Sample Complexity and Overparameterization Bounds for Temporal Difference Learning with Neural Network Approximation

Author(s):
Cayci, Semih; Satpathi, Siddhartha; He, Niao; Srikant, Rayadurgam

Publication date:
2021

Permanent link:
https://doi.org/10.3929/ethz-b-000555353

Sample Complexity and Overparameterization Bounds for Temporal Difference Learning with Neural Network Approximation

Semih Cayci [1] **Siddhartha Satpathi** [1] **Niao He** [2] **R. Srikant** [3 1 4]

Abstract

We study the dynamics of temporal difference learning with neural network-based value function approximation over a general state space, namely, *Neural TD learning*. We consider two practically used algorithms, *projection-free* and *max-norm regularized* Neural TD learning, and establish the first convergence bounds for these algorithms. An interesting observation from our results is that max-norm regularization can dramatically improve the performance of TD learning algorithms, both in terms of sample complexity and overparameterization. In particular, we prove that max-norm regularization achieves state-of-the-art sample complexity and overparameterization bounds by exploiting the geometry of the neural tangent kernel (NTK) function class. The results in this work rely on a novel Lyapunov drift analysis of the network parameters as a stopped and controlled random process.

1. Introduction

Recently, deep reinforcement learning (RL) algorithms have achieved significant breakthroughs in challenging high-dimensional problems in a broad spectrum of applications including video gaming (Mnih et al., 2013; Silver et al., 2017b;a), natural language processing (Li et al., 2016), and robotics (Gu et al., 2017; Kalashnikov et al., 2018). An important component of these success stories lies in the power and versatility provided by neural networks in function approximation. Despite the impressive empirical success, the convergence properties of RL algorithms with neural network approximation are not yet fully understood due to their inherent nonlinearity.

In this paper, we investigate the convergence of temporal-difference (TD) learning algorithm equipped with neural network approximation, namely Neural TD learning, which is an important building block of many deep RL algorithms (Konda & Tsitsiklis, 2000; Wang et al., 2019). Despite the theoretical insights provided by recent studies (Cai et al., 2019; Xu & Gu, 2020; Brandfonbrener & Bruna, 2020; Agazzi & Lu, 2019; Sirignano & Spiliopoulos, 2019), there is still a gap between theory and practice. In order to address this, in this paper, we consider two practically used Neural TD learning algorithms, projection-free and max-norm regularized, and establish sharp convergence bounds. Interestingly, our theoretical findings show that regularization based on max-norm geometry yields sharp convergence bounds, which justifies the success of max-norm regularization methods in practical applications.

1.1. Main Contributions

The paper presents a non-asymptotic analysis of TD learning with neural network approximation. We elaborate on some of the contributions in this paper below:

• **Analysis of Neural TD learning:** We analyze two practically used Neural TD learning algorithms: (i) vanilla projection-free Neural TD and (ii) max-norm regularized Neural TD. We prove, for the first time, that both algorithms achieve any given target error within a provably rich function class, which is dense in the space of continuous functions over a compact state space. In particular, we establish explicit bounds on the required number of samples, step-size and network width to achieve a given target error.

• **Improved convergence bounds:** We show that projection-free and max-norm regularized Neural TD improve the prior state-of-the-art overparameterization bounds by factors of $1/\epsilon^2$ and $1/\epsilon^6$, respectively, for a given target error ϵ. Notably, we prove that max-norm regularized Neural TD achieves the sharpest overparameterization and sample complexity bounds in the literature, which theoretically supports its empirical effectiveness.

• **Key insights on regularization:** Our analysis reveals that using regularization based on ℓ_∞ geometry leads to

[1]Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA [2]Department of Computer Science, ETH Zurich, Zurich, Switzerland [3]c3.ai DTI [4]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. Correspondence to: Semih Cayci <scayci@illinois.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

considerably improved overparameterization and sample complexity bounds compared to the ℓ_2-regularization over a provably rich function class in the NTK regime.

• **Analytical techniques:** We propose a novel Lyapunov drift analysis to track the evolution of neural network parameters and the error simultaneously using martingale concentration and stopping times.

1.2. Related Work

Neural TD learning has been considered in a variety of papers. For a detailed comparison, please refer to Appendix A.

It is shown in (Cai et al., 2019; Wang et al., 2019) that Neural TD learning with ℓ_2 projection, equipped with a ReLU network of width $O(1/\epsilon^8)$ achieves an error $\epsilon + \epsilon_m$ after $O(1/\epsilon^4)$ iterations. Unlike (Cai et al., 2019; Wang et al., 2019), our Neural TD learning algorithms converge to the *true* value function in a provably rich function class without any approximation error. We show that the algorithms that we consider in this paper achieve improved overparameterization bounds $\widetilde{O}(1/\epsilon^6)$ and $\widetilde{O}(1/\epsilon^2)$ for a given target error ϵ, which improve the existing results by $1/\epsilon^2$ to $1/\epsilon^6$. The results with ℓ_2-regularization were generalized to deep Q-learning setting in (Xu & Gu, 2020).

In another line of work, (Agazzi & Lu, 2019; Brandfonbrener & Bruna, 2020) consider Neural TD learning; however, these works only deal with finite state-space problems in the infinite-width regime. Since these results rely on the positive definiteness of the limiting kernel, the required overparameterization is much larger than the size of the state space which negates the benefits of Neural TD learning over tabular TD learning.

2. System Model

For simplicity, we consider a Markov reward process $\{(s_t, r_t) : t = 0, 1, \ldots\}$, where the Markov chain s_t takes on values in the state space $\mathcal{S}$, and there is an associated reward $r_t = r(s_t)$ in every time-step for a reward function $r : \mathcal{S} \to [0, 1]$. The process $\{s_t : t \geq 0\}$ evolves according to the transition probabilities $P(s, A) = \mathbb{P}(s_{t+1} \in A | s_t = s)$ for any $s \in \mathcal{S}$, $A \subset \mathcal{S}$ and $t \geq 0$. We assume that the Markov chain $\{s_t : t \geq 0\}$ is an ergodic unichain, therefore there exists a stationary probability distribution π:

$$\pi(A) = \lim_{t\to\infty} \mathbb{P}(s_t \in A | s_0 = s), \ \forall s \in \mathcal{S}, A \subset \mathcal{S}.$$

The value function associated with the Markov reward process $\{(s_t, r_t) : t \geq 0\}$ is defined as follows:

$$V(s) = \mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma^t r_t | s_0 = s\Big], \ \forall s \in \mathcal{S}, \tag{1}$$

where $\gamma \in (0, 1)$ is the discount factor. The Bellman operator for this Markov reward process, denoted by $\mathcal{T}$, is defined as follows:

$$\mathcal{T}\widehat{V}(s) = r(s) + \gamma \int_{s' \in \mathcal{S}} \widehat{V}(s') P(s, ds'), \ \forall s \in \mathcal{S}. \tag{2}$$

For simplicity, we consider iid distributed samples from the stationary distribution π. Namely, we obtain (s_t, s'_t) where $s_t \sim \pi$ and $s'_t \sim P(s_t, \cdot)$. We denote $\mathcal{F}_t = \sigma(\{(s_j, s'_j) : j = 0, 1, \ldots, t\})$ to be the history up to (including) time t. The case where the samples are generated by the Markov reward process can be handled as in (Srikant & Ying, 2019), but we do not consider that here.

Assumption 1. *For any state $x \in \mathcal{X}$, we assume $\|x\|_2 \leq 1$.*

In the next subsection, we introduce the neural network architecture that will be used to approximate the value function.

2.1. Neural Network Architecture for Value Function Approximation

Throughout the paper, we consider the two-layer ReLU network to approximate the value function V:

$$Q(x; W, a) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \sigma(W_i^\top x). \tag{3}$$

where $\sigma(z) = \max\{0, z\} = z \cdot \mathbb{I}\{z \geq 0\}$ is the ReLU activation function, $a_i \in \mathbb{R}$ and $W_i \in \mathbb{R}^d$ for $i \in [m]$. We include a bias term in W_i's, and express x as (x, c) for a constant $c \in (0, 1)$.

Symmetric initialization: The NTK regime is established by random initialization. In this paper, we consider the symmetric initialization, which was proposed in (Bai & Lee, 2019): $a_i = -a_{i+m/2} \overset{\text{iid}}{\sim} \text{Unif}\{-1, +1\}$ and $W_i(0) = W_{i+m/2}(0) \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ iid over $i = 1, 2, \ldots, m/2$, and independent from each other. The benefit of the symmetric initialization is that it provides $Q(x; W(0), a) = 0$ with probability 1 for all $x \in \mathcal{X}$.

In the following, inspired by (Ji & Telgarsky, 2019), we define the function class that we consider in this paper, which can be realized by a neural network in the NTK regime.

Function class: Define the space

$$\mathcal{H} = \Big\{ v : \mathbb{R}^d \to \mathbb{R}^d \,\Big|\, \mathbb{E}\big[\|v(w_0)\|_2^2\big] < \infty, w_0 \sim \mathcal{N}(0, I_d) \Big\}.$$

We assume that the value function V lies in the following function class.

Assumption 2. *There exists a vector $v \in \mathcal{H}$ and $\bar{\nu} \geq 0$ such that:*

$$V(x) = \mathbb{E}[v^\top(w_0)\phi(x; w_0)], \ w_0 \sim \mathcal{N}(0, I_d), \ \forall x \in \mathcal{X}, \tag{4}$$

where $\sup_{w \in \mathbb{R}^d} \|v(w)\|_2 \leq \bar{\nu}$ and $\phi(x; w) = \mathbb{I}\{w^\top x \geq 0\}x$.

Remark 1. If we replace the condition $\sup_{w\in\mathbb{R}^d}\|v(w)\|_2 \leq \bar{\nu}$ in Assumption 2 by $\mathbb{E}\big[\|v(w_0)\|_2^2\big] < \infty$, then it implies that V belongs to the reproducing kernel Hilbert space (RKHS) induced by the Neural Tangent Kernel (NTK). The above kernel can be shown to be a universal kernel (Ji et al., 2019) and hence the RKHS induced by the NTK is dense in the space of continuous functions on compact set $\mathcal{X}$ (Micchelli et al., 2006). Therefore, it is possible to replace Assumption 2 by the more general assumption that V is continuous on a compact state space $\mathcal{X}$. In this case, from Theorem 4.3 in (Ji et al., 2019), one can find a function $\tilde{V}$ in the RKHS associated with the NTK, i.e., $\tilde{V}(x) = \mathbb{E}[\tilde{v}^\top(w_0)\phi(x;w_0)]$, $\forall x \in \mathcal{X}$, such that $\sup_w \|\tilde{v}(w)\|_2 \leq \bar{\nu}$ for some finite $\bar{\nu}$ which approximates V, where $\bar{\nu}$ depends on the approximation error $\sup_x |V(x) - \tilde{V}(x)|$.

