(t) \|_2^2 \right] = O(B^3m^{-1/2}).
\]
Proof. By the definitions of $\mathcal{F}(t)$ and $\mathcal{F}_0(t)$ in (D.2) and (D.3), respectively, we have
\[
\|\mathcal{F}(t) - \mathcal{F}_0(t)\|_2^2 = \mathbb{E}_{\mu_{\exp}} \left[ (\hat{Q}_t(x) - T\hat{Q}_t(x)) \cdot \nabla_W \hat{Q}_t(x) - (\hat{Q}_0(x; W(t)) - T\hat{Q}_0(x; W(t))) \cdot \nabla_W \hat{Q}_0(x; W(t)) \right] \leq \mathbb{E}_{\mu_{\exp}} \left[ (\hat{Q}_t(x) - \gamma \hat{Q}_t^*(s') - \hat{Q}_0(x; W(t)) + \gamma \hat{Q}_0^*(s'; W(t))) \cdot \nabla_W \hat{Q}_t(x) \right].
\]

For notational simplicity, we define $\hat{Q}_t^*(s) = \max_{a \in A} \hat{Q}_t(s, a)$. Recall that $\hat{Q}_0^*(s; W)$ is similarly defined in Assumption 6.1. Then on the right-hand side of (D.5), we have
\[
\mathbb{E}_{\mu_{\exp}} \left[ (\hat{Q}_t(x) - \gamma \hat{Q}_t^*(s') - \hat{Q}_0(x; W(t)) + \gamma \hat{Q}_0^*(s'; W(t))) \cdot \nabla_W \hat{Q}_t(x) \right].
\]

Thus, from (D.5) we obtain
\[
\|\mathcal{F}(t) - \mathcal{F}_0(t)\|_2^2 \leq 2\mathbb{E}_{\mu_{\exp}} \left[ (\hat{Q}_t(x) - \gamma \hat{Q}_t^*(s') - \hat{Q}_0(x; W(t)) + \gamma \hat{Q}_0^*(s'; W(t)))^2 \right]
\]
\[
+ 2\mathbb{E}_{\mu_{\exp}} \left[ |\hat{Q}_0(x; W(t)) - r(x) - \gamma \hat{Q}_0^*(s'; W(t))|^2 \right] \cdot \mathbb{E}_{\mu_{\exp}} \left[ \|\nabla_W \hat{Q}_t(x) - \nabla_W \hat{Q}_0(x; W(t))\|_2^2 \right].
\]

Here we use the fact that, for any $t \in [T]$,
\[
\|\nabla_W \hat{Q}_t(x)\|_2 \leq \|x\|_2 = 1.
\]

Taking expectation on the both sides of (D.6) with respect to the random initialization, we obtain
\[
\mathbb{E}_{\text{init}} \left[ \|\mathcal{F}(t) - \mathcal{F}_0(t)\|_2^2 \right]
\]
\[
\leq 4\mathbb{E}_{\text{init,\mu}_{\exp}} \left[ |\hat{Q}_t(x) - \hat{Q}_0(x; W(t))|^2 \right] + 4\gamma^2 \mathbb{E}_{\text{init,\mu}_{\exp}} \left[ |\hat{Q}_t^*(s) - \hat{Q}_0^*(s; W(t))|^2 \right]
\]
\[
+ 2\mathbb{E}_{\text{init}} \mathbb{E}_{\mu_{\exp}} \left[ |\hat{Q}_0(x; W(t)) - r(x) - \gamma \hat{Q}_0^*(s'; W(t))|^2 \right] \cdot \mathbb{E}_{\mu_{\exp}} \left[ \|\nabla_W \hat{Q}_t(x) - \nabla_W \hat{Q}_0(x; W(t))\|_2^2 \right].
\]

Similar to the proof of Lemma 5.2, we characterize the three terms on the right-hand side of (D.7). For (i) in (D.7), recall that Lemma 5.1 gives
\[
\mathbb{E}_{\text{init,\mu}_{\exp}} \left[ |\hat{Q}_t(x) - \hat{Q}_0(x; W(t))|^2 \right] \leq 4c_1 B^3 \cdot m^{-1/2}.
\]

We establish a similar upper bound of (ii) in (D.7). Note that
\[
|\hat{Q}_t^*(s) - \hat{Q}_0^*(s; W(t))| = \max_{a \in A} |\hat{Q}_t(s, a) - \hat{Q}_0(s, a; W(t))|.
\]

Similar to Lemma 5.1, we have the following lemma for characterizing the right-hand side of (D.8).
Lemma D.3. Under Assumption 6.2, there exists a constant $c_4 > 0$ such that for any $t \in [T]$, it holds that
\[
\mathbb{E}_{\text{init}, s \sim \mu_{\text{exp}}} \left[ \max_{a \in A} \left| \hat{Q}_t(s, a) - \hat{Q}_0(s, a; W(t)) \right|^2 \right] \leq c_4 B_3 \cdot m^{-1/2}.
\]

Proof. See Appendix F.1 for a detailed proof.

By Lemma D.3, (ii) in (D.7) satisfies
\[
\mathbb{E}_{\text{init}, s \sim \mu_{\text{exp}}} \left[ \left| \hat{Q}_t^2(s) - \hat{Q}_0^2(s; W(t)) \right|^2 \right] \leq \mathbb{E}_{\text{init}, s \sim \mu_{\text{exp}}} \left[ \max_{a \in A} \left| \hat{Q}_t(s, a) - \hat{Q}_0(s, a; W(t)) \right|^2 \right] \leq c_4 B_3 \cdot m^{-1/2}.
\]
For (iii) in (D.7), in the proof of Lemma 5.2, we show that
\[
\mathbb{E}_{\mu_{\text{exp}}} \left[ \left\| \nabla_W \hat{Q}_t(x) - \nabla_W \hat{Q}_0(x; W(t)) \right\|_2^2 \right] \leq \mathbb{E}_{\mu_{\text{exp}}} \left[ \frac{1}{m} \sum_{r=1}^{m} 1 \{ \| W_r(0)^\top x \| \leq \| W_r(t) - W_r(0) \|_2 \} \right].
\]
Meanwhile, by the definition that $\hat{Q}_t^0(s; W) = \max_{a \in A} \hat{Q}_0(s, a; W)$ and (4.7), we have
\[
\left| \hat{Q}_0(x; W(t)) - r(x) - \gamma \hat{Q}_t^0(s'; W(t)) \right|^2 \\
\leq 3\| \hat{Q}_0(x; W(t)) \|^2 + 3\left| \hat{Q}_0(s', a^\prime_{\text{max}}; W(t)) \right|^2 \\
\leq 6\| \hat{Q}_0(x; W(0)) \|^2 + 3\left| \hat{Q}_0(s', a^\prime_{\text{max}}; W(0)) \right|^2 + 12B^2,
\]
where $a^\prime_{\text{max}} = \arg\max_{a' \in A} \hat{Q}_0(s', a'; W(t))$. Then applying Lemmas F.1, F.2, and F.3, we obtain that (iii) in (D.7) satisfies
\[
2\mathbb{E}_{\text{init}} \left[ \mathbb{E}_{\mu_{\text{exp}}} \left[ \left| \hat{Q}_0(x; W(t)) - r(x) - \gamma \hat{Q}_t^0(s'; W(t)) \right|^2 \right] \cdot \mathbb{E}_{\mu_{\text{exp}}} \left[ \left\| \nabla_W \hat{Q}_t(x) - \nabla_W \hat{Q}_0(x; W(t)) \right\|_2^2 \right] \right] \\
= O(B^3 m^{-1/2}).
\]
Combining the upper bounds of (i)-(iii) in (D.7), we complete the proof of Lemma D.2.

