Can Temporal-Difference and Q-Learning Learn Representation? A Mean-Field Analysis

Yufeng Zhang
Northwestern University
Evanston, IL 60208
yufengzhang2023@u.northwestern.edu

Qi Cai
Northwestern University
Evanston, IL 60208
qicai2022@u.northwestern.edu

Zhuoran Yang
Princeton University
Princeton, NJ 08544
zy6@princeton.edu

Yongxin Chen
Georgia Institute of Technology
Atlanta, GA 30332
yongchen@gatech.edu

Zhaoran Wang
Northwestern University
Evanston, IL 60208
zhaoranwang@gmail.com

Abstract

Temporal-difference and Q-learning play a key role in deep reinforcement learning, where they are empowered by expressive nonlinear function approximators such as neural networks. At the core of their empirical successes is the learned feature representation, which embeds rich observations, e.g., images and texts, into the latent space that encodes semantic structures. Meanwhile, the evolution of such a feature representation is crucial to the convergence of temporal-difference and Q-learning.

In particular, temporal-difference learning converges when the function approximator is linear in a feature representation, which is fixed throughout learning, and possibly diverges otherwise. We aim to answer the following questions:

When the function approximator is a neural network, how does the associated feature representation evolve? If it converges, does it converge to the optimal one?

We prove that, utilizing an overparameterized two-layer neural network, temporal-difference and Q-learning globally minimize the mean-squared projected Bellman error at a sublinear rate. Moreover, the associated feature representation converges to the optimal one, generalizing the previous analysis of [21] in the neural tangent kernel regime, where the associated feature representation stabilizes at the initial one. The key to our analysis is a mean-field perspective, which connects the evolution of a finite-dimensional parameter to its limiting counterpart over an infinite-dimensional Wasserstein space. Our analysis generalizes to soft Q-learning, which is further connected to policy gradient.

1 Introduction

Deep reinforcement learning achieves phenomenal empirical successes, especially in challenging applications where an agent acts upon rich observations, e.g., images and texts. Examples include video gaming [56], visuomotor manipulation [51], and language generation [39]. Such empirical successes are empowered by expressive nonlinear function approximators such as neural networks, which are used to parameterize both policies (actors) and value functions (critics) [46]. In particular, the neural network learned from interacting with the environment induces a data-dependent feature representation, which embeds rich observations into a latent space encoding semantic structures

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

[12, 40, 49, 75]. In contrast, classical reinforcement learning mostly relies on a handcrafted feature representation that is fixed throughout learning [65].

In this paper, we study temporal-difference (TD) [64] and Q-learning [71], two of the most prominent algorithms in deep reinforcement learning, which are further connected to policy gradient [73] through its equivalence to soft Q-learning [37, 57, 58, 61]. In particular, we aim to characterize how an overparameterized two-layer neural network and its induced feature representation evolve in TD and Q-learning, especially their rate of convergence and global optimality. A fundamental obstacle, however, is that such an evolving feature representation possibly leads to the divergence of TD and Q-learning. For example, TD converges when the value function approximator is linear in a feature representation, which is fixed throughout learning, and possibly diverges otherwise [10, 18, 67].

To address such an issue of divergence, nonlinear gradient TD [15] explicitly linearizes the value function approximator locally at each iteration, that is, using its gradient with respect to the parameter as an evolving feature representation. Although nonlinear gradient TD converges, it is unclear whether the attained solution is globally optimal. On the other hand, when the value function approximator in TD is an overparameterized multi-layer neural network, which is required to be properly scaled, such a feature representation stabilizes at the initial one [21], making the explicit local linearization in nonlinear gradient TD unnecessary. Moreover, the implicit local linearization enabled by overparameterization allows TD (and Q-learning) to converge to the globally optimal solution. However, such a required scaling, also known as the neural tangent kernel (NTK) regime [43], effectively constrains the evolution of the induced feature presentation to an infinitesimal neighborhood of the initial one, which is not data-dependent.

Contribution. Going beyond the NTK regime, we prove that, when the value function approximator is an overparameterized two-layer neural network, TD and Q-learning globally minimize the mean-squared projected Bellman error (MSPBE) at a sublinear rate. Moreover, in contrast to the NTK regime, the induced feature representation is able to deviate from the initial one and subsequently evolve into the globally optimal one, which corresponds to the global minimizer of the MSPBE. We further extend our analysis to soft Q-learning, which is connected to policy gradient.

The key to our analysis is a mean-field perspective, which allows us to associate the evolution of a finite-dimensional parameter with its limiting counterpart over an infinite-dimensional Wasserstein space [4, 5, 68, 69]. Specifically, by exploiting the permutation invariance of the parameter, we associate the neural network and its induced feature representation with an empirical distribution, which, at the infinite-width limit, further corresponds to a population distribution. The evolution of such a population distribution is characterized by a partial differential equation (PDE) known as the continuity equation. In particular, we develop a generalized notion of one-point monotonicity [38], which is tailored to the Wasserstein space, especially the first variation formula therein [5], to characterize the evolution of such a PDE solution, which, by a discretization argument, further quantifies the evolution of the induced feature representation.

Related Work. When the value function approximator is linear, the convergence of TD is extensively studied in both continuous-time [16, 17, 42, 47, 67] and discrete-time [14, 29, 48, 63] settings. See [31] for a detailed survey. Also, when the value function approximator is linear, [25, 55, 78] study the convergence of Q-learning. When the value function approximator is nonlinear, TD possibly diverges [10, 18, 67]. [15] propose nonlinear gradient TD, which converges but only to a locally optimal solution. See [13, 36] for a detailed survey. When the value function approximator is an overparameterized multi-layer neural network, [21] prove that TD converges to the globally optimal solution in the NTK regime. See also the independent work of [1, 19, 20, 62], where the state space is required to be finite. In contrast to the previous analysis in the NTK regime, our analysis allows TD to attain a data-dependent feature representation that is globally optimal.