3. Neural Temporal Difference Learning Algorithms

For a given function $\mu = [\mu(x)]_{x\in\mathcal{X}}$, we denote the weighted ℓ_2-norm of any function $\widehat{V}$ as:

$$\|\widehat{V}\|_\mu = \sqrt{\int_{x\in\mathcal{X}} |\widehat{V}(x)|^2 \mu(dx)}.$$

TD learning aims to minimize mean-squared Bellman error, which is defined as follows:

$$\begin{aligned} L(W,a) &= \|Q(W,a) - \mathcal{T}Q(W,a)\|_\pi^2 \\ &= \int_{x\in\mathcal{X}} \Big(Q(x;W,a) - \mathcal{T}Q(x;W,a)\Big)^2 \pi(dx), \end{aligned} \tag{5}$$

for any $W_i \in \mathbb{R}^d, a_i \in \mathbb{R}$ for $i = 1,2,\ldots,m$, where $Q(W,a) = [Q(x;W,a)]_{x\in\mathcal{X}}$, π is the stationary distribution of the Markov chain, and $\mathcal{T}$ is the Bellman operator.

Given the initialization $\{\big(a_i, W_i(0)\big) : i \in [m]\}$, the parameter update is performed as follows:

$$W(t+1/2) = W(t) + \alpha\Big(r_t + \gamma Q_t(x_t') - Q_t(x_t)\Big)\nabla_W Q_t(x_t),$$

where $\alpha > 0$ is the step-size, $Q_t(x) = Q(x; W(t), a)$ is the network at time step $t \geq 0$. The algorithm is summarized in Algorithm 1. We consider two variants of the Neural TD learning algorithm:

(1) **Projection-free Neural TD learning (PF-NTD):** The network parameters are updated as follows:

$$W(t+1) = W(t+1/2). \tag{6}$$

For regularization, we utilize early stopping, i.e., the number of samples T is chosen as a function of the problem parameters and target error, which we will specify in Theorem 1.

Algorithm 1 PF/MN-Neural TD Learning

Initialization: $-a_i = a_{i+m/2} \sim \text{Unif}\{-1,+1\}$, $W_i(0) = W_{i+m/2}(0) \sim \mathcal{N}(0, I_d), \forall i \in [\frac{m}{2}]$
for $t < T-1$ **do**
  Observe $x_t = \psi(s_t), r_t = r(s_t)$ and $x_t' = \psi(s_t')$
  Compute stochastic semi-gradient: g_t
  Take a semi-gradient step: $W(t+1/2) = W(t) + \alpha g_t$
  if projection-free **then**
    $W(t+1) = W(t+1/2)$
  end if
  if max-norm regularization **then**
    $W_i(t+1) = \Pi_{\mathcal{G}_{m,R}^i} W_i(t+1/2), \forall i \in [m]$
  end if
  Update iterate:

$$\widehat{W}(t+1) = \Big(1 - \frac{1}{t+2}\Big)\widehat{W}(t) + \frac{1}{t+2}W(t+1)$$

end for
Output: $\overline{Q}_T(x) = Q(x; \widehat{W}(T-1), a)$ for all $x \in \mathcal{X}$

(2) **Max-norm regularized Neural TD learning (MN-NTD):** For a given parameter $R > 0$, let the set of parameters for max-norm regularization be defined as:

$$\mathcal{G}_{m,R}^i = \{W_i \in \mathbb{R}^d : \|W_i - W_i(0)\|_2 \leq \frac{R}{\sqrt{m}}\}, \forall i \in [m]. \tag{7}$$

Then, the network parameters are updated as follows:

$$W_i(t+1) = \Pi_{\mathcal{G}_{m,R}^i} W_i(t+1/2), \forall i \in [m]. \tag{8}$$

where $\Pi_{\mathcal{G}}(\cdot)$ is the projection operator onto set $\mathcal{G}$.

Max-norm regularization was introduced in (Srebro & Shraibman, 2005; Srebro et al., 2005), and has been widely used in training neural networks (Srivastava et al., 2014; Goodfellow et al., 2013). Unlike the ℓ_2-projection in (Wang et al., 2019; Cai et al., 2019), max-norm regularization can be performed in parallel for all $i \in [m]$, thus it is computationally more feasible. Also, it implies projection onto a well-chosen subset, which leads to state-of-the-art overparameterization and sample complexity bounds as we will show in Theorem 2.

4. Main Results

In the following, we present our main results on the performance of Neural TD learning algorithms described in Section 3.

Theorem 1. *Under Assumptions 1 and 2, for any (possibly infinite) state-space $\mathcal{X}$, target error $\epsilon > 0$ and error probability $\delta \in (0,1)$, let $\lambda = \frac{3\bar{\nu}^2}{(1-\gamma)\epsilon\delta}$, $\ell(m,\delta) =$*

$4\sqrt{\log(2m+1)} + \sqrt{\log(1/\delta)}$,

$$m_0 = \frac{16\Big(\bar{\nu} + \big(\lambda + \ell(m_0,\delta)\big)\big(\bar{\nu}+\lambda\big)\Big)^2}{(1-\gamma)^2\epsilon^2},$$

$$\alpha_0 = \frac{(1-\gamma)\epsilon^2}{(1+2\lambda)^2}\min\Big\{\frac{\lambda^2}{32\bar{\nu}^2(\sqrt{d}+\sqrt{2\log(m_0/\delta)})^2}, 1\Big\}.$$

Then, for any width $m \geq m_0$, PF-NTD with step-size $\alpha \leq \alpha_0$ yields the following bound after $T = \frac{\bar{\nu}^2}{4\alpha(1-\gamma)\epsilon^2}$ iterations:

$$\mathbb{E}\Big[\big\|\overline{Q}_T - V\big\|_\pi; \mathcal{E}_T\Big] \leq \frac{1}{T}\sum_{t<T}\mathbb{E}[\|Q_t - V\|_\pi; \mathcal{E}_T] + \epsilon \leq 4\epsilon,$$

where $Q_t = [Q_t(x)]_{x\in\mathcal{X}}$, $V = [V(x)]_{x\in\mathcal{X}}$, and $\mathbb{P}(\mathcal{E}_T) > 1 - 4\delta$.

Theorem 1 implies that there exists a set $\mathcal{E}_T$ of trajectories which occurs with probability at least $1-4\delta$ such that PF-NTD achieves target error ϵ under the event $\mathcal{E}_T$ for $m = \frac{poly(\bar{\nu})}{\epsilon^6}$ and $T = \frac{poly(\bar{\nu},\frac{1}{\delta})}{\epsilon^6}$.

Theorem 2. *Under Assumptions 1-2, for any $R > \bar{\nu}$, $T \in \mathbb{N}$, $m > \frac{16\Big(\bar{\nu}+\big(R+\ell(m,\delta)\big)\big(\bar{\nu}+R\big)\Big)^2}{(1-\gamma)^2\epsilon^2}$, and step-size $\alpha = \frac{\epsilon^2(1-\gamma)}{(1+2R)^2}$, MN-NTD yields:*

$$\mathbb{E}\Big[\|\overline{Q}_T - V\|_\pi; E_1\Big] \leq \frac{(1+2R)\bar{\nu}}{\epsilon\sqrt{T}} + 3\epsilon,$$

where $E_1 \in \mathcal{F}_{init}$ holds with probability at least $1-\delta$.

Theorem 2 implies that there exists a set E_1 of trajectories which occurs with probability at least $1-\delta$ such that MN-NTD achieves target error ϵ under the event E_1 for $m = \frac{poly(\bar{\nu})}{\epsilon^4}$ and $T = \frac{poly(\bar{\nu},\log\frac{1}{\delta})}{\epsilon^4}$.

The analysis of PF-NTD relies on the following stopping time:

Definition 1. *For λ as given in Theorem 1, let*

$$t_1 = \inf\Big\{t > 0 : \max_{i\in[m]}\|W_i(t) - W_i(0)\|_2 > \frac{\lambda}{\sqrt{m}}\Big\}, \quad (9)$$

be the stopping time.

Proof outline: Below, we outline the proof steps for Theorem 1. The methodology is inspired by (Ji & Telgarsky, 2019), which considers binary classification problem with cross-entropy loss. The details can be found in Appendix C.

1. First, we prove a Lyapunov drift bound for $\|W(t) - \overline{W}\|_2$ which holds for all $t < t_1$ where $\overline{W} \in \mathbb{R}^{md}$ is such that $\nabla_W^\top Q_0(x)\overline{W} \approx V(x)$ for all $x \in \mathcal{X}$. A novel symmetrization and concentration argument enables establishing this bound for infinite $\mathcal{X}$.

2. In the second step, we use the drift bound in conjunction with a stopped martingale concentration argument to show that $t_1 \geq T$ occurs with high probability, thus the drift bound holds for all $t < T$ under that event.

3. Finally, we will use the drift bound again to show that the approximation error is bounded as in Theorem 1 under the high-probability event considered in Step 2.

5. Remarks and Conclusions

Theorems 1 and 2 provide, to the best of our knowledge, the first explicit convergence bounds for PF-NTD and MN-NTD. Below we list some further implications.

ℓ_2 vs. ℓ_∞ regularizations: Both PF-NTD and MN-NTD yield improved bounds on m compared to the algorithms in (Cai et al., 2019; Wang et al., 2019) over the provably rich NTK function class. A key insight from our analysis is that this improvement is mainly because both PF-NTD and MN-NTD are designed to control $\max_{i\in[m]}\|W_i(t) - W_i(0)\|_2$ via the choice of the stopping time (PF-NTD) or max-norm projection (MN-NTD), while ℓ_2 regularization in (Cai et al., 2019; Wang et al., 2019) is designed to control $\|W(t) - W(0)\|_2$. Notably, MN-NTD achieves the sharpest overparameterization and sample complexity bounds among all existing NTD variants, which justifies the empirical success of max-norm regularization for training ReLU networks in practice (Srivastava et al., 2014; Goodfellow et al., 2013).

Approximation power: PF-NTD fully exploits the expressive power of the neural network approximation in practice since the parameters are not strictly constrained. On the other hand, MN-NTD confines the network parameters within the sets $\mathcal{G}^i_{m,R}$ with a fixed radius $R/\sqrt{m}$, which may limit the expressive power of the neural network, especially for small radius R. A similar loss of approximation power arise for the projection-based NTD studied in (Cai et al., 2019; Wang et al., 2019) for the same reason.