D.1 Proof of Theorem 6.3

Proof. Recall that $z(t)$, $\tau(t)$, $\tau_0(t)$, and $\tau_0$ are defined in (D.1)-(D.4), respectively. Similar to the proof of Theorem 4.4, we have
\[
\| W(t + 1) - W^* \|_2^2 = \| \Pi_{S_0} \left( W(t) - \eta \cdot z(t) \right) - \Pi_{S_0} (W^* - \eta \cdot \tau_0) \|_2^2 \\
\leq \| W(t) - W^* \|_2^2 - 2\eta \cdot (z(t) - \tau_0)^\top (W(t) - W^*) + \eta^2 \cdot \| z(t) - \tau_0 \|_2^2.
\]

To characterize the inner product on the right-hand side of (D.10), we take conditional expectation and obtain
\[
\mathbb{E}_{\mu_{\text{exp}}} \left[ (z(t) - \tau_0)^\top (W(t) - W^*) \mid W(t) \right] \\
= (\tau_0 - \tau_0)^\top (W(t) - W^*) + (\tau(t) - \tau_0(t))^\top (W(t) - W^*) \\
\geq (\tau_0 - \tau_0)^\top (W(t) - W^*) - B \cdot \| \tau(t) - \tau_0(t) \|_2.
\]

29
We establish a lower bound of \((\overline{z}_0(t) - \overline{z}_0)^\top (W(t) - W^*)\) as follows. By (D.3) and (D.4), we have
\[
(\overline{z}_0(t) - \overline{z}_0)^\top (W(t) - W^*)
= \mathbb{E}_{\mu}\left[\left(\hat{Q}_0(x; W(t)) - \gamma \hat{Q}_0^\prime(s'; W(t)) - \hat{Q}_0(x; W^*) + \gamma \hat{Q}_0^\prime(s'; W^*) \right) \cdot \nabla_W \hat{Q}_0(x; W(t))\right]^\top (W(t) - W^*)
= \mathbb{E}_{\mu}\left[\left(\hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*)\right)^2\right]
- \gamma \mathbb{E}_{\mu}\left[\left(\hat{Q}_0^\prime(s'; W(t)) - \hat{Q}_0^\prime(s'; W^*)\right) \cdot \left(\hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*)\right)\right].
\]

Applying Hölder’s inequality to the second term on the right-hand side of (D.12), we obtain
\[
\mathbb{E}_{\mu}\left[\left(\hat{Q}_0^\prime(s'; W(t)) - \hat{Q}_0^\prime(s'; W^*)\right) \cdot \left(\hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*)\right)\right]
\leq \mathbb{E}_{\mu}\left[\left(\hat{Q}_0^\prime(s'; W(t)) - \hat{Q}_0^\prime(s'; W^*)\right)^2\right]^{1/2} \cdot \mathbb{E}_{\mu}\left[\left(\hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*)\right)^2\right]^{1/2}.
\]

By Assumption 6.1, we have
\[
\mathbb{E}_{\mu}\left[\left(\hat{Q}_0^\prime(s'; W(t)) - \hat{Q}_0^\prime(s'; W^*)\right)^2\right]^{1/2} \leq 1/(\gamma + \nu) \cdot \mathbb{E}_{\mu}\left[\left(\hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*)\right)^2\right]^{1/2},
\]
which implies that, on the right-hand side of (D.12),
\[
\gamma \mathbb{E}_{\mu}\left[\left(\hat{Q}_0^\prime(s'; W(t)) - \hat{Q}_0^\prime(s'; W^*)\right) \cdot \left(\hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*)\right)\right]
\leq \gamma/(\gamma + \nu) \cdot \mathbb{E}_{\mu}\left[\left(\hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*)\right)^2\right].
\]
Therefore, from (D.12) we obtain
\[
(\overline{z}_0(t) - \overline{z}_0)^\top (W(t) - W^*) \geq \nu/(\gamma + \nu) \cdot \mathbb{E}_{\mu}\left[\left(\hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*)\right)^2\right]. \tag{D.13}
\]

Similar to the proof of Theorem 4.4, for \(\|z(t) - \overline{z}_0\|_2^2\) on the right-hand side of (D.10), we have
\[
\mathbb{E}_{W,\mu}\left[\|z(t) - \overline{z}_0\|_2^2\right] \leq \mathbb{E}_{W,\mu}\left[\|z(t) - \overline{z}_0\|_2^2\right] + \mathbb{E}_{W}\left[2\|z(t) - \overline{z}_0(t)\|_2^2 + 2\|\overline{z}_0(t) - \overline{z}_0\|_2^2\right], \tag{D.14}
\]
where the expectation of \(\|\overline{z}(t) - \overline{z}_0(t)\|_2^2\) on the right-hand side is characterized by Lemma D.2, while the expectation of \(\|\overline{z}_0(t) - \overline{z}_0\|_2^2\) has the following upper bound,
\[
\mathbb{E}_{W}\left[\|\overline{z}_0(t) - \overline{z}_0\|_2^2\right]
= \mathbb{E}_{W,\mu}\left[\left(\hat{Q}_0(x; W(t)) - \gamma \hat{Q}_0^\prime(s'; W(t)) - \hat{Q}_0(x; W^*) + \gamma \hat{Q}_0^\prime(s'; W^*) \right) \cdot \nabla_W \hat{Q}_0(x; W(t))\right]^2
\leq \mathbb{E}_{W,\mu}\left[\left(\hat{Q}_0(x; W(t)) - \hat{Q}_0^\prime(s'; W(t)) - \hat{Q}_0(x; W^*) + \gamma \hat{Q}_0^\prime(s'; W^*) \right)^2\right]
\leq \mathbb{E}_{W,\mu}\left[\left(\hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*)\right)^2\right]. \tag{D.15}
\]
Here the last inequality follows from Assumption 6.1 as \(\gamma/(\gamma + \nu) < 1\).
Plugging (D.11), (D.13), (D.14), and (D.15) into (D.10) yields the following inequality, which parallels Lemma 5.4,

\[ \mathbb{E}_{W,\eta} \left[ \|W(t + 1) - W^*\|^2 \right] \]
\[ \leq \mathbb{E}_{W} \left[ \|W(t) - W^*\|^2 \right] - (2\eta(\gamma + \nu) - 8\eta^2) \cdot \mathbb{E}_{W,\eta} \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right] \]
\[ + \mathbb{E}_{W} \left[ 2\eta^2 \cdot \|\hat{z}(t) - \hat{z}_0(t)\|^2 + 2\eta B \cdot \|\hat{z}(t) - \hat{z}_0(t)\|_2 \right] + \mathbb{E}_{W,\eta} \left[ \eta^2 \cdot \|\hat{z}(t) - \hat{z}(t)\|^2 \right]. \quad \text{(D.16)}