Meanwhile, our analysis is related to the recent breakthrough in the mean-field analysis of stochastic gradient descent (SGD) for the supervised learning of an overparameterized two-layer neural network [23, 27, 34, 35, 44, 53, 54, 72]. See also the previous analysis in the NTK regime [2, 3, 7–9, 22, 24, 26, 30, 32, 33, 43, 45, 50, 52, 76, 77]. Specifically, the previous mean-field analysis casts SGD as the Wasserstein gradient flow of an energy functional, which corresponds to the objective function in supervised learning. In contrast, TD follows the stochastic semigradient of the MSPBE [65], which is biased. As a result, there does not exist an energy functional for casting TD as its Wasserstein

gradient flow. Instead, our analysis combines a generalized notion of one-point monotonicity [38] and the first variation formula in the Wasserstein space [5], which is of independent interest.

Notations. We denote by $\mathscr{B}(\mathcal{X})$ the Borel σ-algebra over the space $\mathcal{X}$. Let $\mathscr{P}(\mathcal{X})$ be the set of Borel probability measures over the measurable space $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$. We denote by $[N] = \{1, 2, \ldots, N\}$ for any $N \in \mathbb{N}_+$. Also, we denote by $\mathcal{B}^n(x; r) = \{y \in \mathbb{R}^n \,|\, \|y - x\| \leq r\}$ the closed ball in $\mathbb{R}^n$. Given a curve $\rho : \mathbb{R} \to \mathcal{X}$, we denote by $\rho'_s = \partial_t \rho_t \,|_{\,t=s}$ its derivative with respect to the time. For a function $f : \mathcal{X} \to \mathbb{R}$, we denote by $\mathrm{Lip}(f) = \sup_{x,y \in \mathcal{X}, x \neq y} |f(x) - f(y)| / \|x - y\|$ its Lipschitz constant. For an operator $F : \mathcal{X} \to \mathcal{X}$ and a measure $\mu \in \mathscr{P}(\mathcal{X})$, we denote by $F_\sharp \mu = \mu \circ F^{-1}$ the push forward of μ through F. We denote by D_{KL} and D_{χ^2} the Kullback-Leibler (KL) divergence and the χ^2 divergence, respectively.

2 Background

2.1 Policy Evaluation

We consider a Markov decision process $(\mathcal{S}, \mathcal{A}, P, R, \gamma, \mathcal{D}_0)$, where $\mathcal{S} \subseteq \mathbb{R}^{d_1}$ is the state space, $\mathcal{A} \subseteq \mathbb{R}^{d_2}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathcal{S})$ is the transition kernel, $R : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathbb{R})$ is the reward distribution, $\gamma \in (0, 1)$ is the discount factor, and $\mathcal{D}_0 \in \mathscr{P}(\mathcal{S})$ is the initial state distribution. An agent following a policy $\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})$ interacts with the environment in the following manner. At a state s_t, the agent takes an action a_t according to $\pi(\cdot \,|\, s_t)$ and receives from the environment a random reward r_t following $R(\cdot \,|\, s_t, a_t)$. Then, the environment transits into the next state s_{t+1} according to $P(\cdot \,|\, s_t, a_t)$. We measure the performance of a policy π via the expected cumulative reward $J(\pi)$, which is defined as follows,

$$J(\pi) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \cdot r_t \,\Big|\, s_0 \sim \mathcal{D}_0, a_t \sim \pi(\cdot \,|\, s_t), r_t \sim R(\cdot \,|\, s_t, a_t), s_{t+1} \sim P(\cdot \,|\, s_t, a_t)\Big]. \tag{2.1}$$

In policy evaluation, we are interested in the state-action value function (Q-function) $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which is defined as follows,

$$Q^\pi(s, a) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \cdot r_t \,\Big|\, s_0 = s, a_0 = a, a_t \sim \pi(\cdot \,|\, s_t), r_t \sim R(\cdot \,|\, s_t, a_t), s_{t+1} \sim P(\cdot \,|\, s_t, a_t)\Big].$$

We learn the Q-function by minimizing the mean-squared Bellman error (MSBE), which is defined as follows,

$$\mathrm{MSBE}(Q) = \frac{1}{2} \cdot \mathbb{E}_{(s,a) \sim \mathcal{D}}\Big[\big(Q(s, a) - \mathcal{T}^\pi Q(s, a)\big)^2\Big].$$

Here $\mathcal{D} \in \mathscr{P}(\mathcal{S} \times \mathcal{A})$ is the stationary distribution induced by the policy π of interest and $\mathcal{T}^\pi$ is the corresponding Bellman operator, which is defined as follows,

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

However, $\mathcal{T}^\pi Q$ may be not representable by a given function class $\mathcal{F}$. Hence, we turn to minimizing a surrogate of the MSBE over $Q \in \mathcal{F}$, namely the mean-squared projected Bellman error (MSPBE), which is defined as follows,

$$\mathrm{MSPBE}(Q) = \frac{1}{2} \cdot \mathbb{E}_{(s,a) \sim \mathcal{D}}\Big[\big(Q(s, a) - \Pi_{\mathcal{F}} \mathcal{T}^\pi Q(s, a)\big)^2\Big], \tag{2.2}$$

where $\Pi_{\mathcal{F}}$ is the projection onto $\mathcal{F}$ with respect to the $\mathcal{L}_2(\mathcal{D})$-norm. The global minimizer of the MSPBE is the fixed point solution to the projected Bellman equation $Q = \Pi_{\mathcal{F}} \mathcal{T}^\pi Q$.