Convergence rate: Regularization of PF-NTD relies on early stopping, whereas MN-NTD utilizes more aggressive max-norm regularization. Without any strict control over $\max_{i\in[m]}\|W_i(t) - W_i(0)\|_2$, PF-NTD requires considerably smaller step-sizes for convergence. Consequently, the sample complexity and required width for PF-NTD to achieve a target error ϵ are worse than MN-NTD for which larger step-sizes can be chosen.

Future work: In this work, we proved that any target error ϵ can be achieved by projection-free TD learning under early stopping for a specific $T = T(\epsilon, \bar{\nu}, \delta)$. The behavior of Neural TD learning beyond this T requires the analysis of $W(t)$ leaving the vicinity of $W(0)$, i.e., going beyond the NTK regime, and is an important open problem. The benefit of increasing the number of layers is also an open question.

References

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

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Bai, Y. and Lee, J. D. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations*, 2019.

Bartlett, P. L. Glivenko-cantelli classes. *Lecture Notes, CS281B/Stat241B, Statistical Learning Theory*, 2003.

Bertsekas, D. P. Dynamic programming and optimal control 3rd edition, volume ii. *Belmont, MA: Athena Scientific*, 2011.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pp. 1691–1692. PMLR, 2018.

Brandfonbrener, D. and Bruna, J. Geometric insights into the convergence of non-linear td learning. In *International Conference on Learning Representations*, 2020.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. Maxout networks. In *International conference on machine learning*, pp. 1319–1327. PMLR, 2013.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3389–3396. IEEE, 2017.

Jacot, A., Gabriel, F., and Hongler, C. Neural Tangent Kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Ji, Z. and Telgarsky, M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2019.

Ji, Z., Telgarsky, M., and Xian, R. Neural tangent kernels, transportation mappings, and universal approximation. In *International Conference on Learning Representations*, 2019.

Juditsky, A. and Nemirovski, A. S. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Advances in neural information processing systems*, pp. 1008–1014. Citeseer, 2000.

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., and Jurafsky, D. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

Massart, P. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pp. 245–303, 2000.

Micchelli, C. A., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Oymak, S. and Soltanolkotabi, M. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pp. 4951–4960. PMLR, 2019.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017a.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017b.

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

Srebro, N. and Shraibman, A. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pp. 545–560. Springer, 2005.

Srebro, N., Rennie, J., and Jaakkola, T. S. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pp. 1329–1336, 2005.

Srikant, R. and Ying, L. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pp. 2803–2830. PMLR, 2019.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Tsitsiklis, J. N. and Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.

Williams, D. *Probability with martingales*. Cambridge university press, 1991.

Xu, P. and Gu, Q. A finite-time analysis of Q-learning with neural network function approximation. In *International Conference on Machine Learning*, pp. 10555–10565. PMLR, 2020.

A. Comparison with Previous Results

Variants of Neural TD learning have been analyzed in the literature. For a quantitative comparison in terms of the required sample complexity and overparameterization bounds to achieve a given target error, please see Table 1.

The first result on the convergence of Neural TD learning was presented in (Cai et al., 2019). Their work builds upon the analysis in (Tsitsiklis & Van Roy, 1997; Bhandari et al., 2018; Du et al., 2018; Arora et al., 2019) and requires constraining the network parameter within a compact set through the ℓ_2-projection at each iteration. They prove convergence to a stationary point in a *random* function class $\mathcal{F}_{B,m}$ where m is the network width and B is a given projection radius. Consequently, the algorithm suffers from an approximation error $\epsilon_m = O(\mathbb{E}[\|V - \Pi_{\mathcal{F}_{B,m}} V\|_\mu])$, which is not explicitly bounded, and possibly non-vanishing even with increasing width and projection radius. It is shown in (Cai et al., 2019; Wang et al., 2019) that this variant of Neural TD learning with projection, equipped with a ReLU network of width $O(1/\epsilon^8)$ achieves an error $\epsilon + \epsilon_m$ after $O(1/\epsilon^4)$ iterations. Unlike (Cai et al., 2019; Wang et al., 2019), our Neural TD learning algorithms converge to the *true* value function in a provably rich function class without any approximation error. We show that the algorithms that we consider in this paper achieve improved overparameterization bounds $\widetilde{O}(1/\epsilon^6)$ and $\widetilde{O}(1/\epsilon^2)$ for a given target error ϵ, which improve the existing results by $1/\epsilon^2$ to $1/\epsilon^6$. The results with ℓ_2-regularization were generalized to deep Q-learning setting in (Xu & Gu, 2020).

In practice, projection-free (Mnih et al., 2013) and max-norm regularized (Srivastava et al., 2014; Goodfellow et al., 2013; Srebro et al., 2005) algorithms are often adopted in training neural networks because of their computational efficiency and expressive power, which we consider in this work. In contrast, the Neural TD with ℓ_2-projection considered in (Cai et al., 2019; Wang et al., 2019) can be computationally expensive for high-dimensional state-spaces as it cannot be performed in parallel.

Projection-free Neural TD learning has also been considered in (Agazzi & Lu, 2019; Brandfonbrener & Bruna, 2020); however, these works only deal with finite state-space problems in the infinite-width regime, i.e., they do not provide bounds on the amount of overparameterization required. Since these results rely on the positive definiteness of the limiting kernel, the required overparameterization is much larger than the size of the state space which negates the benefits of Neural TD learning over tabular TD learning.

Our work is related to the analysis of (stochastic) gradient descent in the NTK regime. It is shown in (Du et al., 2018; Jacot et al., 2018) that the network parameters trained by gradient descent lie inside a ball around their initialization. However, they require massive overparameterization to ensure the positive definiteness of the neural tangent kernel, which would imply finite state and width much larger than the size of the state space in Neural TD learning. To establish such a result for stochastic gradient descent (and with modest overparameterization) requires additional work, and this problem has been considered for supervised learning tasks in (Ji & Telgarsky, 2019; Oymak & Soltanolkotabi, 2019). Our paper deviates from this line of work as we consider TD learning over an infinite state space, which has significantly different dynamics than supervised learning. Firstly, the stochastic semi-gradient in TD learning update is not a real gradient (or its unbiased estimate) because of bootstrapping, while the SGD update in supervised learning instances correspond to unbiased estimates of the true gradient. This leads to significant difficulties in the analysis of projection-free algorithms. Also, the exponential tails of the cross-entropy loss in (Ji & Telgarsky, 2019) lead

to $m = O(poly(\log(1/\epsilon)))$ dependency, which is not the case for TD learning because the objective is mean-squared Bellman error.

B. Analysis of PF-NTD: Proof of Theorem 1

In this section, we will prove Theorem 1. Before starting the proof, let us define a quantity that will be central throughout the proof.

Definition 2. *For λ as given in Theorem 1, let*

$$t_1 = \inf\Big\{t > 0 : \max_{i\in[m]} \|W_i(t) - W_i(0)\|_2 > \frac{\lambda}{\sqrt{m}}\Big\}, \quad (10)$$

be the stopping time at which there exists $i \in [m]$ such that $W_i(t) \notin \mathcal{B}\big(W_i(0), \lambda/\sqrt{m}\big)$ for the first time.

Since the updates, g_t, are random in the Neural TD Learning Algorithm (see Algorithm 1), the stopping time t_1 is random, which constitutes the main challenge in the proof. As we will show, for any $t < t_1$, the drift of $W(t)$ can be controlled. Therefore, we will prove that $t_1 > T$ with high probability to prove the error bounds in Theorem 1.

Proof outline: Below, we outline the proof steps for Theorem 1.

1. First, we will prove a drift bound for $\|W(t) - \overline{W}\|_2$ which holds for all $t < t_1$ where $\overline{W} \in \mathbb{R}^{md}$ is a weight vector such that $\nabla_W^\top Q_0(x)\overline{W} \approx V(x)$ for all $x \in \mathcal{X}$.

2. In the second step, we will use the drift bound obtained in the first step in conjunction with a stopped martingale concentration argument to show that $t_1 \geq T$ occurs with high probability, thus the drift bound holds for all $t < T$ under that event.

3. Finally, we will use the drift bound again to show that the approximation error is bounded as in Theorem 1 under the high-probability event considered in Step 2.

B.1. Step 1: Lyapunov Drift bound for $W(t)$

We first prove a drift bound on the weight vector $W(t)$, a common step in the analysis of stochastic gradient descent and TD learning with function approximation (Cai et al., 2019; Ji & Telgarsky, 2019; Bhandari et al., 2018; Xu & Gu, 2020). Define the point of attraction as follows:

$$\overline{W} = \Big[W_i(0) + a_i \frac{v\big(W_i(0)\big)}{\sqrt{m}}\Big]_{i\in[m]}, \quad (11)$$

where $W(0)$ is the initial weight vector. Intuitively, $\lim_{m\to\infty} \nabla_W^\top Q_0(x)\overline{W} = V(x)$ for any $x \in \mathcal{X}$ under the symmetric initialization, which guarantees $\nabla_W^\top Q_0(x)W(0) = Q_0(x) = 0$ for all $x \in \mathcal{X}$. For error probability $\delta \in (0,1)$, recall that we define $\ell(\delta, m) = 4\sqrt{\log(m+1)} + \sqrt{\log(1/\delta)}$, and let

$$E_1 = \Big\{\sup_{x\in\mathcal{X}} \frac{1}{m}\sum_{i=1}^m \mathbb{1}\{|W_i^\top(0)x| \leq \frac{\lambda}{\sqrt{m}}\} \leq \frac{\lambda + \ell(m,\delta)}{\sqrt{m}}\Big\},$$

and $\mathcal{E}_t = E_1 \cap \{t < t_1\}$ for any $t < T$.

The following key proposition is used to establish the drift bound.

Proposition 1. *Denote $\Delta_t = r_t + \gamma Q_t(x_t') - Q_t(x_t)$ as the Bellman error. Under Assumptions 1-2, we have the following inequalities:*

(1) $\mathbb{E}\big[\Delta_t\big(Q_t(x_t) - V(x_t)\big); \mathcal{E}_t\big] \leq -(1-\gamma)z_t^2$.