Rearranging terms of (D.16), we get

\[ \mathbb{E}_{W,\eta} \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right] \]
\[ \leq (2\eta(\gamma + \nu) - 8\eta^2)^{-1} \left( \mathbb{E}_{W} \left[ \|W(t) - W^*\|^2 \right] - \mathbb{E}_{W,\eta} \left[ \|W(t + 1) - W^*\|^2 \right] \right) \]
\[ + \mathbb{E}_{W} \left[ 2\eta^2 \cdot \|\hat{z}(t) - \hat{z}_0(t)\|^2 + 2\eta B \cdot \|\hat{z}(t) - \hat{z}_0(t)\|_2 \right] + \mathbb{E}_{W,\eta} \left[ \eta^2 \cdot \|\hat{z}(t) - \hat{z}(t)\|^2 \right]. \quad \text{(D.17)}

In parallel with Lemma 4.5, we establish an upper bound of the variance \( \mathbb{E}_{W,\eta} \left[ \|z(t) - \hat{z}(t)\|^2 \right] \) in the right-hand side of (D.17), which is independent of \( t \) and \( m \). Note that by \( \|\nabla W \hat{Q}_t(s, a)\|_2 \leq 1 \), we have

\[ \|z(t)\|^2 \leq \left| \hat{Q}_t(s, a) - r(s, a) - \gamma \max_{a' \in A} \hat{Q}_t(s', a') \right|^2 \cdot \|\nabla W \hat{Q}_t(s, a)\|_2^2 \]
\[ \leq 3 \hat{Q}_t(s, a)^2 + 3\eta^2 + 3 \max_{a' \in A} \hat{Q}_t(s', a')^2. \quad \text{(D.18)}

To characterize \( \hat{Q}_t(s, a)^2 \) we have

\[ \hat{Q}_t(s, a)^2 = \left( \hat{Q}_t(s, a) - \hat{Q}_0(s, a) + \hat{Q}_0(s, a) \right)^2 \]
\[ \leq 2 \left| \hat{Q}_t(s, a) - \hat{Q}_0(s, a) \right|^2 + 2\hat{Q}_0(s, a)^2 \]
\[ \leq 2B^2 + 2\hat{Q}_0(s, a)^2. \quad \text{(D.19)}

where the second inequality comes from (B.5). Similarly, to characterize \( \max_{a' \in A} \hat{Q}_t(s', a') \) in (D.18), we take maximum on the both side of (D.19) and obtain

\[ \max_{a' \in A} \hat{Q}_t(s', a')^2 \leq 2B^2 + 2 \max_{a' \in A} \hat{Q}_0(s', a')^2. \quad \text{(D.20)}

Plugging (D.19) and (D.20) into (D.18), we obtain

\[ \|z(t)\|^2 \leq 6\hat{Q}_0(s, a)^2 + 3\eta^2 + 6 \max_{a' \in A} \hat{Q}_0(s', a')^2 + 12B^2. \quad \text{(D.21)}

To upper bound the expectation of (D.21), it remains to characterize \( \mathbb{E}_{W,\eta} \left[ \hat{Q}_0(s, a)^2 \right] \) and

\[ \mathbb{E}_{W,\eta} \left[ \max_{a' \in A} \hat{Q}_0(s', a')^2 \right]. \]

In fact, for any \((s, a), \mathbb{E}_{W} \left[ \hat{Q}_0(s, a)^2 \right] \) has the following uniform upper bound

\[ \mathbb{E}_{\text{init}} \left[ \hat{Q}_0(s, a)^2 \right] = \mathbb{E}_{w \sim N(0, I_d/d)} [\sigma(w^\top x)^2] \leq \mathbb{E}_{w \sim N(0, I_d/d)} [\|w\|_2^2] = 1. \quad \text{(D.22)}

31
For $\mathbb{E}_{\text{init}, s' \sim \mu_{\text{exp}}} \left[ \max_{a' \in A} \hat{Q}_0(s', a')^2 \right]$, we use the inequality
\[
\mathbb{E}_{\text{init}, s' \sim \mu_{\text{exp}}} \left[ \max_{a' \in A} \hat{Q}_0(s', a')^2 \right] \\
\leq \mathbb{E}_{\text{init}, s' \sim \mu_{\text{exp}}} \left[ \sum_{a' \in A} \hat{Q}_0(s', a')^2 \right] \\
= \sum_{a' \in A} \mathbb{E}_{s' \sim \mu_{\text{exp}}} \left[ \mathbb{E}_{\text{init}}[\hat{Q}_0(s', a')^2 \mid (s', a')] \right] \\
\leq |A| \cdot \mathbb{E}_{\|w\| \sim \mathcal{N}(0, I_d/d)} \left[ \|w\|^2 \right],
\]
which gives us the same upper bound as in (D.22) but with an additional factor $|A|$. Therefore, we know that the variance $\mathbb{E}_{\text{init}, \mu_{\exp}}[\|z(t) - \bar{z}(t)\|^2_2]$ in the expectation of (D.17) has an upper bound $\sigma_z^2 = O(B^2)$, according to the fact that $\mathbb{E}_{\text{init}, \mu_{\exp}}[\|z(t) - \bar{z}(t)\|^2_2] \leq \mathbb{E}_{\text{init}, \mu_{\exp}}[\|z(t)\|^2_2].$

With the variance term upper bounded, we take expectation on (D.17) over the random initialization and obtain
\[
\mathbb{E}_{\text{init}, \mu_{\exp}} \left[ \left( \hat{Q}_0(x; W(t)) - \hat{Q}_0(x; W^*) \right)^2 \right] \\
\leq \left( \frac{2\eta}{(\gamma + \nu) - 8\eta^2} \right)^{-1} \left( \mathbb{E}_{\text{init}}[\|W(t) - W^*\|^2_2] - \mathbb{E}_{\text{init}}[\|W(t + 1) - W^*\|^2_2] + \eta^2 \sigma_z^2 \right) \\
+ O(B^3m^{-1/2} + B^5/2m^{-1/4}). \tag{D.23}
\]
We let $\eta = \min\{1/\sqrt{T}, \nu/8(\gamma + \nu)\}$ and telescope (D.23) for $t = 0, \ldots, T - 1$. Then the rest proof mirrors that of Theorem 4.6. Thus, we complete the proof of Theorem 6.3. \hfill \Box