In temporal-difference (TD) learning, corresponding to the MSPBE defined in (2.2), we parameterize the Q-function with $\widehat{Q}(\cdot; \theta)$ and update the parameter θ via stochastic semigradient descent [65],

$$\theta' = \theta - \epsilon \cdot \big(\widehat{Q}(s, a; \theta) - r - \gamma \cdot \widehat{Q}(s', a'; \theta)\big) \cdot \nabla_\theta \widehat{Q}(s, a; \theta), \tag{2.3}$$

where $\epsilon > 0$ is the stepsize and $(s, a, r, s', a') \sim \widetilde{\mathcal{D}}$. Here we denote by $\widetilde{\mathcal{D}} \in \mathscr{P}(\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S} \times \mathcal{A})$ the distribution of (s, a, r, s', a'), where $(s, a) \sim \mathcal{D}$, $r \sim R(\cdot \,|\, s, a)$, $s' \sim P(\cdot \,|\, s, a)$, and $a' \sim \pi(\cdot \,|\, s')$.

2.2 Wasserstein Space

Let $\Theta \subseteq \mathbb{R}^D$ be a Polish space. We denote by $\mathscr{P}_2(\Theta) \subseteq \mathscr{P}(\Theta)$ the set of probability measures with finite second moments. Then, the Wasserstein-2 distance between $\mu, \nu \in \mathscr{P}_2(\Theta)$ is defined as follows,

$$\mathcal{W}_2(\mu, \nu) = \inf\Big\{\mathbb{E}\big[\|X - Y\|^2\big]^{1/2} \,\Big|\, \mathrm{law}(X) = \mu, \mathrm{law}(Y) = \nu\Big\}, \tag{2.4}$$

where the infimum is taken over the random variables X and Y on Θ. Here we denote by $\mathrm{law}(X)$ the distribution of a random variable X. We call $\mathcal{M} = (\mathscr{P}_2(\Theta), \mathcal{W}_2)$ the Wasserstein space, which is an infinite-dimensional manifold [69]. In particular, such a structure allows us to write any tangent vector at $\mu \in \mathcal{M}$ as ρ_0' for a corresponding curve $\rho : [0, 1] \to \mathscr{P}_2(\Theta)$ that satisfies $\rho_0 = \mu$. Here ρ_0' denotes $\partial_t \rho_t \,|_{\,t=0}$. Specifically, under certain regularity conditions, for any curve $\rho : [0, 1] \to \mathscr{P}_2(\Theta)$, the continuity equation $\partial_t \rho_t = -\operatorname{div}(\rho_t v_t)$ corresponds to a vector field $v : [0, 1] \times \Theta \to \mathbb{R}^D$, which endows the infinite-dimensional manifold $\mathscr{P}_2(\Theta)$ with a weak Riemannian structure in the following sense [69]. Given any tangent vectors u and $\widetilde{u}$ at $\mu \in \mathcal{M}$ and the corresponding vector fields $v, \widetilde{v}$, which satisfy $u + \operatorname{div}(\mu v) = 0$ and $\widetilde{u} + \operatorname{div}(\mu \widetilde{v}) = 0$, respectively, we define the inner product of u and $\widetilde{u}$ as follows,

$$\langle u, \widetilde{u} \rangle_\mu = \int \langle v, \widetilde{v} \rangle \,\mathrm{d}\mu, \tag{2.5}$$

which yields a Riemannian metric. Here $\langle v, \widetilde{v} \rangle$ is the inner product on $\mathbb{R}^D$. Such a Riemannian metric further induces a norm $\|u\|_\mu = \langle u, u \rangle_\mu^{1/2}$ for any tangent vector $u \in T_\mu \mathcal{M}$ at any $\mu \in \mathcal{M}$, which allows us to write the Wasserstein-2 distance defined in (2.4) as follows,

$$\mathcal{W}_2(\mu, \nu) = \inf\left\{\left(\int_0^1 \|\rho_t'\|_{\rho_t}^2 \,\mathrm{d}t\right)^{1/2} \,\middle|\, \rho : [0, 1] \to \mathcal{M}, \rho_0 = \mu, \rho_1 = \nu\right\}. \tag{2.6}$$

Here ρ_s' denotes $\partial_t \rho_t \,|_{\,t=s}$ for any $s \in [0, 1]$. In particular, the infimum in (2.6) is attained by the geodesic $\widetilde{\rho} : [0, 1] \to \mathscr{P}_2(\Theta)$ connecting $\mu, \nu \in \mathcal{M}$. Moreover, the geodesics on $\mathcal{M}$ are constant-speed, that is,

$$\|\widetilde{\rho}_t'\|_{\widetilde{\rho}_t} = \mathcal{W}_2(\mu, \nu), \quad \forall t \in [0, 1]. \tag{2.7}$$

3 Temporal-Difference Learning

For notational simplicity, we write $\mathbb{R}^d = \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $\mathcal{X} = \mathcal{S} \times \mathcal{A} \subseteq \mathbb{R}^d$, and $x = (s, a) \in \mathcal{X}$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Parameterization of Q-Function. We consider the parameter space $\mathbb{R}^D$ and parameterize the Q-function with the following two-layer neural network,

$$\widehat{Q}(x; \theta^{(m)}) = \frac{\alpha}{m} \sum_{i=1}^m \sigma(x; \theta_i), \tag{3.1}$$

where $\theta^{(m)} = (\theta_1, \ldots, \theta_m) \in \mathbb{R}^{D \times m}$ is the parameter, $m \in \mathbb{N}_+$ is the width, $\alpha > 0$ is the scaling parameter, and $\sigma : \mathbb{R}^d \times \mathbb{R}^D \to \mathbb{R}$ is the activation function. Assuming the activation function in (3.1) takes the form of $\sigma(x; \theta) = b \cdot \widetilde{\sigma}(x; w)$ for $\theta = (w, b)$, we recover the standard form of two-layer neural networks, where $\widetilde{\sigma}$ is the rectified linear unit or the sigmoid function. Such a parameterization is also used in [23, 26, 53]. For $\{\theta_i\}_{i=1}^m$ independently sampled from a distribution $\rho \in \mathscr{P}(\mathbb{R}^D)$, we have the following infinite-width limit of (3.1),

$$Q(x; \rho) = \alpha \cdot \int \sigma(x; \theta) \,\mathrm{d}\rho(\theta). \tag{3.2}$$

For the empirical distribution $\widehat{\rho}^{(m)} = m^{-1} \cdot \sum_{i=1}^m \delta_{\theta_i}$ corresponding to $\{\theta_i\}_{i=1}^m$, we have that $Q(x; \widehat{\rho}^{(m)}) = \widehat{Q}(x; \theta^{(m)})$.