(2) $\mathbb{E}\big[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big); \mathcal{E}_t\big] \leq \frac{4\bar{\nu}}{\sqrt{m}} z_t$,

(3) For $\ell(m,\delta)$ defined in Theorem 1:

$$\mathbb{E}\big[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top \overline{W}; \mathcal{E}_t\big] \leq \frac{4\big(\bar{\nu}+\lambda\big)\big(\lambda + \ell(m,\delta)\big)z_t}{\sqrt{m}}, \quad (12)$$

where $z_t = \sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}$, $\mathbb{E}$ is the expectation over random initialization and trajectory, $\mathbb{E}_t[.] = \mathbb{E}[.|\mathcal{F}_{t-1}]$ with $\mathcal{F}_{-1} = \mathcal{F}_{init}$.

The proof of Proposition 1 is given in Appendix C. The first inequality in Proposition 1 follows from the fact that the Bellman operator $\mathcal{T}$ is a contraction with respect to $\|.\|_\pi$, and V is the fixed point of $\mathcal{T}$ (Tsitsiklis & Van Roy, 1997). The second inequality holds since $\nabla_W^\top Q_0(x)\overline{W}$ turns into an empirical estimate of V with $m/2$ iid samples, where the variance of each term is at most $\bar{\nu}^2$. The last inequality is the most challenging one as it reflects the evolution of the network output over TD learning steps, and it is essential to have $W_i(t) \in \mathcal{B}(W_i(0), \lambda/\sqrt{m})$ to prove that part.

Now we present the main drift bound for the TD update.

Lemma 1 (Drift Bound)**.** *For any $t \geq 0$, we have the following inequalities:*

$$\mathbb{E}[\mathbb{E}_t\|W(t+1) - \overline{W}\|_2^2; t < t_1] \leq \mathbb{E}[\|W(t) - \overline{W}\|_2^2; t < t_1] - 2\alpha(1-\gamma)z_t^2 + \alpha^2(1+2\lambda)^2 + 8\alpha z_t\Big(\frac{\bar{\nu} + \big(\bar{\nu}+\lambda\big)\big(\lambda + \ell(m,\delta)\big)}{\sqrt{m}}\Big), \quad (13)$$

where $\overline{W}$ is as defined in (11), $z_t = \sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}$.

Lemma 1 implies that for $t < t_1$, i.e., as long as $W_i(t) \in \mathcal{B}(W_i(0), \lambda/\sqrt{m})$ for all $i \in [m]$, the drift can be made negative by sufficiently large width m and sufficiently small step-size α.

Paper	State sp.	Width	Sample complexity	Error	Regularization
(Cai et al., 2019)	General	$O(1/\epsilon^8)$	$O(1/\epsilon^4)$	$\epsilon+\epsilon_m$	ℓ_2-projection
(Wang et al., 2019)	General	$O(1/\epsilon^8)$	$O(1/\epsilon^4)$	$\epsilon+\epsilon_\infty$	ℓ_2-projection
(Agazzi & Lu, 2019)	Finite	$poly(\lvert\mathcal{X}\rvert)$	$O(\log(1/\epsilon))$	ϵ	$poly(\lvert\mathcal{X}\rvert)$ width
This paper (PF-NTD)	General	$\widetilde{O}(1/\epsilon^6)$	$O(1/\epsilon^6)$	ϵ	Early stopping
This paper (MN-NTD)	General	$\widetilde{O}(1/\epsilon^2)$	$O(1/\epsilon^4)$	ϵ	Max-norm projection

Table 1. The overparameterization and sample complexity bounds for neural TD-learning algorithms. PF-NTD denotes projection-free, MN-NTD denotes max-norm regularized Neural TD learning algorithm. $\epsilon_m = \mathbb{E}\|V - \Pi_{\mathcal{F}_{B,m}} V\|_\pi$ denotes the approximation error of the random function class $\mathcal{F}_{B,m}$ for a given value function V.

Proof. Recall that

$$\begin{aligned} g_t &= \big(r_t + \gamma Q_t(x_t') - Q_t(x_t)\big)\nabla_W Q_t(x_t) \\ &= \Delta_t \nabla_W Q_t(x_t), \end{aligned} \tag{14}$$

is the semi-gradient, where $\Delta_t = r_t + \gamma Q_t(x_t') - Q_t(x_t)$ is the Bellman error. Since $W(t+1) = W(t) + \alpha g_t$, we have the following relation:

$$\|W(t+1) - \overline{W}\|_2^2 = \|W(t) - \overline{W}\|_2^2 + 2\alpha\Big[g_t^\top \big(W(t) - \overline{W}\big)\Big] + \alpha^2 \|g_t\|_2^2.$$

We can write the expected drift in the following form:

$$\begin{aligned} &\mathbb{E}_t[\|W(t+1) - \overline{W}\|_2^2; \mathcal{E}_t] = \|W(t) - \overline{W}\|_2^2 \mathbb{1}_{\mathcal{E}_t} \\ &+ 2\alpha \underbrace{\mathbb{E}_t[g_t^\top](W(t) - \overline{W})}_{(i)} \mathbb{1}_{\mathcal{E}_t} + \alpha^2 \underbrace{\mathbb{E}_t\|g_t\|_2^2}_{(ii)} \mathbb{1}_{\mathcal{E}_t}. \end{aligned} \tag{15}$$

Bounding (i) in (15): In order to bound (i), we expand it as follows. For any $t < t_1$:

$$\begin{aligned} \mathbb{E}_t[g_t^\top \big(W(t) - \overline{W}\big)] &= \mathbb{E}_t[\Delta_t \cdot \big(Q_t(x_t) - V(x_t)\big)] \\ &+ \mathbb{E}_t[\Delta_t \cdot \big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big)] \\ &+ \mathbb{E}_t[\Delta_t \cdot \big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top \overline{W}], \end{aligned} \tag{16}$$

Then, we obtain the inequality in Lemma 1 by applying Proposition 1.

Bounding (ii) in (15): The next argument follows the proof of Lemma 4.5 in (Cai et al., 2019):

$$\begin{aligned} \|g_t\|_2 &= \|(r_t + \gamma Q_t(x_t') - Q_t(x_t))\nabla_W Q_t(x_t)\|_2, \\ &\le |r_t + \gamma Q_t(x_t') - Q_t(x_t)|, \\ &\le 1 + 2\max_{x\in\mathcal{X}} |Q_t(x)|, \\ &\le 1 + 2\|W(t) - W(0)\|_2 \le 1 + 2\lambda, \end{aligned} \tag{17}$$

where the first inequality follows since $\|\nabla_W Q_t(x)\|_2 \le 1$ for any t, x, the second inequality follows since $r(x) \in [0,1]$ for all $x \in \mathcal{X}$, and the last inequality holds since $|Q_t(x)| = |Q_t(x) - Q_0(x)| \le \|W(t) - W(0)\|_2$ and $t < t_1$. Consequently, $\|g_t\|_2^2 \le (1+2\lambda)^2$.

The result in (13) immediately follows by combining these two bounds. □

B.2. Step 2: Stopping time $t_1 \ge T$ with high probability

Now, we will use the drift result in Step 1 to show that $t_1 \ge T$ with high probability.

Lemma 2. *Under Assumptions 1-2, we have:*

$$t_1 = \inf\Big\{t > 0 : \max_{i\in[m]} \|W_i(t) - W_i(0)\|_2 > \frac{\lambda}{\sqrt{m}}\Big\} \ge T,$$

with probability at least $1 - \delta$.

Proof. First, invoking Lemma 1 with the values for T, λ and m specified in Theorem 1, we have the following inequality for any t:

$$\begin{aligned} \mathbb{E}[\mathbb{E}_t\|W(t+1) - \overline{W}\|_2^2; \mathcal{E}_t] &\le \mathbb{E}[\|W(t) - \overline{W}\|_2^2; \mathcal{E}_t] \\ &- 2\alpha(1-\gamma)z_t^2 + 2\alpha(1-\gamma)\epsilon^2 + 4\alpha(1-\gamma)\epsilon z_t, \end{aligned} \tag{18}$$

where $z_t = \sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}$. The step-size α is chosen sufficiently small so that, by (17), $\alpha^2\|g_t\|^2 \le 2\alpha(1-\gamma)\epsilon^2$.

Claim 1. *Telescoping sum of (18) over $t < T$ yields:*

$$0 \le \bar{\nu}^2 - 2\alpha(1-\gamma)\sum_{t<T}(z_t - \epsilon)^2 + 4\alpha(1-\gamma)\epsilon^2 T.$$

Proof of Claim 1. Recall the notation $\mathbb{E}_t[.] = \mathbb{E}[.|\mathcal{F}_{t-1}]$. Let $\overline{\delta}(t) = W(t) - \overline{W}$ and ζ_T be defined as:

$$\begin{aligned} \zeta_T &= \sum_{t<T}\Big(\mathbb{E}_t[\|\overline{\delta}(t+1)\|_2^2]\mathbb{1}_{\mathcal{E}_t} - \|\overline{\delta}(t+1)\|_2^2 \mathbb{1}_{\mathcal{E}_{t+1}}\Big), \\ &\ge \sum_{t<T} \mathbb{1}_{\mathcal{E}_t}\Big(\mathbb{E}_t[\|\overline{\delta}(t+1)\|_2^2] - \|\overline{\delta}(t+1)\|_2^2\Big) = \zeta_T', \end{aligned}$$

for $T \ge 1$ with $\zeta_0 = \zeta_0' = 0$, where the inequality holds since $\mathbb{1}_{\mathcal{E}_{t+1}} \le \mathbb{1}_{\mathcal{E}_t}$. Note that ζ_T' is a martingale over the filtration $\{\mathcal{F}_t\}$ since each summand constitutes a martingale difference sequence, and $\mathbb{1}_{\mathcal{E}_t} \in \mathcal{F}_{t-1}$ is predictable and nonnegative. Then, we have:

$$\sum_{t<T}\Big(\mathbb{E}_t[\|\overline{\delta}(t+1)\|_2^2] - \|\overline{\delta}(t)\|_2^2\Big)\mathbb{1}_{\mathcal{E}_t} \ge \zeta_T - \bar{\nu}^2 + \underbrace{\|W(T) - \overline{W}\|_2^2 \mathbb{1}_{\mathcal{E}_T^c}}_{\ge 0},$$

which follows from $\|W(0) - \overline{W}\|_2 \leq \bar{\nu}$. Since $\zeta_T \geq \zeta'_T$ for any $T \geq 1$, and ζ'_T is a martingale with $\zeta'_0 = 0$, we have $\mathbb{E}[\zeta_T] \geq \mathbb{E}[\zeta'_T] = 0$. Hence,

$$\sum_{t<T} \Big(\mathbb{E}[\|W(t+1)-\overline{W}\|_2^2; \mathcal{E}_t] - \mathbb{E}[\|W(t)-\overline{W}\|_2^2; \mathcal{E}_t] \Big) \geq -\bar{\nu}^2,$$

and therefore the claim follows. □

Applying Claim 1 and Jensen's inequality, we have:

$$\sum_{t<T} z_t \leq 3\epsilon T = \frac{3\bar{\nu}^2}{4\alpha(1-\gamma)\epsilon}. \tag{19}$$