## E From Neural Soft Q-Learning to Policy Gradient

### E.1 Global Convergence of Neural Soft Q-Learning

We extend the global convergence of neural Q-learning in Section 6 to neural soft Q-learning, where the max operator is replaced by a more general softmax operator. More specifically, we consider the soft Bellman optimality operator
\[
\mathcal{T}_\beta(Q)(s, a) = \mathbb{E} \left[ r(s, a) + \gamma \cdot \text{softmax}_{a' \in A} Q(s', a') \mid s' \sim \mathcal{P}(\cdot \mid s, a) \right],
\]
where $\text{softmax}_{a' \in A} Q(s', a') = \beta^{-1} \cdot \log \sum_{a' \in A} \exp \left( \beta \cdot Q(s', a') \right), \tag{E.1}$

which in parallel with (6.2) corresponds to the update
\[
\theta' \leftarrow \theta - \eta \cdot \left( \hat{Q}_\theta(s, a) - r(s, a) - \gamma \cdot \text{softmax}_{a' \in A} \hat{Q}_\theta(s', a') \right) \cdot \nabla \theta \hat{Q}_\theta(s, a). \tag{E.2}
\]

See Algorithm 3 for a detailed description of such neural soft Q-learning algorithm.

In parallel with Assumption 6.1, we require the following regularity condition on the exploration policy $\pi_{\text{exp}}$.

**Assumption E.1** (Regularity of Exploration Policy $\pi_{\text{exp}}$). There exists a constant $\nu' > 0$ such that for any $W_1, W_2 \in S_B$, it holds that
\[
\mathbb{E}_{x \sim \mu_{\exp}} \left[ \left( \hat{Q}_0(x; W_1) - \hat{Q}_0(x; W_2) \right)^2 \right] \\
\geq (\gamma + \nu')^2 \cdot \mathbb{E}_{s' \sim \mu_{\exp}} \left[ \left( \text{softmax}_{a \in A} \hat{Q}_0(s, a; W_1) - \text{softmax}_{a \in A} \hat{Q}_0(s, a; W_2) \right)^2 \right]. \tag{E.3}
\]
We remark that, when $\beta \to \infty$, the softmax operator converges to the max operator, which implies that Assumptions E.1 and 6.1 are equivalent.

The approximate stationary point $W^*$ of the projected soft Q-learning satisfies

$$\hat{Q}_0(\cdot; W^*) = \Pi_{\mathcal{F}_B} T_\beta \hat{Q}_0(\cdot; W^*),$$

which uniquely exist by the same proof of Lemma D.1. Under the above regularity condition, we can extend Theorem 6.3 to cover neural soft Q-learning.

**Theorem E.2** (Convergence of Stochastic Update). We set $\eta$ to be of order $T^{-1/2}$ in Algorithm 3. Under Assumptions E.1 and 6.2, the output $\hat{Q}_\text{out}$ of Algorithm 3 satisfies

$$\mathbb{E}_{\text{init,}\mu_{\text{exp}}} \left( \left( \hat{Q}_\text{out}(x) - \hat{Q}_0(x; W^*) \right)^2 \right) = O( B^2 T^{-1/2} + B^3 m^{-1/2} + B^{5/2} m^{-1/4}).$$

**Proof.** The proof of Theorem E.2 mirrors that of Theorem 6.3 with the max operator replaced by the softmax operator in (E.1). We prove that the same claim of Lemma D.2 holds under Assumption 6.2, for which it suffices to upper bound (i), (ii), and (iii) in (D.7). Note that (i) does not involve the max operator. For (ii), we lay out the following lemma.

**Lemma E.3.** For any $W \in S_B$ and the constant $c_4$ in Lemma D.3, we have

$$\mathbb{E}_{\text{init}, s \sim \mu_{\text{exp}}} \left[ \left( \text{softmax} \hat{Q}_t(s, a) - \text{softmax} \hat{Q}_0(s, a; W(t)) \right)^2 \right] \leq c_4 B^3 \cdot m^{-1/2}.$$

**Proof.** The softmax operator has the following duality. For any function $Q(s, a)$, it holds that

$$\text{softmax} Q(s, a) = \max_{a \in \mathcal{A}} \mathbb{E}_{a \sim \pi(s)} \left[ Q(s, a) \right] + \beta^{-1} \cdot H(\pi(s)), \quad (E.4)$$

where $\Delta$ is the set of all probability distributions over $\mathcal{A}$ and $H(\pi(s))$ is the entropy of $\pi(s)$. Hence, we obtain

$$\left| \text{softmax} \hat{Q}_t(s, a) - \text{softmax} \hat{Q}_0(s, a; W(t)) \right| \leq \max_{\pi(s) \in \Delta} \left| \mathbb{E}_{a \sim \pi(s)} \left[ \hat{Q}_t(s, a) - \hat{Q}_0(s, a; W(t)) \right] \right|$$

$$= \max_{a \in \mathcal{A}} \left| \hat{Q}_t(s, a) - \hat{Q}_0(s, a; W(t)) \right|.$$

By applying Lemma D.3, we complete the proof of Lemma E.3. \hfill \Box

Meanwhile, note that the upper bound in (D.9) of (iii) in (D.7) still holds by Lemma F.3. Thus, the claim of Lemma D.2 holds for neural soft Q-learning. Moreover, (D.17) also holds for neural soft Q-learning under Assumption E.1. To further extend the upper bound of $\mathbb{E}_{\text{init,} \mu_{\text{exp}}} \left[ \| z(t) - \bar{z}(t) \|^2 \right]$ to neural soft Q-learning, it remains to upper bound $\text{softmax}_{a' \in \mathcal{A}} \hat{Q}_t(s', a')$, which replaces $\max_{a' \in \mathcal{A}} \hat{Q}_t(s', a')$ in (D.21), by

$$\left| \text{softmax}_{a' \in \mathcal{A}} \hat{Q}_t(s', a') \right| \leq \max_{a' \in \mathcal{A}} \left| \hat{Q}_t(s', a') \right| + \beta^{-1} \cdot \log |\mathcal{A}|,$$

where $\beta$ is the parameter of the softmax operator. Here the inequality follows from (E.4). The additional term $\beta^{-1} \cdot \log |\mathcal{A}|$ is independent of $t$ and $m$. Thus, we obtain an upper bound of the variance $\mathbb{E}_{\text{init,} \mu_{\text{exp}}} \left[ \| z(t) - \bar{z}(t) \|^2 \right]$, which is independent of $t$ and $m$. With (D.17) and Lemma D.2, the proof of Theorem E.2 follows from that of Theorem 6.3. \hfill \Box
In this section, we briefly summarize the equivalence between policy gradient algorithms and neural soft Q-learning (Schulman et al., 2017; Haarnoja et al., 2018), which implies that our results are extendable to characterize a variant of the policy gradient algorithm.