TD Dynamics. In what follows, we consider the TD dynamics,

$$\begin{aligned}&\theta_i(k+1)\\&\quad=\theta_i(k)-\eta\epsilon\cdot\alpha\cdot\Big(\widehat{Q}\big(x_k;\theta^{(m)}(k)\big)-r_k-\gamma\cdot\widehat{Q}\big(x_k';\theta^{(m)}(k)\big)\Big)\cdot\nabla_\theta\sigma\big(x_k;\theta_i(k)\big),\end{aligned}\tag{3.3}$$

where $i\in[m]$, $(x_k,r_k,x_k')\sim\widetilde{\mathcal{D}}$, and $\epsilon>0$ is the stepsize with the scaling parameter $\eta>0$. Without loss of generality, we assume that (x_k,r_k,x_k') is independently sampled from $\widetilde{\mathcal{D}}$, while our analysis straightforwardly generalizes to the setting of Markov sampling [14, 74, 78]. For an initial distribution $\rho_0\in\mathscr{P}(\mathbb{R}^D)$, we initialize $\{\theta_i\}_{i=1}^m$ as $\theta_i\overset{\text{i.i.d.}}{\sim}\rho_0$ $(i\in[m])$. See Algorithm 1 in §A for a detailed description.

Mean-Field Limit. Corresponding to $\epsilon\to0^+$ and $m\to\infty$, the continuous-time and infinite-width limit of the TD dynamics in (3.3) is characterized by the following partial differential equation (PDE) with ρ_0 as the initial distribution,

$$\partial_t\rho_t=-\eta\cdot\operatorname{div}\big(\rho_t\cdot g(\cdot;\rho_t)\big).\tag{3.4}$$

Here $g(\cdot;\rho_t):\mathbb{R}^D\to\mathbb{R}^D$ is a vector field, which is defined as follows,

$$g(\theta;\rho)=-\alpha\cdot\mathbb{E}_{(x,r,x')\sim\widetilde{\mathcal{D}}}\Big[\big(Q(x;\rho)-r-\gamma\cdot Q(x';\rho)\big)\cdot\nabla_\theta\sigma(x;\theta)\Big].\tag{3.5}$$

Note that (3.4) holds in the sense of distributions [5]. See [6, 53, 54] for the existence, uniqueness, and regularity of the PDE solution ρ_t in (3.4). In the sequel, we refer to the continuous-time and infinite-width limit with $\epsilon\to0^+$ and $m\to\infty$ as the mean-field limit. Let $\widehat{\rho}_k^{(m)}=m^{-1}\cdot\sum_{i=1}^m\delta_{\theta_i(k)}$ be the empirical distribution corresponding to $\{\theta_i(k)\}_{i=1}^m$ in (3.3). The following proposition proves that the PDE solution ρ_t in (3.4) well approximates the TD dynamics $\theta^{(m)}(k)$ in (3.3).

Proposition 3.1 (Informal Version of Proposition D.1)**.** Let the initial distribution ρ_0 be the standard Gaussian distribution $N(0,I_D)$. Under certain regularity conditions, $\widehat{\rho}_{\lfloor t/\epsilon\rfloor}^{(m)}$ weakly converges to ρ_t as $\epsilon\to0^+$ and $m\to\infty$.

The proof of Proposition 3.1 is based on the propagation of chaos [53, 54, 66]. In contrast to [53, 54], the PDE in (3.4) can not be cast as a gradient flow, since there does not exist a corresponding energy functional. Thus, their analysis is not directly applicable to our setting. We defer the detailed discussion on the approximation analysis to §D. Proposition 3.1 allows us to convert the TD dynamics over the finite-dimensional parameter space to its counterpart over the infinite-dimensional Wasserstein space, where the infinitely wide neural network $Q(\cdot;\rho)$ in (3.2) is linear in the distribution ρ.

Feature Representation. We are interested in the evolution of the feature representation

$$\Big(\nabla_\theta\sigma\big(x;\theta_1(k)\big)^\top,\ldots,\nabla_\theta\sigma\big(x;\theta_m(k)\big)^\top\Big)^\top\in\mathbb{R}^{Dm}\tag{3.6}$$

corresponding to $\theta^{(m)}(k)=(\theta_1(k),\ldots,\theta_m(k))\in\mathbb{R}^{D\times m}$. Such a feature representation is used to analyze the TD dynamics $\theta^{(m)}(k)$ in (3.3) in the NTK regime [21], which corresponds to setting $\alpha=\sqrt{m}$ in (3.1). Meanwhile, the nonlinear gradient TD dynamics [15] explicitly uses such a feature representation at each iteration to locally linearize the Q-function. Moreover, up to a rescaling, such a feature representation corresponds to the kernel

$$\mathbb{K}(x,x';\widehat{\rho}_k^{(m)})=\int\nabla_\theta\sigma(x;\theta)^\top\nabla_\theta\sigma(x';\theta)\,\mathrm{d}\widehat{\rho}_k^{(m)}(\theta),$$

which by Proposition 3.1 further induces the kernel

$$\mathbb{K}(x,x';\rho_t)=\int\nabla_\theta\sigma(x;\theta)^\top\nabla_\theta\sigma(x';\theta)\,\mathrm{d}\rho_t(\theta)\tag{3.7}$$

at the mean-field limit with $\epsilon\to0^+$ and $m\to\infty$. Such a correspondence allows us to use the PDE solution ρ_t in (3.4) as a proxy for characterizing the evolution of the feature representation in (3.6).

4 Main Results

We first introduce the assumptions for our analysis.