This bound on the total error will be the fundamental quantity in the proof. Now, by using (19), we will show that the event $\mathcal{E}'_T = \{t_1 < T\} \cap E_1$ occurs with low probability. For any $i \in [m]$, let $\overline{g}_i(t+1) = W_i(t+1) - W_i(t)$. Then, we have:

$$\begin{aligned} \|\overline{g}_i(t+1)\|_2 &= \| \sum_{t<t_1} \overline{g}_i(t+1)\|_2, \\ &\leq \| \sum_{t<t_1} \overline{g}_i(t+1) - \sum_{t<t_1} \mathbb{E}_t[\overline{g}_i(t+1)]\|_2 \qquad (20) \\ &\quad + \| \sum_{t<t_1} \mathbb{E}_t[\overline{g}_i(t+1)]\|_2. \qquad (21) \end{aligned}$$

Bounding (20): For any t, let

$$D_{i,t} = W_i(t+1) - W_i(t) - \mathbb{E}_t[W_i(t+1) - W_i(t)], \tag{22}$$

which forms a martingale difference sequence with respect to the filtration $\mathcal{F}_t$ since $\mathbb{E}_t[D_{i,t}] = 0$. Let $X_{i,t'} = \sum_{t<t'} D_{i,t}$. Since $D_{i,t}$ is a martingale difference sequence, $X_{i,t}$ is a martingale. Thus, bounding (20) is equivalent to bounding $\|X_{i,t_1}\|_2$, under the event $\mathcal{E}'_T$. In order to achieve this, we use a concentration inequality for vector-valued martingales Theorem 2.1 in (Juditsky & Nemirovski, 2008), which is given in the following.

Proposition 2 (Concentration for Vector Martingales)**.** *Consider a martingale difference sequence $\{D_t \in \mathbb{R}^d : t \geq 0\}$, and let $X_T = \sum_{t<T} D_t$. If $\|D_t\|_2 \leq \sigma$ almost surely for all t, then for any T and $\beta > 0$, we have the following inequality:*

$$\mathbb{P}\Big(\|X_T\| \geq \big(\sqrt{2d} + \beta\sqrt{2}\big)\sigma\sqrt{T} \Big) \leq \exp(-\beta^2/2). \tag{23}$$

Since $\sup_{x\in\mathcal{X}} |Q_t(x)| \leq \|W(t) - W(0)\|_2 \leq \lambda$ for all $t < t_1$, we have $\|D_{i,t}\|_2 \leq \frac{2\alpha(1+2\lambda)}{\sqrt{m}}$. Define the stopped martingale $\widetilde{X}_{i,t} = X_{i,\min\{t,t_1\}}$, which is again a martingale with a corresponding martingale difference sequence $\widetilde{D}_{i,t}$ that satisfies $\|\widetilde{D}_{i,t}\|_2 \leq \|D_{i,t}\|_2$ (Williams, 1991). Since

$$\|X_{i,t_1}\|_2 \cdot \mathbb{I}_{\mathcal{E}'_T} \leq \|\widetilde{X}_{i,T}\|_2,$$

the following inequality holds:

$$\mathbb{P}\Big(\|X_{i,t_1}\|_2 \geq \sqrt{2}(\sqrt{d}+\beta)\frac{2\alpha(1+2\lambda)\sqrt{T}}{\sqrt{m}}; \mathcal{E}'_T \Big) \leq e^{-\beta^2/2},$$

which follows from

$$\begin{aligned} &\{\|X_{i,t_1}\|_2 \geq \sqrt{2}(\sqrt{d}+\beta)\frac{2\alpha(1+2\lambda)\sqrt{T}}{\sqrt{m}}\} \cap \mathcal{E}'_T \\ &\qquad \subset \{\|\widetilde{X}_{i,T}\|_2 \geq \sqrt{2}(\sqrt{d}+\beta)\frac{2\alpha(1+2\lambda)\sqrt{T}}{\sqrt{m}}\}, \end{aligned}$$

and

$$\mathbb{P}\Big(\|\widetilde{X}_{i,T}\|_2 \geq \sqrt{2}(\sqrt{d}+\beta)\frac{2\alpha(1+2\lambda)\sqrt{T}}{\sqrt{m}} \Big) \leq e^{-\beta^2/2},$$

by Proposition 2. Therefore, by using union bound:

$$\begin{aligned} \mathbb{P}(\|X_{i,t_1}\|_2 > \big(\sqrt{2d}+2\sqrt{\log(\frac{m}{\delta})}\big)\sqrt{\frac{T}{m}}2\alpha(1+2\lambda); \mathcal{E}'_T) \\ \leq \delta, \end{aligned} \tag{24}$$

The step-size α is chosen to satisfy

$$\big(\sqrt{2d} + 2\sqrt{\log(m/\delta)}\big)\sqrt{T} \cdot 2\alpha(1+2\lambda) \leq \lambda/2.$$

Bounding (21): Note that we can bound (21) as follows:

$$\| \sum_{t<t_1} \mathbb{E}_t[\overline{g}_i(t+1)]\mathbb{1}_{\mathcal{E}'_T}\|_2 \leq \sum_{t<t_1} \frac{2\alpha \mathbb{1}_{\mathcal{E}'_T}}{\sqrt{m}} \|Q_t - V\|_\pi, \tag{25}$$

for all $i \in [m]$ under $\mathcal{E}'_T$ since $\sup_{i,t,x} \|\nabla_{W_i} Q_t(x)\|_2 \leq 1/\sqrt{m}$. The expectation of the RHS above is bounded as follows:

$$\begin{aligned} \frac{2\alpha}{\sqrt{m}}\mathbb{E}[\sum_{t<t_1} \|Q_t - V\|_\pi \mathbb{1}_{\mathcal{E}'_T}] &\leq \frac{2\alpha}{\sqrt{m}} \sum_{t<T} \mathbb{E}[\|Q_t - V\|_\pi; \mathcal{E}_t] \\ &\leq \frac{2\alpha}{\sqrt{m}} \sum_{t<T} z_t, \end{aligned}$$

by the law of iterated expectations as the event $\{t < t_1\} \cap E_1 \in \mathcal{F}_{t-1}$ as $\|W_i(t) - W_i(0)\| \in \mathcal{F}_{t-1}$. Note that the RHS of the previous inequality is upper bounded by (19). Therefore, we have:

$$\frac{2\alpha}{\sqrt{m}}\mathbb{E}[\sum_{t<t_1} \|Q_t - V\|_\pi; \mathcal{E}'_T] \leq \frac{6T\epsilon\alpha}{\sqrt{m}}.$$

Hence, we have the following:

$$\bigcup_{i\in[m]}\Big\{\|\sum_{t<t_1}\mathbb{E}_t[\overline{g}_i(t+1)]\mathbb{1}_{\mathcal{E}'_T}\|_2>\frac{6\alpha T\epsilon}{\sqrt{m}\delta}\Big\}\cap\mathcal{E}'_T \subset\Big\{\sum_{t<t_1}\frac{2\alpha\|Q_t-V\|_\pi\mathbb{1}_{\mathcal{E}'_T}}{\sqrt{m}}>\frac{6\alpha T\epsilon}{\sqrt{m}\delta}\Big\},$$

which implies that

$$\mathbb{P}(\bigcup_{i\in[m]}\Big\{\|\sum_{t<t_1}\mathbb{E}_t[\overline{g}_i(t+1)]\|_2>\frac{6\alpha T\epsilon}{\sqrt{m}\delta}\Big\};\mathcal{E}'_T)\le\delta, \quad (26)$$

by Markov's inequality. Now, using (24) and (26) in (20) and (21), we conclude that $\mathbb{P}(\mathcal{E}'_T)\le 2\delta$. Since $\mathcal{E}_T^c=\mathcal{E}'_T\cup E_1^c$ and $\mathbb{P}(E_1^c)\le\delta$ by Lemma 3, we conclude that $\mathcal{E}_T$ holds with probability at least $1-3\delta$.

□

B.3. Step 3: Error bound

In Step 2, we have shown that the event $\{t_1\ge T\}$ occurs with high probability. Since $\mathcal{E}_T=\{t_1\ge T\}\cap E_1\subset\mathcal{E}_t$ for any $t<T$, we have the following inequality:

$$\mathbb{E}[\|Q_t-V\|_\pi;\mathcal{E}_T]\le\sqrt{\mathbb{E}[\|Q_t-V\|_\pi^2;\mathcal{E}_T]}\le z_t=\sqrt{\mathbb{E}[\|Q_t-V\|_\pi^2;\mathcal{E}_t]},$$

for any $t<T$. Consequently, by using (19) and Jensen's inequality, we have:

$$\mathbb{E}[\|\frac{1}{T}\sum_{t<T}Q_t-V\|_\pi;\mathcal{E}_T]\le\frac{1}{T}\sum_{t<T}\mathbb{E}[\|Q_t-V\|_\pi;\mathcal{E}_T]\le\frac{1}{T}\sum_{t<T}z_t\le 3\epsilon.$$

In the final step, by following similar steps as (Cai et al., 2019), we use Proposition 3 in Appendix C to show the proximity of $\overline{Q}_T$ and $\frac{1}{T}\sum_{t<T}Q_t$ to $\nabla_W^\top Q_0\overline{W}$, and conclude that $\mathbb{E}[\|\overline{Q}_T-\frac{1}{T}\sum_{t<T}Q_t\|_\pi;\mathcal{E}_T]\le\epsilon$, which implies $\mathbb{E}[\|\overline{Q}_T-V\|_\pi;\mathcal{E}_T]\le 4\epsilon$ by triangle inequality.

C. Technical Results

In this section, we will provide a uniform tail inequality which will be used in the analysis of Neural TD learning algorithms. The argument is inspired by the proof of Glivenko-Cantelli theorem (Wainwright, 2019; Bartlett, 2003).