We define \( \pi_\theta \) as the Boltzmann policy corresponding to the Q-function \( Q_\theta \), which is parametrized by \( \theta \),

\[
\pi_\theta(a | s) = \pi(a | s) \cdot \exp\left(\beta \cdot (Q_\theta(s, a) - V_\theta(s))\right),
\]

where \( V_\theta(s) = \beta^{-1} \cdot \log \left( \mathbb{E}_{a \sim \pi(a | s)} \left[ \exp(\beta \cdot Q_\theta(s, a)) \right] \right). \]

(E.5)

Here \( \pi \) is the uniform policy and \( V_\theta \) is the partition function. In the context of neural soft Q-learning, we use the parametrization

\[
Q_\theta(s, a) = \tilde{Q}(x; W) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} b_r \sigma(W_r^\top x).
\]

From (E.5) we have

\[
Q_\theta(s, a) = V_\theta(s) + \beta^{-1} \cdot \log \left( \frac{\pi_\theta(a | s)}{\pi(a | s)} \right).
\]

(E.7)

In the sequel, we show that the population semigradient in soft Q-learning, which is defined in (D.2), equals a variant of population policy gradient, given that the exploration policy \( \pi_{\text{exp}} \) in (D.2) is \( \pi_\theta \). Recall that the population semigradient in soft Q-learning is given by

\[
\mathbb{E}_{(s, a, s') \sim \mu_\theta} \left[ \nabla_\theta Q_\theta(s, a) \cdot (Q_\theta(s, a) - r(s, a) - \gamma \cdot \text{softmax}_{a' \in A} Q_\theta(s', a')) \right]
\]

\[
= \mathbb{E}_{(s, a, s') \sim \mu_\theta} \left[ \nabla_\theta Q_\theta(s, a) \cdot (Q_\theta(s, a) - r(s, a) - \gamma \cdot V_\theta(s')) \right],
\]

(E.8)

where \( \mu_\theta \) is the stationary distribution of \( \pi_\theta \). For notational simplicity, we define

\[
\xi = r(s, a) - \beta^{-1} \cdot D_{KL}(\pi_\theta(a | s) \| \pi(a | s)) + \gamma \cdot V_\theta(s') - V_\theta(s).
\]

(E.9)
Plugging (E.7) and (E.9) into (E.8), we obtain
\[
\mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ \nabla_\theta Q_\theta(s, a) \cdot (Q_\theta(s, a) - r(s, a) - \gamma \cdot V_\theta(s')) \right] \\
= \mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ \nabla_\theta Q_\theta(s, a) \cdot \left( \beta^{-1} \cdot \log(\pi_\theta(a | s)/\pi(a | s)) - \beta^{-1} \cdot D_{\text{KL}}(\pi_\theta(\cdot | s) \| \pi(\cdot | s)) - \xi \right) \right].
\]
Taking gradient on both sides of (E.7), we obtain
\[
\nabla_\theta Q_\theta(s, a) = \nabla_\theta V_\theta(s) + \beta^{-1} \cdot \nabla_\theta \log(\pi_\theta(a | s)).
\]
Plugging (E.11) into the right-hand side of (E.10) yields
\[
\mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ \nabla_\theta Q_\theta(s, a) \cdot (Q_\theta(s, a) - r(s, a) - \gamma \cdot V_\theta(s')) \right] \\
= \mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ \beta^{-1} \cdot \nabla_\theta V_\theta(s) \cdot \left( \log(\pi_\theta(a | s)/\pi(a | s)) - D_{\text{KL}}(\pi_\theta(\cdot | s) \| \pi(\cdot | s)) \right) - \nabla_\theta V_\theta(s) \cdot \xi \\
+ \beta^{-2} \cdot \nabla_\theta \log(\pi_\theta(a | s)) \cdot \log(\pi_\theta(a | s)/\pi(a | s)) - \beta^{-1} \cdot \nabla_\theta \log(\pi_\theta(a | s)) \cdot D_{\text{KL}}(\pi_\theta(\cdot | s) \| \pi(\cdot | s)) \\
- \beta^{-1} \cdot \nabla_\theta \log(\pi_\theta(a | s)) \cdot \xi \right].
\]
By the definition of the KL-divergence, we have
\[
\mathbb{E}_{a \sim \pi_\theta(\cdot | s)} \left[ \log(\pi_\theta(a | s)/\pi(a | s)) \right] = D_{\text{KL}}(\pi_\theta(\cdot | s) \| \pi(\cdot | s)).
\]
Thus, on the right-hand side of (E.12), we have
\[
\mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ \beta^{-1} \cdot \nabla_\theta V_\theta(s) \left( \log(\pi_\theta(a | s)/\pi(a | s)) - D_{\text{KL}}(\pi_\theta(\cdot | s) \| \pi(\cdot | s)) \right) \right] = 0.
\]
Also, since
\[
\mathbb{E}_{a \sim \pi_\theta(\cdot | s)} \left[ \nabla_\theta \log(\pi_\theta(a | s)) \right] = \sum_{a \in A} \nabla_\theta \pi_\theta(a | s) = \nabla_\theta 1 = 0,
\]
on the right-hand side of (E.12), we have
\[
\mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ \beta^{-2} \cdot \nabla_\theta \log(\pi_\theta(a | s)) \cdot D_{\text{KL}}(\pi_\theta(\cdot | s) \| \pi(\cdot | s)) \right] = 0.
\]
Plugging (E.13) and (E.15) into (E.12), we obtain
\[
\mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ \nabla_\theta Q_\theta(s, a) \cdot (Q_\theta(s, a) - r(s, a) - \gamma \cdot V_\theta(s')) \right] \\
= \underbrace{\mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ \beta^{-2} \cdot \nabla_\theta \log(\pi_\theta(a | s)) \cdot \log(\pi_\theta(a | s)/\pi(a | s)) - \beta^{-1} \cdot \nabla_\theta \log(\pi_\theta(a | s)) \cdot \xi \right]}_{(ii)} \\
+ \underbrace{\mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ -\nabla_\theta V_\theta(s) \cdot \xi \right]}_{(i)}.
\]
We characterize (i) and (ii) on the right-hand side of (E.16). For (i), by the definition of \( \xi \) in (E.9), we have
\[
\mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ -\nabla_\theta V_\theta(s) \cdot \xi \right] \\
= \mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ -\nabla_\theta V_\theta(s) \cdot \left( r(s, a) - \beta^{-1} \cdot D_{\text{KL}}(\pi_\theta(\cdot | s) \| \pi(\cdot | s)) + \gamma \cdot V_\theta(s') - V_\theta(s) \right) \right] \\
= \mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ -\nabla_\theta \left( V_\theta(s) - \frac{\mathbb{E}_s}{\| \nabla_\theta V_\theta(s) \|} \right)^2 / 2 \right]_{\bar{s} = \pi_\theta, \bar{a} = \theta'}. 
\]