Assumption 4.1. We assume that the state-action pair $x = (s, a)$ satisfies $\|x\| \le 1$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Assumption 4.1 can be ensured by normalizing all state-action pairs. Such an assumption is commonly used in the mean-field analysis of neural networks [6, 23, 27, 34, 35, 53, 54]. We remark that our analysis straightforwardly generalizes to the setting where $\|x\| \le C$ for an absolute constant $C > 0$.

Assumption 4.2. We assume that the activation function σ in (3.1) satisfies

$$\big|\sigma(x;\theta)\big| \le B_0, \quad \big\|\nabla_\theta \sigma(x;\theta)\big\| \le B_1 \cdot \|x\|, \quad \big\|\nabla^2_{\theta\theta}\sigma(x;\theta)\big\|_{\mathrm{F}} \le B_2 \cdot \|x\|^2 \tag{4.1}$$

for any $x \in \mathcal{X}$. Also, we assume that the reward r satisfies $|r| \le B_r$.

Assumption 4.2 holds for a broad range of neural networks. For example, let $\theta = (w, b) \in \mathbb{R}^{D-1} \times \mathbb{R}$. The activation function

$$\sigma^\dagger(x;\theta) = B_0 \cdot \tanh(b) \cdot \mathrm{sigmoid}(w^\top x) \tag{4.2}$$

satisfies (4.1) in Assumption 4.2. Moreover, the infinitely wide neural network in (3.2) with the activation function $\sigma^\dagger$ in (4.2) induces the following function class,

$$\mathcal{F}^\dagger = \bigg\{ \int \beta \cdot \mathrm{sigmoid}(w^\top x)\, \mathrm{d}\mu(w,\beta) \,\bigg|\, \mu \in \mathscr{P}\big(\mathbb{R}^{D-1} \times [-B_0, B_0]\big) \bigg\},$$

where $\beta = B_0 \cdot \tanh(b) \in [-B_0, B_0]$. By the universal approximation theorem [11, 60], $\mathcal{F}^\dagger$ captures a rich class of functions.

Throughout the rest of this paper, we consider the following function class,

$$\mathcal{F} = \bigg\{ \int \sigma_0(b) \cdot \sigma_1(x;w)\, \mathrm{d}\rho(w,b) \,\bigg|\, \rho \in \mathscr{P}_2(\mathbb{R}^{D-1} \times \mathbb{R}) \bigg\}, \tag{4.3}$$

which is induced by the infinitely wide neural network in (3.2) with $\theta = (w, b) \in \mathbb{R}^{D-1} \times \mathbb{R}$ and the following activation function,

$$\sigma(x;\theta) = \sigma_0(b) \cdot \sigma_1(x;w).$$

We assume that σ_0 is an odd function, that is, $\sigma_0(b) = -\sigma_0(-b)$, which implies $\int \sigma(x;\theta)\, \mathrm{d}\rho_0(\theta) = 0$. Note that the set of infinitely wide neural networks taking the forms of (3.2) is $\alpha \cdot \mathcal{F}$, which is larger than $\mathcal{F}$ in (4.3) by the scaling parameter $\alpha > 0$. Thus, α can be viewed as the degree of "overrepresentation". Without loss of generality, we assume that $\mathcal{F}$ is complete. The following theorem characterizes the global optimality and convergence of the PDE solution ρ_t in (3.4).

Theorem 4.3. There exists a unique fixed point solution to the projected Bellman equation $Q = \Pi_{\mathcal{F}} \mathcal{T}^\pi Q$, which takes the form of $Q^*(x) = \int \sigma(x;\theta)\, \mathrm{d}\bar{\rho}(\theta)$. Also, Q^* is the global minimizer of the MSPBE defined in (2.2). We assume that $D_{\chi^2}(\bar{\rho}\,\|\,\rho_0) < \infty$ and $\bar{\rho}(\theta) > 0$ for any $\theta \in \mathbb{R}^D$. Under Assumptions 4.1 and 4.2, it holds for $\eta = \alpha^{-2}$ in (3.4) that

$$\inf_{t \in [0,T]} \mathbb{E}_{x \sim \mathcal{D}}\Big[\big(Q(x;\rho_t) - Q^*(x)\big)^2\Big] \le \frac{D_{\chi^2}(\bar{\rho}\,\|\,\rho_0)}{2(1-\gamma)\cdot T} + \frac{C_*}{(1-\gamma)\cdot\alpha}, \tag{4.4}$$

where $C_* > 0$ is a constant that depends on $D_{\chi^2}(\bar{\rho}\,\|\,\rho_0)$, B_1, B_2, and B_r.

Proof. See §5 for a detailed proof. □

Theorem 4.3 proves that the optimality gap $\mathbb{E}_{x\sim\mathcal{D}}[(Q(x;\rho_t) - Q^*(x))^2]$ decays to zero at a sublinear rate up to the error of $O(\alpha^{-1})$, where $\alpha > 0$ is the scaling parameter in (3.1). Varying α leads to a tradeoff between such an error of $O(\alpha^{-1})$ and the deviation of ρ_t from ρ_0. Specifically, in §5 we prove that ρ_t deviates from ρ_0 by the divergence $D_{\chi^2}(\rho_t\,\|\,\rho_0) \le O(\alpha^{-2})$. Hence, a smaller α allows ρ_t to move further away from ρ_0, inducing a feature representation that is more different from the initial one [34, 35]. See (3.6)-(3.7) for the correspondence of ρ_t with the feature representation and

the kernel that it induces. On the other hand, a smaller α yields a larger error of $O(\alpha^{-1})$ in (4.4) of Theorem 4.3. In contrast, the NTK regime [21], which corresponds to setting $\alpha = \sqrt{m}$ in (3.1), only allows ρ_t to deviate from ρ_0 by the divergence $D_{\chi^2}(\rho_t \,\|\, \rho_0) \le O(m^{-1}) = o(1)$. In other words, the NTK regime fails to induce a feature representation that is significantly different from the initial one. In summary, our analysis goes beyond the NTK regime, which allows us to characterize the evolution of the feature representation towards the (near-)optimal one. Moreover, based on Proposition 3.1 and Theorem 4.3, we establish the following corollary, which characterizes the global optimality and convergence of the TD dynamics $\theta^{(m)}(k)$ in (3.3).