Lemma 3. *For any $\delta\in(0,1)$ and $m\in\mathbb{N}$, let*

$$\ell_0(m,\delta)=\sqrt{8\log(m+1)}+\sqrt{\log(1/\delta)}.$$

Then, for any $\epsilon>0$, if $W_i\overset{\text{iid}}{\sim}\mathcal{N}(0,I_d)$ for all $i\in[m]$, we have:

$$\sup_{x:\|x\|_2\le 1}\frac{1}{m}\sum_{i=1}^m\mathbb{1}\{|W_i^\top x|\le\epsilon\}\le\sqrt{\frac{2}{\pi}}\epsilon+\frac{\ell_0(m,\delta)}{\sqrt{m}}, \quad (27)$$

with probability at least $1-\delta$ over the random initialization.

Proof. Let $\mathcal{X}=\{x\in\mathbb{R}^d:\|x\|_2\le 1\}$. For any $[W_i]_{i\in[m]}$, ϵ and $x\in\mathcal{X}$, let

$$Z_\epsilon(x)=\frac{1}{m}\sum_{i=1}^m\mathbb{1}\{|W_i^\top x|\le\epsilon\},$$

and

$$f(W_1,W_2,\dots,W_m)=Z_\epsilon(x).$$

For any $w,w'\in\mathbb{R}^{md}$, let

$$w_i^{\setminus j}(u)=\begin{cases}w_i, & i=j,\\ u_j, & else,\end{cases}$$

Then, we have:

$$\sup_{u\in\mathbb{R}^{md}}|f(w)-f(w^{\setminus i}(u))|\le\frac{1}{m}, \quad (28)$$

for any $i\in[m]$. Therefore, by Azuma-Hoeffding inequality, we have:

$$\sup_{x\in\mathcal{X}}|Z_\epsilon(x)-\mathbb{E}Z_\epsilon(x)|>\mathbb{E}\sup_{x\in\mathcal{X}}|Z_\epsilon(x)-\mathbb{E}Z_\epsilon(x)|+\sqrt{\frac{\log(\frac{1}{\delta})}{2m}}, \quad (29)$$

with probability at least $1-\delta$.

The following standard symmetrization argument, which follows from in Theorem 4.10 in (Wainwright, 2019), will provide an upper bound for $\mathbb{E}\sup_{x\in\mathcal{X}}|Z_\epsilon(x)-\mathbb{E}Z_\epsilon(x)|$.

Claim 2. *Let $\sigma_i\overset{\text{iid}}{\sim}$ Rademacher be a sequence of random variables independent from W. Then,*

$$\mathbb{E}\sup_{x\in\mathcal{X}}|Z_\epsilon(x)-\mathbb{E}Z_\epsilon(x)|\le 2\mathbb{E}\sup_{x\in\mathcal{X}}\Big|\frac{1}{m}\sum_{i=1}^m\sigma_i\mathbb{1}\{|W_i^\top x|\le\epsilon\}\Big|.$$

In the following, we show that there is an underlying structure, which will be the key in deriving the uniform bound. For any given $w\in\mathbb{R}^{md}$, we can order w as follows:

$$|w_{(i)}^\top x|\le|w_{(i+1)}^\top x|\Rightarrow\mathbb{1}\{|w_{(i)}^\top x|\le\epsilon\}\ge\mathbb{1}\{|w_{(i+1)}^\top x|\le\epsilon\},$$

for any $i<m$. The existence of such a transformation implies that, for any $w\in\mathbb{R}^{md}$,

$$\big[\mathbb{1}\{|w_i^\top x|\le\epsilon\}\big]_{i\in[m]}\in A\subset\{0,1\}^m,$$

where $|A| \leq m+1$ since

$$\big[\mathbb{1}\{|w_{(i)}^\top x| \leq \epsilon\}\big]_{i\in[m]} \in \{[j_i]_{i\in[m]} \in \{0,1\}^m, j_i \geq j_{i+1}\}.$$

Furthermore, for any $w \in \mathbb{R}^{md}$, $x \in \mathcal{X}$ and $\epsilon > 0$,

$$\Big\|\big[\mathbb{1}\{|w_i^\top x| \leq \epsilon\}\big]_{i\in[m]}\Big\|_2 \leq \sqrt{m}.$$

Therefore, by Massart's finite class lemma (Massart, 2000), we have:

$$\mathbb{E}\Big[\sup_x \Big|\frac{1}{m}\sum_{i=1}^m \sigma_i \mathbb{1}\{|W_i^\top x| \leq \epsilon\}\Big|\Big|[W_i]_{i\in[m]}\Big] \leq \sqrt{\frac{2\log(m+1)}{m}}. \quad (30)$$

Therefore, by the law of iterated expectations:

$$\begin{aligned}&\mathbb{E}\sup_{x\in\mathcal{X}}\Big|\frac{1}{m}\sum_{i=1}^m \sigma_i \mathbb{1}\{|W_i^\top x| \leq \epsilon\}\Big|,\\ &= \mathbb{E}\mathbb{E}\Big[\sup_x \Big|\frac{1}{m}\sum_{i=1}^m \sigma_i \mathbb{1}\{|W_i^\top x| \leq \epsilon\}\Big|\Big|[W_i]_{i\in[m]}\Big],\\ &\leq \sqrt{\frac{2\log(m+1)}{m}}.\end{aligned}$$

By using this result in conjunction with Claim 2, we obtain the following bound:

$$\mathbb{E}\sup_{x\in\mathcal{X}}|Z_\epsilon(x) - \mathbb{E}Z_\epsilon(x)| \leq 2\sqrt{\frac{2\log(m+1)}{m}}. \quad (31)$$

Now, (29) and (31) together imply that:

$$\sup_{x\in\mathcal{X}}|Z_\epsilon(x) - \mathbb{E}Z_\epsilon(x)| \leq \sqrt{\frac{8\log(m+1)}{m}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}, \quad (32)$$

holds with probability at least $1-\delta$. By Gaussian anti-concentration, it is obvious that

$$\mathbb{E}Z_\epsilon(x) = \mathbb{P}(|W_1^\top x| \leq \epsilon) \leq \sqrt{\frac{2}{\pi}}\epsilon,$$

for any $x \in \mathcal{X}$. Using this result, we conclude the proof. □

C.1. Proof of Proposition 1

Part (1) Let $\ell(\delta, m) = 2\sqrt{\log(2m+1)} + \sqrt{\log(1/\delta)}/2$,

$$E_1 = \Big\{\sup_{x\in\mathcal{X}}\frac{1}{m}\sum_{i=1}^m \mathbb{1}\{|W_i^\top(0)x| \leq \epsilon\} \leq \sqrt{\frac{2}{\pi}}\epsilon + \frac{\ell(m,\delta)}{\sqrt{m}}\Big\},$$

For any $t < T$, let $\mathcal{E}_t = E_1 \cap \{t < t_1\}$. (1) For any $t < t_1$, we have the following inequality:

$$\mathbb{E}_t[\Delta_t\big(Q_t(x_t) - V(x_t)\big)] \leq -(1-\gamma)\|Q_t - V\|_\pi^2.$$

Proof. The proof follows the strategy first proposed in (Tsitsiklis & Van Roy, 1997), and then used for the convergence proofs in (Bhandari et al., 2018; Cai et al., 2019; Xu & Gu, 2020). Let $\mathbb{E}_t[.] = \mathbb{E}[.|\mathcal{F}_{t-1}]$, i.e., the expectation is over (x_t, x_t'). Then, we have

$$\begin{aligned}&\mathbb{E}_t[\Delta_t(Q_t(x_t) - V(x_t))]\\ &\quad= \mathbb{E}_t[(\mathcal{T}Q_t(x_t) - Q_t(x_t))(Q_t(x_t) - V(x_t))],\end{aligned}$$

by taking expectation over x_t' first, which implies the following:

$$\begin{aligned}&\mathbb{E}_t\Big[\big(\mathcal{T}Q_t(x_t) - Q_t(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)\Big]\\ &= \mathbb{E}_t\Big[\big(\mathcal{T}Q_t(x_t) - \mathcal{T}V(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)\Big]\\ &\quad - \mathbb{E}_t\Big[\big(Q_t(x_t) - V(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)\Big],\end{aligned}$$

since $\mathcal{T}V(x) = V(x)$ for any $x \in \mathcal{X}$. Therefore, we have:

$$\begin{aligned}&\mathbb{E}_t\Big[\big(\mathcal{T}Q_t(x_t) - Q_t(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)\Big]\\ &\qquad\leq \eta_t - \|Q_t - V\|_\pi^2.\end{aligned}$$

where $\eta_t = \mathbb{E}_t[\big(\mathcal{T}Q_t(x_t) - \mathcal{T}V(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)]$. Since $\|.\|_\pi$ defines a norm, by Cauchy-Schwarz inequality, we have:

$$\begin{aligned}\eta_t &= \mathbb{E}_t[\big(\mathcal{T}Q_t(x_t) - \mathcal{T}V(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)]\\ &\leq \|\mathcal{T}Q_t - \mathcal{T}V\|_\pi \cdot \|Q_t - V\|_\pi.\end{aligned}$$

From Lemma 6.3.1 in (Bertsekas, 2011), we have $\|\mathcal{T}Q_t - \mathcal{T}V\|_\pi \leq \gamma\|Q_t - V\|_\pi$, which implies the result. □

Part (2) For any t, we have:

$$\mathbb{E}[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big); \mathcal{E}_t] \leq \frac{4\overline{\nu}}{\sqrt{m}}\sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}.$$

Proof. Let $\nabla_W^\top Q_0\overline{W} = \big[\nabla_W^\top Q_0(x)\overline{W}\big]_{x\in\mathcal{X}}$. Then, for any t, we have:

$$\begin{aligned}&\mathbb{E}_t[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big)]\\ &= \mathbb{E}_t[(\mathcal{T}Q_t(x_t) - Q_t(x_t))\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big)]\end{aligned}$$

By using Cauchy-Schwarz inequality, we have:

$$\begin{aligned}&\mathbb{E}_t[(\mathcal{T}Q_t(x_t) - Q_t(x_t))\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big)]\\ &\quad\leq \|\mathcal{T}Q_t - Q_t\|_\pi\sqrt{\mathbb{E}_t[(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W})^2]}.\end{aligned}$$

Then, by using the contraction of τ with respect to $\|\cdot\|_\pi$ by Lemma 6.3.1 in (Bertsekas, 2011),

$$\|\mathcal{T}Q_t - Q_t\|_\pi \leq (1+\gamma)\|Q_t - V\|_\pi.$$

By the law of iterated expectations,

$$\mathbb{E}[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big); \mathcal{E}_t] \leq (1+\gamma)\mathbb{E}[\|Q_t - V\|_\pi \mathbb{1}_{\mathcal{E}_t}\|V - \nabla_W^\top Q_0 \overline{W}\|_\pi],$$

since $\mathbb{1}_{\mathcal{E}_t} \in \mathcal{F}_{t-1}$. Hence, by Cauchy-Schwarz inequality, we have the following:

$$\mathbb{E}[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big); \mathcal{E}_t] \leq 2\sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}\sqrt{\mathbb{E}[\|V - \nabla_W^\top Q_0\overline{W}\|_\pi^2]}. \quad (33)$$