35
Here the operator $\mathcal{T}_{\text{KL}}^\pi$ is defined as

$$\mathcal{T}_{\text{KL}}^\pi V(s) = \mathbb{E} \left[ r(s, a) - \beta^{-1} \cdot D_{\text{KL}}(\tilde{\pi}(\cdot | s) \| \pi(\cdot | s)) + \gamma \cdot V(s') \middle| a \sim \tilde{\pi}(\cdot | s), s' \sim \mathcal{P}(\cdot | s, a) \right],$$

which is the Bellman evaluation operator for the value function $V^\pi(s)$ associated with the KL-regularized reward

$$r_{\text{KL}}(s, a, s') = r(s, a) - \beta^{-1} \cdot D_{\text{KL}}(\tilde{\pi}(\cdot | s) \| \pi(\cdot | s)). \quad (E.18)$$

By (E.17), $\mathbb{E}_{(s,a,s') \sim \mu_\theta}[-\nabla_\theta V_\theta(s) \cdot \xi]$ is the population semigradient for the evaluation of policy $\pi_\theta$. Now we characterize (ii) in (E.16). First, by the definition of the KL-divergence, we have

$$\nabla_\theta D_{\text{KL}}(\pi_\theta(\cdot | s) \| \pi(\cdot | s)) = \sum_{a \in A} \nabla_\theta [\pi_\theta(a | s) \cdot \log(\pi_\theta(a | s) / \pi(a | s))]$$

$$= \sum_{a \in A} \nabla_\theta \pi_\theta(a | s) \cdot \log(\pi_\theta(a | s) / \pi(a | s))$$

$$= \mathbb{E}_{a \sim \pi_\theta(\cdot | s)} [\nabla_\theta \log \pi_\theta(a | s) \cdot \log(\pi_\theta(a | s) / \pi(a | s))].$$

Here the second equality follows from (E.14). Hence, (ii) in (E.16) takes the form

$$\mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ \beta^{-2} \cdot \nabla_\theta \log(\pi_\theta(a | s)) \cdot \log(\pi_\theta(a | s) / \pi(a | s)) - \beta^{-1} \cdot \nabla_\theta \log(\pi_\theta(a | s)) \cdot \xi \right]$$

$$= -\beta^{-1} \cdot \mathbb{E}_{(s,a,s') \sim \mu_\theta} \left[ \nabla_\theta \log(\pi_\theta(a | s)) \cdot \xi - \nabla_\theta \left( \beta^{-1} \cdot D_{\text{KL}}(\pi_\theta(\cdot | s) \| \pi(\cdot | s)) \right) \right]. \quad (ii.a)$$

We show that (ii) in (E.16) is the population policy gradient. For (ii).a, note that $\xi$ defined in (E.9) is an unbiased estimator of the advantage function $Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s)$ associated with the KL-regularized reward $r_{\text{KL}}$ defined in (E.18). If we denote by $J_{\text{KL}}(\pi_\theta)$ the expected total reward, then (ii).a is an estimator of the population policy gradient $\nabla_\theta J_{\text{KL}}(\pi_\theta)$. For (ii).b, it is the gradient of the entropy regularization

$$\beta^{-1} \cdot \mathbb{E}_{s \sim \mu_\theta} \left[ D_{\text{KL}}(\pi_\theta(\cdot | s) \| \pi(\cdot | s)) \right] = \beta^{-1} \cdot \mathbb{E}_{s \sim \mu_\theta} [H(\pi_\theta(\cdot | s))].$$

Therefore, we recover the policy gradient update in the Q-learning updating scheme. Combining (i) and (ii) in (E.16), we obtain a variant of the policy gradient algorithm, which is connected with the soft actor-critic algorithm (Haarnoja et al., 2018). Hence, our global convergence of neural soft Q-learning extends to a variant of the actor-critic algorithm. See Algorithm 4 for a detailed description of such an algorithm with $\hat{Q}_\theta$ parametrized by a two-layer neural network. In parallel with $V_\theta(s)$ in (E.6), we define

$$\hat{V}(s; W) = \beta^{-1} \cdot \log \left( \mathbb{E}_{a \sim \pi(\cdot | s)} \left[ \exp(\beta \cdot \hat{Q}(s, a; W)) \right] \right).$$

### F Auxiliary Lemmas

Under Assumption 4.3, we establish the following auxiliary lemmas on the random initialization $W(0)$ and the stationary distribution $\mu$, which plays a key role in quantifying the error of local linearization.
Algorithm 4 Neural Soft Actor-Critic

1: Initialization: \( b_r \sim \text{Unif}(-1,1) \), \( W_r(0) \sim N(0,I_d/d) \) \((r \in [m])\), \( W = W(0) \)

2: For \( t = 0 \) to \( T - 2 \):

3: Policy Update: \( \pi_t(a \mid s) \propto \pi(a \mid s) \cdot \exp(\beta \cdot \hat{Q}(s,a;W(t))) \)

4: Sample a tuple \((s,a,r,s')\) from the stationary distribution \( \mu_t \) of policy \( \pi_t \)

5: Reward regularization: \( r_{\text{KL}} = r - \beta^{-1} \cdot D_{\text{KL}}(\pi_t(\cdot \mid s) \parallel \pi(\cdot \mid s)) \)

6: Bellman residual calculation: \( \xi = r_{\text{KL}} + \gamma \hat{V}(s';W(t)) - \hat{V}(s;W(t)) \)

7: Actor Update:

8: \[ \hat{W}(t+1) \leftarrow W(t) + \eta \cdot (\beta^{-1} \cdot \xi \cdot \nabla W \log \pi_t(a \mid s) - \beta^{-2} \cdot \nabla W D_{\text{KL}}(\pi_t(\cdot \mid s) \parallel \pi(\cdot \mid s))) \]

9: Critic update: \( \hat{W}'(t+1) \leftarrow \hat{W}(t+1) + \eta \cdot \xi \cdot \nabla W \hat{V}(s;W(t)) \)

10: Projection: \( W(t+1) \leftarrow \arg \min_{W \in S_B} \| W - \hat{W}'(t+1) \|_2 \)

11: Averaging: \( \overline{W} \leftarrow \frac{t+1}{t+2} \cdot \hat{W} + \frac{1}{t+2} \cdot W(t+1) \)