Corollary 4.4. Under the same conditions of Theorem 4.3, it holds with probability at least $1 - \delta$ that

$$\min_{\substack{k \le T/\epsilon \\ (k \in \mathbb{N})}} \mathbb{E}_{x \sim \mathcal{D}} \Big[\Big(\widehat{Q}\big(x; \theta^{(m)}(k)\big) - Q^*(x) \Big)^2 \Big] \le \frac{D_{\chi^2}(\bar{\rho} \,\|\, \rho_0)}{2(1-\gamma) \cdot T} + \frac{C_*}{(1-\gamma) \cdot \alpha} + \Delta(\epsilon, m, \delta, T), \quad (4.5)$$

where $C_* > 0$ is the constant of (4.4) in Theorem 4.3 and $\Delta(\epsilon, m, \delta, T) > 0$ is an error term such that

$$\lim_{m \to \infty} \lim_{\epsilon \to 0^+} \Delta(\epsilon, m, \delta, T) = 0.$$

Proof. See §D.2 for a detailed proof. □

In (4.5) of Corollary 4.4, the error term $\Delta(\epsilon, m, \delta, T)$ characterizes the error of approximating the TD dynamics $\theta^{(m)}(k)$ in (3.3) using the PDE solution ρ_t in (3.4). In particular, such an error vanishes at the mean-field limit.

5 Proof of Main Results

We first introduce two technical lemmas. Recall that $\mathcal{F}$ is defined in (4.3), $Q(x; \rho)$ is defined in (3.2), and $g(\theta; \rho)$ is defined in (3.5).

Lemma 5.1. There exists a unique fixed point solution to the projected Bellman equation $Q = \Pi_{\mathcal{F}} \mathcal{T}^\pi Q$, which takes the form of $Q^*(x) = \int \sigma(x; \theta) \, \mathrm{d}\bar{\rho}(\theta)$. Also, there exists $\rho^* \in \mathscr{P}_2(\mathbb{R}^D)$ that satisfies the following properties,

(i) $Q(x; \rho^*) = Q^*(x)$ for any $x \in \mathcal{X}$,

(ii) $g(\cdot; \rho^*) = 0$ for $\bar{\rho}$-a.e., and

(iii) $\mathcal{W}_2(\rho^*, \rho_0) \le \alpha^{-1} \cdot \bar{D}$, where $\bar{D} = D_{\chi^2}(\bar{\rho} \,\|\, \rho_0)^{1/2}$.

Proof. See §C.1 for a detailed proof. The proof of (iii) is adopted from [23], which focuses on supervised learning. □

Lemma 5.1 establishes the existence of the fixed point solution Q^* to the projected Bellman equation $Q = \Pi_{\mathcal{F}} \mathcal{T}^\pi Q$. Furthermore, such a fixed point solution Q^* can be parameterized with the infinitely wide neural network $Q(\cdot; \rho^*)$ in (3.2). Meanwhile, the Wasserstein-2 distance between ρ^* and the initial distribution ρ_0 is upper bounded by $O(\alpha^{-1})$. Based on the existence of Q^* and the property of ρ^* in Lemma 5.1, we establish the following lemma that characterizes the evolution of $\mathcal{W}_2(\rho_t, \rho^*)$, where ρ_t is the PDE solution in (3.4).

Lemma 5.2. We assume that $\mathcal{W}_2(\rho_t, \rho^*) \le 2\mathcal{W}_2(\rho_0, \rho^*)$, $D_{\chi^2}(\bar{\rho} \,\|\, \rho_0) < \infty$, and $\bar{\rho}(\theta) > 0$ for any $\theta \in \mathbb{R}^D$. Under Assumptions 4.1 and 4.2, it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathcal{W}_2(\rho_t, \rho^*)^2}{2} \le -(1-\gamma) \cdot \eta \cdot \mathbb{E}_{x \sim \mathcal{D}} \Big[\big(Q(x; \rho_t) - Q^*(x) \big)^2 \Big] + C_* \cdot \alpha^{-1} \cdot \eta, \quad (5.1)$$

where $C_* > 0$ is a constant depending on $D_{\chi^2}(\bar{\rho} \,\|\, \rho_0)$, B_1, B_2, and B_r.

Proof. See §C.2 for a detailed proof. □

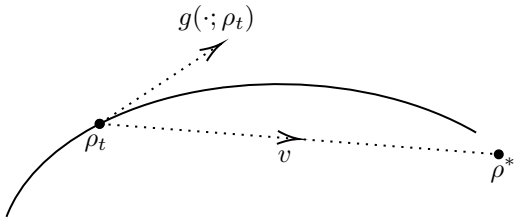

Figure 1: We illustrate the first variation formula $\frac{\mathrm{d}\mathcal{W}_2(\rho_t,\rho^*)^2}{2} = -\langle g(\cdot;\rho_t), v\rangle_{\rho_t}$, where v is the vector field corresponding to the geodesic that connects ρ_t and ρ^*. See Lemma E.2 for details.

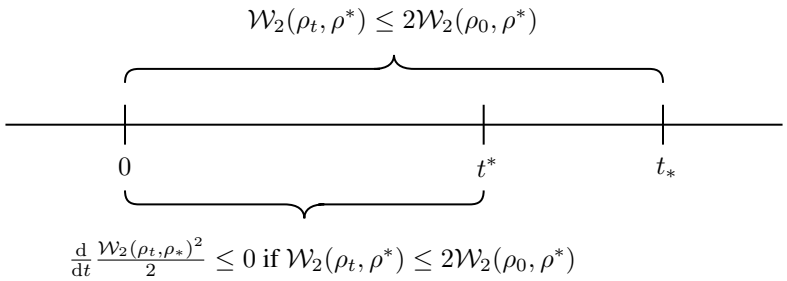

Figure 2: For any $0 \le t \le \min\{t^*, t_*\}$, (5.1) of Lemma 5.2 holds and $\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t,\rho_*)^2}{2} \le 0$.