In the following, we will bound $\sqrt{\mathbb{E}[\|V - \nabla_W^\top Q_0\overline{W}\|_\pi^2]}$. For any $x \in \mathcal{X}$, we have:

$$V(x) - \nabla_W^\top Q_0(x)\overline{W} = \frac{1}{m}\sum_{i=1}^m \Big(V(x) - \widehat{V}_i(x)\Big). \quad (34)$$

where $\widehat{V}_i(x) = \mathbb{I}\{W_i^\top(0)x \geq 0\}v^\top(W_i(0))x$. Recall that $V(x) = \mathbb{E}[\mathbb{I}\{W_i^\top(0)x \geq 0\}v^\top(W_i(0))x] = \mathbb{E}[\widehat{V}_i(x)]$ by Assumption 2. Hence, for any $i \in [m]$,

$$\mathbb{E}[V(x) - \widehat{V}_i(x)] = 0,$$

and for $i, j \in [m/2]$, we have:

$$Cov\Big(\widehat{V}_i(x), \widehat{V}_j(x)\Big) \leq \mathbb{1}\{i=j\}\mathbb{E}[\|v(W_1(0))\|_2^2].$$

Under symmetric initialization, $W_i(0) = W_{i+m/2}(0)$ for all $i \in [m/2]$. Therefore, by using the above result along with Fubini's theorem, we have:

$$\begin{aligned}\mathbb{E}\|V - \nabla_W^\top Q_0\overline{W}\|_\pi^2 &= \mathbb{E}\Big[\int_{x\in\mathcal{X}}\Big(\frac{1}{m}\sum_{i=1}^m\Big(V(x) - \widehat{V}_i(x)\Big)\Big)^2\pi(dx)\Big], \\ &\leq \int_{x\in\mathcal{X}}\frac{4}{m^2}\sum_{i=1}^m\mathbb{E}\Big[\Big|V(x) - \widehat{V}_i(x)\Big|^2\Big]\pi(dx), \\ &\leq 4\int_{x\in\mathcal{X}}\frac{\mathbb{E}[\|v(W_1(0))\|_2^2]}{m}\pi(dx) \leq \frac{4\overline{\nu}^2}{m}, \quad (35)\end{aligned}$$

since $Var(\widehat{V}_i(x)) \leq \mathbb{E}[\|v(W_1(0))\|_2^2] \leq \overline{\nu}^2$ by Assumption 2 and $\|x\|_2 \leq 1$ for all $x \in \mathcal{X}$ by Assumption 1. The extra factor is due to the symmetric initialization. By substituting (35) into (33), we have:

$$\mathbb{E}\Big[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big); \mathcal{E}_t\Big] \leq \frac{4\overline{\nu}}{\sqrt{m}}\sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}.$$

□

Part (3) Let

$$\overline{U}_i = a_i\frac{v(W_i(0))}{\sqrt{m}}, i \in [m],$$

with $\overline{U} = [\overline{U}_i]_{i\in[m]}$, which implies $\overline{W} = W(0) + \overline{U}$. Note that under symmetric initialization, $\nabla_W Q_0^\top(x)W(0) = Q_0(x) = 0$ for all $x \in \mathcal{X}$. Then, for any t, we have:

$$\mathbb{E}[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top\overline{U}; \mathcal{E}_t] \leq \frac{4\bar{\nu}\big(\lambda + \ell(m,\delta)\big)}{\sqrt{m}}z_t, \quad (36)$$

and

$$\mathbb{E}[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top W(0); \mathcal{E}_t] \leq \frac{4\lambda\big(\lambda + \ell(m,\delta)\big)}{\sqrt{m}}z_t, \quad (37)$$

with probability at least $1-\delta$ over the random initialization.

Proof. In order to prove (36), we have the following bound by using Lemma 6.3.1 in (Bertsekas, 2011):

$$\mathbb{E}_t[\Delta_t \cdot \big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top\overline{U}]\mathbb{1}_{\mathcal{E}_t} \leq (1+\gamma)\|Q_t - V\|_\pi\|\nabla_W^\top Q_t\overline{U} - \nabla_W^\top Q_0\overline{U}\|_\pi \quad (38)$$

For any $x \in \mathcal{X}$, we have:

$$\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top\overline{U} = \sum_{i\in[m]}\Big(\mathbb{I}\{W_i^\top(0)x \geq 0\} - \mathbb{I}\{W_i^\top(t)x \geq 0\}\Big)\frac{v^\top(W_i(0))x}{m}.$$

Let

$$S_x(t) = \Big\{i \in [m] : \mathbb{I}\{W_i^\top(0)x \geq 0\} \neq \mathbb{I}\{W_i^\top(t)x \geq 0\}\Big\}. \quad (39)$$

For any $x \in \mathcal{X}$ and $i \in S_x(t)$, we have:

$$|W_i^\top(0)x| \leq |W_i^\top(0)x - W_i^\top(t)x| \leq \|W_i(0) - W_i(t)\|_2,$$

since $i \in S_x(t)$ implies $W_i^\top(0)x$ and $W_i^\top(t)x$ have different signs. Therefore, we have the following relation:

$$\begin{aligned}S_x(t) &\subset \Big\{i \in [m] : |W_i^\top(0)x| \leq \|W_i(0) - W_i(t)\|_2\Big\}, \\ &\subset \Big\{i \in [m] : |W_i^\top(0)x| \leq \lambda/\sqrt{m}\Big\},\end{aligned} \quad (40)$$

for any $t < t_1$. With this definition, we have:

$$\big|\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top\overline{U}\big| \leq \frac{1}{m}\sum_{i\in[m]}\mathbb{I}\{i \in S_x(t)\}\bar{\nu} \leq \frac{4\bar{\nu}}{m}\widetilde{S}(x). \quad (41)$$

since $v(w) \leq \bar{\nu}$ for any $w \in \mathbb{R}^d$ by Assumption 2, where

$$\widetilde{S}(x) = \sum_{i=1}^{m/2} \mathbb{1}\Big\{|W_i^\top(0)x| \leq \lambda/\sqrt{m}\Big\}, \tag{42}$$

for any $x \in \mathcal{X}$. By Lemma 3, under $E_1 \cap \{t < t_1\}$, we have:

$$\frac{2\widetilde{S}(x)}{m} \leq \frac{\lambda}{\sqrt{m}} + \frac{\sqrt{2}\ell(m/2,\delta)}{\sqrt{m}}. \tag{43}$$

Therefore, we can bound (38) as follows:

$$\mathbb{E}_t[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top \overline{U}]\mathbb{1}_{\mathcal{E}_t} \leq \frac{4\bar{\nu}(\lambda + \ell(m,\delta))}{\sqrt{m}}\|Q_t - V\|_\pi \mathbb{1}_{\mathcal{E}_t}.$$

By taking expectation and using Cauchy-Schwarz inequality, we obtain:

$$\mathbb{E}[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top \overline{U}] \leq \frac{4\bar{\nu}\big(\lambda + \ell(m,\delta)\big)}{\sqrt{m}} z_t.$$

In order to prove (37), we use Lemma 6.3.1 in (Bertsekas, 2011) to obtain the following inequality:

$$\begin{aligned}&\mathbb{E}_t[\Delta_t\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top W(0)] \\ &\leq 2\|Q_t - V\|_\pi \|\big(\nabla_W^\top Q_0 W(0) - \nabla_W^\top Q_t W(0)\|_\pi.\end{aligned} \tag{44}$$

For any $x \in \mathcal{X}$, we have:

$$\begin{aligned}&\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top W(0) \\ &= \frac{1}{\sqrt{m}} \sum_{i\in[m]} a_i \Big(\mathbb{I}\{W_i^\top(0)x \geq 0\} - \mathbb{I}\{W_i^\top(t)x \geq 0\}\Big) W_i^\top(0)x.\end{aligned}$$

Recall the definition of $S_x(t)$ in (39). By using triangle inequality:

$$\big|\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top W(0)\big| \leq \frac{1}{\sqrt{m}} \sum_{i\in[m]} \mathbb{I}\{i \in S_x(t)\} \cdot |W_i^\top(0)x|.$$

For any $x \in \mathcal{X}$ and $i \in S_x(t)$, we have:

$$|W_i^\top(0)x| \leq |W_i^\top(0)x - W_i^\top(t)x| \leq \|W_i(0) - W_i(t)\|_2,$$

since $i \in S_x(t)$ implies $W_i^\top(0)x$ and $W_i^\top(t)x$ have different signs. The correlation between $\mathbb{1}\{i \in S_x(t)\}$ and $\|W_i(t) - W_i(0)\|_2$ creates the main problem in the proof, which we resolve under the event $\{t < t_1\}$. For $t < t_1$, we have $\|W_i(0) - W_i(t)\|_2 \leq \lambda/\sqrt{m}$. Thus, we have:

$$\begin{aligned}\big|\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top W(0)\big| &\leq \frac{\lambda}{m} \sum_{i\in[m]} \mathbb{I}\{i \in S_x(t)\}, \\ &\leq \frac{\lambda}{m}|S_x(t)| \leq \frac{4\lambda}{m}\widetilde{S}(x),\end{aligned}$$

where $\widetilde{S}(x)$ is defined in (42). Using Lemma 3 similar to (43), under $E_1 \cap \{t < t_1\}$, we have:

$$\mathbb{E}[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top \overline{U}] \leq \frac{4\lambda\big(\lambda + \ell(m,\delta)\big)}{\sqrt{m}} z_t.$$

□

C.2. Proximity of $\overline{Q}_T$ and $\frac{1}{T}\sum_{t<T} Q_t$

In this section, we will show that the output of Algorithm 1, $\overline{Q}_T(x) = Q(x; \frac{1}{T}\sum_{t<T} W(t), a)$, is close to $\frac{1}{T}\sum_{t<T} Q_t(x)$ in expectation, which will prove that $\overline{Q}_T$ achieves the target error. The idea is based on (Cai et al., 2019), and aims to use the linear approximation $\nabla_W^\top Q_0(x)\widehat{W}(T-1)$ as an auxiliary function to show the proximity of $\overline{Q}_T$ and $\frac{1}{T}\sum_{t<T} Q_t$.