12: End For

13: Output: \( \hat{Q}_{\text{out}}(\cdot) \leftarrow \hat{Q}(\cdot;\overline{W}) \), \( \pi_{\text{out}}(a \mid s) \propto \pi(a \mid s) \cdot \exp(\beta \cdot \hat{Q}_{\text{out}}(s,a)) \)

Lemma F.1. There exists a constant \( c_1 > 0 \) such that for any random vector \( W \) with \( \| W - W(0) \|_2 \leq B \), it holds that

\[
\mathbb{E}_{\text{init,} \mu} \left[ \frac{1}{m} \sum_{r=1}^{m} \mathbb{I} \{ |W_r(0)^T x| \leq \| W_r - W_r(0) \|_2 \} \right] \leq c_1 B \cdot m^{-1/2}. \tag{F.1}
\]

Proof. By Assumption 4.3, we have

\[
\mathbb{E}_{\text{init,} \mu} \left[ \frac{1}{m} \sum_{r=1}^{m} \mathbb{I} \{ |W_r(0)^T x| \leq \| W_r - W_r(0) \|_2 \} \right] 
\leq \mathbb{E}_{\text{init}} \left[ \frac{1}{m} \sum_{r=1}^{m} c_0 \cdot \| W_r - W_r(0) \|_2/\| W_r(0) \|_2 \right]. \tag{F.2}
\]

Applying Hölder’s inequality to the right-hand side, we obtain

\[
\mathbb{E}_{\text{init,} \mu} \left[ \frac{1}{m} \sum_{r=1}^{m} \mathbb{I} \{ |W_r(0)^T x| \leq \| W_r - W_r(0) \|_2 \} \right] 
\leq c_0/m \cdot \mathbb{E}_{\text{init}} \left[ \left( \sum_{r=1}^{m} \| W_r - W_r(0) \|_2 \right)^{1/2} \cdot \left( \sum_{r=1}^{m} \frac{1}{\| W_r(0) \|_2^2} \right)^{1/2} \right] 
\leq c_0 B \cdot m^{-1/2} \cdot \mathbb{E}_{w \sim N(0,I_d/d)} \left[ \frac{1}{\| w \|_2^2} \right]^{1/2}, \tag{F.3}
\]

where the second inequality follows from

\[
\mathbb{E}_{\text{init}} \left[ \left( \sum_{r=1}^{m} \frac{1}{\| W_r(0) \|_2^2} \right)^{1/2} \right] \leq m \cdot \mathbb{E}_{w \sim N(0,I_d/d)} \left[ \frac{1}{\| w \|_2^2} \right]^{1/2} = \sqrt{m} \cdot \mathbb{E}_{w \sim N(0,I_d/d)} \left[ \frac{1}{\| w \|_2^2} \right]^{1/2}. \tag{F.4}
\]

Setting \( c_1 = c_0 \cdot \mathbb{E}_{w \sim N(0,I_d/d)} \left[ \frac{1}{\| w \|_2^2} \right]^{1/2} \), we complete the proof of Lemma F.1.

\[\square\]
Lemma F.2. There exists a constant $c_2 > 0$ such that for any random vector $W$ with $\|W - W(0)\|_2 \leq B$, it holds that
\[
\mathbb{E}_{init} \left[ \mathbb{E}_\mu \left( \tilde{Q}_0(x)^2 \right) \cdot \mathbb{E}_\mu \left[ \frac{1}{m} \sum_{r=1}^{m} \mathbb{I} \{ |W_r(0)^\top x| \leq \|W_r - W_r(0)\|_2 \} \right] \right] \leq c_2 B \cdot m^{-1/2}. \tag{F.5}
\]

Proof. By the definition of $\tilde{Q}_0(x) = \tilde{Q}_0(x; W(0))$ in (4.7), we have
\[
\mathbb{E}_\mu [\tilde{Q}_0(x)^2] = 1/m \cdot \mathbb{E}_\mu \left[ \sum_{r=1}^{m} \sigma(W_r(0)^\top x)^2 + \sum_{r \neq s} b_r \sigma(W_r(0)^\top x) \sigma(W_s(0)^\top x) \right].
\]

Following the same derivation of (F.2) and (F.3), we have
\[
\mathbb{E}_{init} \left[ \mathbb{E}_\mu \left( \tilde{Q}_0(x)^2 \right) \cdot \mathbb{E}_\mu \left[ \frac{1}{m} \sum_{r=1}^{m} \mathbb{I} \{ |W_r(0)^\top x| \leq \|W_r - W_r(0)\|_2 \} \right] \right] 
\leq \mathbb{E}_{init} \left[ 1/m \cdot \mathbb{E}_\mu \left[ \sum_{r=1}^{m} \sigma(W_r(0)^\top x)^2 + \sum_{r \neq s} b_r b_s \sigma(W_r(0)^\top x) \sigma(W_s(0)^\top x) \right] \right] 
\cdot c_0/m \cdot \left( \sum_{r=1}^{m} \|W_r - W_r(0)\|_2^2 \right)^{1/2} \cdot \left( \sum_{r=1}^{m} \frac{1}{\|W_r(0)\|_2^2} \right)^{1/2}.
\]

Note that $b_r$ and $b_s$ are independent of $W(0)$ and $\mathbb{E}_{init}[b_r b_s] = 0$. Thus, we obtain
\[
\mathbb{E}_{init} \left[ \mathbb{E}_\mu \left( \tilde{Q}_0(x)^2 \right) \cdot \mathbb{E}_\mu \left[ \frac{1}{m} \sum_{r=1}^{m} \mathbb{I} \{ |W_r(0)^\top x| \leq \|W_r - W_r(0)\|_2 \} \right] \right] 
\leq c_0 B/m^2 \cdot \mathbb{E}_{init} \left[ \mathbb{E}_\mu \left[ \sum_{r=1}^{m} \sigma(W_r(0)^\top x)^2 \right] \cdot \left( \sum_{r=1}^{m} \frac{1}{\|W_r(0)\|_2^2} \right)^{1/2} \right].
\]

By the definition of $\sigma(W_r(0)^\top x)$ and the fact that $\|x\|_2 = 1$, we have
\[
\mathbb{E}_\mu \left[ \sum_{r=1}^{m} \sigma(W_r(0)^\top x)^2 \right] \leq \sum_{r=1}^{m} \|W_r(0)\|_2^2.
\]

Hence, it holds that
\[
\mathbb{E}_{init} \left[ \mathbb{E}_\mu \left( \tilde{Q}_0(x)^2 \right) \cdot \mathbb{E}_\mu \left[ \frac{1}{m} \sum_{r=1}^{m} \mathbb{I} \{ |W_r(0)^\top x| \leq \|W_r - W_r(0)\|_2 \} \right] \right] 
\leq c_0 B/m^2 \cdot \mathbb{E}_{init} \left[ \left( \sum_{r=1}^{m} \|W_r(0)\|_2^2 \right) \cdot \left( \sum_{r=1}^{m} \frac{1}{\|W_r(0)\|_2^2} \right)^{1/2} \right] 
\leq c_0 B/m^2 \cdot \mathbb{E}_{init} \left[ \left( \sum_{r=1}^{m} \|W_r(0)\|_2^2 \right)^{1/2} \cdot \left( \sum_{r=1}^{m} \frac{1}{\|W_r(0)\|_2^2} \right)^{1/2} \right]. \tag{F.6}
\]