The proof of Lemma 5.2 is based on the first variation formula of the Wasserstein-2 distance (Lemma E.2), which is illustrated in Figure 1, and the one-point monotonicity of $g(\cdot;\beta_t)$ along a curve β on the Wasserstein space (Lemma C.1). When the right-hand side of (5.1) is nonpositive, Lemma 5.2 characterizes the decay of $\mathcal{W}_2(\rho_t,\rho^*)$. We are now ready to present the proof of Theorem 4.3.

Proof. We use a continuous counterpart of the induction argument. We define

$$t^* = \inf\left\{\tau \in \mathbb{R}_+ \,\middle|\, \mathbb{E}_{x\sim\mathcal{D}}\Big[(1-\gamma)\cdot\big(Q(x;\rho_\tau) - Q^*(x)\big)^2\Big] < C_*\cdot\alpha^{-1}\right\}. \tag{5.2}$$

In other words, the right-hand side of (5.1) in Lemma 5.2 is nonpositive for any $t \le t^*$, that is,

$$-(1-\gamma)\cdot\mathbb{E}_{x\sim\mathcal{D}}\Big[\big(Q(x;\rho_t) - Q^*(x)\big)^2\Big] + C_*\cdot\alpha^{-1} \le 0. \tag{5.3}$$

Also, we define

$$t_* = \inf\big\{\tau \in \mathbb{R}_+ \,\big|\, \mathcal{W}_2(\rho_\tau,\rho^*) > 2\mathcal{W}_2(\rho_0,\rho^*)\big\}. \tag{5.4}$$

In other words, (5.1) of Lemma 5.2 holds for any $t \le t_*$. Thus, for any $0 \le t \le \min\{t^*, t_*\}$, it holds that $\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t,\rho_*)^2}{2} \le 0$. Figure 2 illustrates the definition of t^* and t_* in (5.2) and (5.4), respectively.

We now prove that $t_* \ge t^*$ by contradiction. By the continuity of $\mathcal{W}_2(\rho_t,\rho^*)^2$ with respect to t [5], it holds that $t_* > 0$, since $\mathcal{W}_2(\rho_0,\rho^*) < 2\mathcal{W}_2(\rho_0,\rho^*)$. For the sake of contradiction, we assume that $t_* < t^*$, by (5.1) of Lemma 5.2 and (5.3), it holds for any $0 \le t \le t_*$ that

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t,\rho^*)^2}{2} \le 0,$$

which implies that $\mathcal{W}_2(\rho_t,\rho^*) \le \mathcal{W}_2(\rho_0,\rho^*)$ for any $0 \le t \le t_*$. This contradicts the definition of t_* in (5.4). Thus, it holds that $t_* \ge t^*$, which implies that (5.1) of Lemma 5.2 holds for any $0 \le t \le t^*$.

If $t^* \le T$, (5.3) implies Theorem 4.3. If $t^* > T$, by (5.1) of Lemma 5.2, it holds for any $0 \le t \le T$ that

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t,\rho^*)^2}{2} \le -(1-\gamma)\cdot\eta\cdot\mathbb{E}_{x\sim\mathcal{D}}\Big[\big(Q(x;\rho_t) - Q^*(x)\big)^2\Big] + C_*\cdot\alpha^{-1}\cdot\eta \le 0,$$

which further implies that

$$\mathbb{E}_{x\sim\mathcal{D}}\Big[\big(Q(x;\rho_t)-Q^*(x)\big)^2\Big] \le -(1-\gamma)^{-1}\cdot\eta^{-1}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathcal{W}_2(\rho_t,\rho^*)^2}{2}+\frac{C_*}{(1-\gamma)\cdot\alpha}. \tag{5.5}$$

Upon telescoping (5.5) and setting $\eta=\alpha^{-2}$, we obtain that

$$\begin{aligned}
&\inf_{t\in[0,T]}\mathbb{E}_{\mathcal{D}}\Big[\big(Q(x;\rho_t)-Q^*(x)\big)^2\Big]\\
&\le T^{-1}\cdot\int_0^T\mathbb{E}_{x\sim\mathcal{D}}\Big[\big(Q(x;\rho_t)-Q^*(x)\big)^2\Big]\,\mathrm{d}t\\
&\le 1/2\cdot(1-\gamma)^{-1}\cdot\eta^{-1}\cdot T^{-1}\cdot\mathcal{W}_2(\rho_0,\rho^*)^2+C_*\cdot(1-\gamma)^{-1}\cdot\alpha^{-1}\\
&\le 1/2\cdot(1-\gamma)^{-1}\cdot\bar{D}^2\cdot T^{-1}+C_*\cdot(1-\gamma)^{-1}\cdot\alpha^{-1},
\end{aligned}$$

where the last inequality follows from the fact that $\eta=\alpha^{-2}$ and (iii) of Lemma 5.1. Thus, we complete the proof of Theorem 4.3. □

6 Extension to Q-Learning and Policy Improvement

In §B, we extend our analysis of TD to Q-learning and soft Q-learning for policy improvement. In §B.1, we introduce Q-learning and its mean-field limit. In §B.2, we establish the global optimality and convergence of Q-learning. In §B.3, we further extend our analysis to soft Q-learning, which is equivalent to a variant of policy gradient [37, 57, 58, 61].

Broader Impact

The popularity of RL creates a responsibility for researchers to design algorithms with guaranteed safety and robustness, which rely on their stability and convergence. In this paper, we provide a theoretical understanding of the global optimality and convergence of the TD and Q-learning with neural network parameterization. We believe that our work is an important step forward in the algorithm design of RL in emerging high-stakes applications, such as autonomous driving, personalized medicine, power systems, and robotics.