Proposition 3. *Let $\widetilde{W} \in \mathbb{R}^{md}$ be a (random) vector of parameters. Also, let $\widehat{Q}(x) = Q(x; \widetilde{W}, a)$ and $\widehat{Q}_0(x) = \nabla_W^\top Q_0(x)\widetilde{W}$ for any $x \in \mathcal{X}$, and the event $\mathcal{A} = \{\max_{i\in[m]} \|\widetilde{W}_i - W_i(0)\|_2 \leq \frac{\lambda}{\sqrt{m}}\} \cap E_1$. Then, we have the following inequality:*

$$\mathbb{E}[\|\widehat{Q} - \widehat{Q}_0\|_\pi; \mathcal{A}] \leq \frac{\lambda(\lambda + \ell(m,\delta))}{\sqrt{m}} \leq \frac{\epsilon}{2}.$$

Consequently, we have:

$$\mathbb{E}\Big[\Big\|\overline{Q}_T - \frac{1}{T}\sum_{t<T} Q_t\Big\|_\pi; \mathcal{E}_T\Big] \leq \epsilon. \tag{45}$$

Proof. First, note that the difference of $\widehat{Q}$ and $\widehat{Q}_0$ can be written as follows:

$$\begin{aligned}&|\widehat{Q}(x) - \widehat{Q}_0(x)| \\ &\leq \frac{1}{\sqrt{m}} \sum_{i\in[m]} \big|\mathbb{1}\{\widetilde{W}_i^\top x \geq 0\} - \mathbb{1}\{W_i^\top(0)x \geq 0\}\big| \cdot |\widetilde{W}_i^\top x|,\end{aligned}$$

for any $x \in \mathcal{X}$. Let

$$S_x = \Big\{i \in [m] : \mathbb{I}\{W_i^\top(0)x \geq 0\} \neq \mathbb{I}\{\widetilde{W}_i^\top x \geq 0\}\Big\}.$$

Then, we have:

$$\begin{aligned}|\widehat{Q}(x) - \widehat{Q}_0(x)| &\leq \frac{1}{\sqrt{m}} \sum_{i\in[m]} \mathbb{1}\{i \in S_x\}|\widetilde{W}_i^\top x| \\ &\leq \frac{1}{\sqrt{m}} \sum_{i\in[m]} \mathbb{1}\{i \in S_x\}\|\widetilde{W}_i - W_i(0)\|_2.\end{aligned}$$

since $i \in S_x$ implies $|\widetilde{W}_i^\top x| \leq |\widetilde{W}_i^\top x - W_i^\top(0)x| \leq \|\widetilde{W}_i - W_i(0)\|_2$. Similarly, we have: $|W_i^\top(0)x| \leq \|\widetilde{W}_i -$

$W_i(0)\|_2$. Then, we have:

$$|\widehat{Q}(x) - \widehat{Q}_0(x)|\mathbb{1}_{\mathcal{A}} \leq \frac{\lambda}{m}|S_x|\mathbb{1}_{\mathcal{A}} \leq \frac{\lambda(\lambda + \ell(m,\delta))}{\sqrt{m}}.$$

Taking the expectation and using Jensen's inequality, we have:

$$\mathbb{E}[\|\widehat{Q} - \widehat{Q}_0\|_\pi; \mathcal{A}] \leq \sqrt{\mathbb{E}[\|\widehat{Q}_T - \widehat{Q}_0\|_\pi^2; \mathcal{A}]} \leq \frac{\lambda(\lambda + \ell(m,\delta))}{\sqrt{m}},$$

which concludes the proof of the first claim.

In order to prove the second claim, consider $\widetilde{W} = \widehat{W}(T-1) = \frac{1}{T}\sum_{t<T} W(t)$, and note that $\widehat{W}(T-1) \in \mathcal{F}_{T-1}$ and $\mathcal{E}_T \subset \mathcal{A}$ by definition. Therefore, the first part implies the following:

$$\mathbb{E}[\|\overline{Q}_T - \nabla_W^\top Q_0 \widehat{W}(T-1)\|_\pi; \mathcal{E}_T] \leq \frac{\lambda(\lambda + \ell(m,\delta))}{\sqrt{m}}. \tag{46}$$

with the usual notation $\nabla_W^\top Q_0 \widehat{W}(T-1) = \big[\nabla_W^\top Q_0(x)\widehat{W}(T-1)\big]_{x\in\mathcal{X}}$. Finally, we have:

$$\mathbb{E}[\|\frac{1}{T}\sum_{t<T} Q_t - \nabla_W^\top Q_0 \widehat{W}(T-1)\|_\pi; \mathcal{E}_T] \leq \frac{1}{T}\sum_{t<T}\mathbb{E}[\|Q_t - \nabla_W^\top Q_0 W(t)\|_\pi; \mathcal{E}_T],$$

by Jensen's inequality. For any $t < T$, letting $\widetilde{W} = W(t)$, and noting that $\mathcal{E}_T \subset \mathcal{A}$, we have $\mathbb{E}[\|Q_t - \nabla_W^\top Q_0 W(t)\|_\pi; \mathcal{E}_t] \leq \frac{\lambda(\lambda+\ell(m,\delta))}{\sqrt{m}}$ by using the first part of the proposition, which implies:

$$\mathbb{E}[\|\frac{1}{T}\sum_{t<T} Q_t - \nabla_W^\top Q_0 \widehat{W}(T-1)\|_\pi; \mathcal{E}_T] \leq \frac{\lambda(\lambda + \ell(m,\delta))}{\sqrt{m}}. \tag{47}$$

Using (46), (47) and triangle inequality together, we conclude that

$$\mathbb{E}[\|\frac{1}{T}\sum_{t<T} Q_t - \overline{Q}_T\|_\pi; \mathcal{E}_T] \leq \epsilon,$$

with the choice of parameters in Theorem 1. □

D. Analysis of MN-NTD: Proof of Theorem 2

The proof of Theorem 2 consists of the same steps as Theorem 1, but it is simpler because the growth of $\|W(t) - \overline{W}\|_2$ is controlled by the max-norm constraint. In the first step, we will prove a Lyapunov drift bound.

D.1. Lyapunov Drift Bound

First, note that for any $R > 0$ and $m \in \mathbb{N}$,

$$\mathcal{G}_{m,R} = \Big\{w \in \mathbb{R}^{md} : \|W_i(0) - w_i\|_2 \leq \frac{R}{\sqrt{m}}, \forall i \in [m]\Big\},$$

is the Cartesian product of convex sets $\mathcal{G}_{m,R}^i$, which is convex. This leads to the following result.

Lemma 4. *For any $t \geq 0$ and $R \geq \bar{\nu}$, we have the following inequalities:*

$$\mathbb{E}[\|W(t+1) - \overline{W}\|_2^2; E_1] \leq \mathbb{E}[\|W(t) - \overline{W}\|_2^2; E_1] - 2\alpha(1-\gamma)z_t^2 + \alpha^2(1+2R)^2 + 8\alpha z_t\Big(\frac{\bar{\nu} + \big(\bar{\nu} + R\big)\big(R + \ell(m,\delta)\big)}{\sqrt{m}}\Big), \tag{48}$$

where $\overline{W}$ is as defined in (11), $z_t = \sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; E_1]}$.

Proof. First, note that $W(t+1) = \Pi_{\mathcal{G}_{m,R}} W(t+1/2)$ by the update rule in (8), and $\mathcal{G}_{m,R}$ is a convex set. Also, note that $R \geq \bar{\nu}$ implies $\overline{W} \in \mathcal{G}_{m,R}$. Therefore, we have:

$$\|W(t+1) - \overline{W}\|_2^2 = \|\Pi_{\mathcal{G}_{m,R}} W(t+1/2) - \Pi_{\mathcal{G}_{m,R}}\overline{W}\|_2^2, \leq \|W(t+1/2) - \overline{W}\|_2^2,$$

which follows since projection is a non-expansive operation for convex subsets. Since $W(t+1/2) = W(t) + \alpha g_t$ and $\|g_t\|_2 \leq 1 + 2R$ by (17), we have:

$$\mathbb{E}_t\|W(t+1) - \overline{W}\|_2^2 \leq \|W(t) - \overline{W}\|_2^2 + 2\alpha\mathbb{E}_t[g_t^\top](W(t) - \overline{W}) + \alpha^2(1+2R)^2.$$

Then, the proof follows by multiplying both sides by $\mathbb{1}_{E_1}$, taking expectation, and using Proposition 1 with λ replaced by R since $\|W_i(t) - W_i(0)\|_2 \leq R/\sqrt{m}$ for all $i \in [m]$ and $t_1 = \infty$. □

D.2. Error Bound

Note that by the choices of step-size α and network width m, we have:

$$\alpha^2(1+2R)^2 = \alpha(1-\gamma)\epsilon^2,$$

and

$$\frac{\bar{\nu} + \big(\bar{\nu} + R\big)\big(R + \ell(m,\delta)\big)}{\sqrt{m}} \leq \epsilon(1-\gamma)/4.$$

Using these in Lemma 4, we have:

$$\mathbb{E}[\|W(t+1) - \overline{W}\|_2^2; E_1] \leq \mathbb{E}[\|W(t) - \overline{W}\|_2^2; E_1] - \alpha(1-\gamma)\Big(z_t - \epsilon\Big)^2 + 2\alpha(1-\gamma)\epsilon^2.$$

By telescoping sum over $t = 0, 1, \ldots, T-1$, the above inequality yields:

$$\frac{1}{T}\sum_{t<T}(z_t - \epsilon)^2 \leq \frac{\mathbb{E}[\|W(0) - \overline{W}\|_2^2; E_1]}{\alpha(1-\gamma)T} + 2\epsilon^2,$$
$$\leq \frac{\bar{\nu}^2}{\alpha(1-\gamma)T} + 2\epsilon^2.$$

By using Jensen's inequality,

$$\Big(\frac{1}{T}\sum_{t<T} z_t - \epsilon\Big)^2 \leq \frac{\bar{\nu}^2}{\alpha(1-\gamma)T} + 2\epsilon^2.$$

The above inequality yields:

$$\frac{1}{T}\sum_{t<T}\mathbb{E}[\|Q_t - V\|_\pi; E_1] \leq \frac{1}{T}\sum_{t<T} z_t$$
$$\leq \frac{\bar{\nu}}{\sqrt{\alpha(1-\gamma)T}} + 3\epsilon.$$

We conclude the proof by using Proposition 3.