By (F.4) and the fact that
\[
\mathbb{E}_{init} \left[ \left( \sum_{r=1}^{m} \|W_r(0)\|_2^2 \right)^{1/2} \right] = m \cdot \mathbb{E}_{w \sim N(0, I_d/d)}[\|w\|_2^2] + m(m - 1) \cdot \mathbb{E}_{w \sim N(0, I_d/d)}[\|w\|_2^2] = O(m^2),
\]

38
the right-hand side of (F.6) is $O(Bm^{-1/2})$. Setting
\[ c_2 = c_0 \cdot \left( \mathbb{E}_{w \sim N(0,I_d/d)}[\|w\|^3] + \mathbb{E}_{w \sim N(0,I_d/d)}[\|w\|^2] \right)^{1/2} \cdot \mathbb{E}_{w \sim N(0,I_d/d)}[1/\|w\|^2]^{1/2}, \]
we complete the proof of Lemma F.2.

\[ \square \]

**Lemma F.3.** For any random vector $W$ with $\|W - W(0)\|_2 \leq B$, we have
\[ \mathbb{E}_{\text{init}} \left[ \mathbb{E}_{s \sim \mu} \max_{a \sim A} \bar{Q}_0(s,a)^2 \right] \cdot \mathbb{E}_\mu \left[ \frac{1}{m} \sum_{r=1}^{m} \mathbb{1}\{ |W_r(0)^\top x| \leq \|W_r - W_r(0)\|_2 \} \right] = O(Bm^{-1/2}), \]
\[ \mathbb{E}_{\text{init}} \left[ \mathbb{E}_{s \sim \mu} \text{softmax} \max_{a \sim A} \bar{Q}_0(s,a)^2 \right] \cdot \mathbb{E}_\mu \left[ \frac{1}{m} \sum_{r=1}^{m} \mathbb{1}\{ |W_r(0)^\top x| \leq \|W_r - W_r(0)\|_2 \} \right] = O(Bm^{-1/2}). \]

**Proof.** The proof mirrors that of Lemma F.2. We utilize the fact that $A$ is finite, so that the expectation of the maximum can be upper bounded by a finite sum of expectations,
\[ \mathbb{E}_{s \sim \mu} \max_{a \in A} \bar{Q}_0(s,a)^2 \leq \mathbb{E}_{s \sim \mu} \left[ \sum_{a \in A} \bar{Q}_0(s,a)^2 \right] \leq \sum_{a \in A} \mathbb{E}_{s \sim \mu} \bar{Q}_0(s,a)^2. \]
\[ \mathbb{E}_{s \sim \mu} \text{softmax} \max_{a \in A} \bar{Q}_0(s,a)^2 \leq \sum_{a \in A} \mathbb{E}_{s \sim \mu} \bar{Q}_0(s,a)^2 + \beta^{-1} \cdot \log |A|. \]

For each expectation $\mathbb{E}_{s \sim \mu} \bar{Q}_0(s,a)^2$, note that the distribution of $(s,a)$ is independent of the initialization. Hence, the same proof of Lemma F.2 is applicable, as $\mathbb{E}_{s \sim \mu} \bar{Q}_0(s,a)^2$ plays the same role of $\mathbb{E}_\mu \bar{Q}_0(x)^2$. Thus, we obtain the same upper bound in Lemmas F.1 and F.2 except for an extra factor involving $|A|$, which however does not change the order of $m$. \[ \square \]

**F.1 Proof of Lemma D.3**

**Proof.** In the proof of Lemma 5.1, we show that, for any $s \in S$ and $a \in A$,
\[ |\hat{Q}_t(s,a) - \bar{Q}_0(s,a;W(t))|^2 \leq \frac{4B^2}{m} \sum_{r=1}^{m} \mathbb{1}\{ |W_r(0)^\top x| \leq \|W_r(t) - W_r(0)\|_2 \}. \]
Taking maximum over $a$, we obtain
\[ \max_{a \in A} |\hat{Q}_t(s,a) - \bar{Q}_0(s,a;W(t))|^2 \leq \frac{4B^2}{m} \sum_{r=1}^{m} \max_{a \in A} \mathbb{1}\{ |W_r(0)^\top (s,a)| \leq \|W_r(t) - W_r(0)\|_2 \}. \]
Taking expectation with respect to the random initialization and the stationary distribution of $s$, we obtain
\[ \mathbb{E}_{\text{init},s \sim \mu_{\exp}} \left[ \max_{a \in A} |\hat{Q}_t(s,a) - \bar{Q}_0(s,a;W(t))|^2 \right] \leq \mathbb{E}_{\text{init},s \sim \mu_{\exp}} \left[ \frac{4B^2}{m} \sum_{r=1}^{m} \max_{a \in A} \mathbb{1}\{ |W_r(0)^\top (s,a)| \leq \|W_r(t) - W_r(0)\|_2 \} \right]. \]
By Assumption 6.2, it holds that

$$\mathbb{E}_{\text{init}, s \sim \mu_{\text{exp}}}[\max_{a \in A}|Q_t(s, a) - \tilde{Q}_0(s, a; W(t))|^2] \leq \mathbb{E}_{\text{init}}\left[\frac{4B^2}{m} \sum_{r=1}^{m} c_3 \cdot \|W_r(t) - W_r(0)\|_2 / \|W_r(0)\|_2\right].$$

Applying Hölder’s inequality to the right-hand side, we obtain

$$\mathbb{E}_{\text{init}, s \sim \mu_{\text{exp}}}[\max_{a \in A}|Q_t(s, a) - \tilde{Q}_0(s, a; W(t))|^2]$$

$$\leq \frac{4B^2 c_3}{m} \cdot \mathbb{E}_{\text{init}}\left[\sum_{r=1}^{m} \|W_r - W_r(0)\|_2^2 \right]^{1/2} \cdot \left(\sum_{r=1}^{m} \frac{1}{\|W_r(0)\|_2^2}\right)^{1/2}$$

$$\leq 4B^3 c_3 \cdot m^{-1/2} \cdot \mathbb{E}_{w \sim \mathcal{N}(0, I_{d/d})}[1 / \|w\|_2^2]^{1/2}. $$

Setting $c_4 = 4c_3 \cdot \mathbb{E}_{w \sim \mathcal{N}(0, I_{d/d})}[1 / \|w\|_2^2]^{1/2}$, we finish the proof of Lemma D.3. □

40