References

[1] Agazzi, A. and Lu, J. (2019). Temporal-difference learning for nonlinear value function approximation in the lazy training regime. *arXiv preprint arXiv:1905.10917*.

[2] 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*.

[3] Allen-Zhu, Z., Li, Y. and Song, Z. (2018). A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*.

[4] Ambrosio, L. and Gigli, N. (2013). A users guide to optimal transport. In *Modelling and Optimisation of Flows on Networks*. Springer, 1–155.

[5] Ambrosio, L., Gigli, N. and Savaré, G. (2008). *Gradient flows: In metric spaces and in the space of probability measures*. Springer.

[6] Araújo, D., Oliveira, R. I. and Yukimura, D. (2019). A mean-field limit for certain deep neural networks. *arXiv preprint arXiv:1906.00193*.

[7] Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R. and Wang, R. (2019). On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*.

[8] 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*.

[9] Bai, Y. and Lee, J. D. (2019). Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*.

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

[11] Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, **39** 930–945.

[12] Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. In *ICML Workshop on Unsupervised and Transfer Learning*.

[13] 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.

[14] 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*.

[15] 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*.

[16] Borkar, V. S. (2009). *Stochastic approximation: A dynamical systems viewpoint*. Springer.

[17] 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.

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

[19] Brandfonbrener, D. and Bruna, J. (2019). Geometric insights into the convergence of nonlinear TD learning. *arXiv preprint arXiv:1905.12185*.

[20] Brandfonbrener, D. and Bruna, J. (2019). On the expected dynamics of nonlinear TD learning. *arXiv preprint arXiv:1905.12185*.

[21] Cai, Q., Yang, Z., Lee, J. D. and Wang, Z. (2019). Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*.

[22] Cao, Y. and Gu, Q. (2019). Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210*.

[23] Chen, Z., Cao, Y., Gu, Q. and Zhang, T. (2020). Mean-field analysis of two-layer neural networks: Non-asymptotic rates and generalization bounds. *arXiv preprint arXiv:2002.04026*.

[24] Chen, Z., Cao, Y., Zou, D. and Gu, Q. (2019). How much over-parameterization is sufficient to learn deep ReLU networks? *arXiv preprint arXiv:1911.12360*.

[25] Chen, Z., Zhang, S., Doan, T. T., Maguluri, S. T. and Clarke, J.-P. (2019). Performance of q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*.

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

[27] Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*.

[28] Conway, J. B. (2019). *A course in functional analysis*, vol. 96. Springer.

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

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

[31] 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.

[32] Du, S. S., Lee, J. D., Li, H., Wang, L. and Zhai, X. (2018). Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*.

[33] Du, S. S., Zhai, X., Poczos, B. and Singh, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

[34] Fang, C., Dong, H. and Zhang, T. (2019). Over parameterized two-level neural networks can learn near optimal feature representations. *arXiv preprint arXiv:1910.11508*.

[35] Fang, C., Gu, Y., Zhang, W. and Zhang, T. (2019). Convex formulation of overparameterized deep neural networks. *arXiv preprint arXiv:1911.07626*.

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

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

[38] 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.

[39] He, J., Chen, J., He, X., Gao, J., Li, L., Deng, L. and Ostendorf, M. (2015). Deep reinforcement learning with a natural language action space. *arXiv preprint arXiv:1511.04636*.

[40] Hinton, G. (1986). Learning distributed representations of concepts. In *Annual Conference of Cognitive Science Society*.

[41] Holte, J. M. (2009). Discrete Gronwall lemma and applications. In *MAA-NCS Meeting at the University of North Dakota*, vol. 24.

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

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

[44] Javanmard, A., Mondelli, M. and Montanari, A. (2019). Analysis of a two-layer neural network via displacement convexity. *arXiv preprint arXiv:1901.01375*.

[45] Ji, Z. and Telgarsky, M. (2019). Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. *arXiv preprint arXiv:1909.12292*.

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

[47] Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer.

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

[49] LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, **521** 436–444.

[50] 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*.

[51] Levine, S., Finn, C., Darrell, T. and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, **17** 1334–1373.

[52] Li, Y. and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*.

[53] Mei, S., Misiakiewicz, T. and Montanari, A. (2019). Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*.

[54] Mei, S., Montanari, A. and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, **115** E7665–E7671.

[55] Melo, F. S., Meyn, S. P. and Ribeiro, M. I. (2008). An analysis of reinforcement learning with function approximation. In *International Conference on Machine Learning*.

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

[57] Nachum, O., Norouzi, M., Xu, K. and Schuurmans, D. (2017). Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*.

[58] O'Donoghue, B., Munos, R., Kavukcuoglu, K. and Mnih, V. (2016). Combining policy gradient and Q-learning. *arXiv preprint arXiv:1611.01626*.

[59] Otto, F. and Villani, C. (2000). Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, **173** 361–400.

[60] Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica*, **8** 143–195.

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

[62] Sirignano, J. and Spiliopoulos, K. (2019). Asymptotics of reinforcement learning with neural networks. *arXiv preprint arXiv:1911.07304*.

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

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

[65] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[66] Sznitman, A.-S. (1991). Topics in propagation of chaos. In *Ecole d'Été de Probabilités de Saint-Flour XIX1989*. Springer, 165–251.

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

[68] Villani, C. (2003). *Topics in optimal transportation*. American Mathematical Society.

[69] Villani, C. (2008). *Optimal transport: Old and new*. Springer.

[70] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

[71] Watkins, C. and Dayan, P. (1992). Q-learning. *Machine Learning*, **8** 279–292.

[72] Wei, C., Lee, J. D., Liu, Q. and Ma, T. (2019). Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*.

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

[74] Xu, T., Zou, S. and Liang, Y. (2019). Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Advances in Neural Information Processing Systems*.

[75] Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*.

[76] Zou, D., Cao, Y., Zhou, D. and Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888*.

[77] Zou, D. and Gu, Q. (2019). An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*.

[78] 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*.