An NTK Theory Approach to UCB Estimation in Semi-gradient TD Learning

Yijun Wu[1] **Pascal R. van der Vaart**[1] **Moritz A. Zanger**[1] **Matthijs T. J. Spaan**[1]

[1]Delft University of Technology, Delft, 2628 XE, The Netherlands

Abstract

Efficient exploration is a central challenge in reinforcement learning (RL), especially under sparse rewards, large state-action spaces, or adversarial environment dynamics. A principled approach to exploration is to choose actions that maximize the upper bound return. Under certain conditions, this can be formulated in the Bayesian framework as exploration based on the posterior variance. In practice, the predictive variance of a neural network is often used as a surrogate for the posterior variance by techniques such as ensembles, randomized priors, and random network distillation. However, for wide neural networks, neural tangent kernel (NTK) theory proves there is a discrepancy between the predictive variance of a wide neural network trained by gradient descent and its posterior variance. This is a known issue in supervised learning and there exist methods to reconcile this discrepancy. In this work, we show that, in contrast to the supervised setting, reconciliation is not possible for semi-gradient temporal difference (TD) learning, which is a commonly used algorithm in deep RL. Instead, we derive the upper confidence bound (UCB) of the value function in semi-gradient TD learning directly and show that this UCB can be estimated by a neural network. To aid our analysis, we extend NTK theory to regularized semi-gradient TD learning.

1 INTRODUCTION

Deep reinforcement learning (RL) has many potential applications, but poor sample efficiency is hindering widespread application. Overcoming this challenge requires the agent to explore the environment in an intelligent manner, which is difficult under sparse rewards, large state-action spaces, or adversarial environment dynamics. Classical exploration strategies include hand-crafted rewards [Schmidhuber, 2010], state visitation counts [Kocsis and Szepesvári, 2006], and random policies [Sutton and Barto, 2018]. Such techniques are designed for finite MDPs or designed based on heuristics from bandit problems and novelty detection, but are often not suitable for deep reinforcement learning. Random policies are inefficient, hand-crafted rewards are difficult to design and can change the optimal policy, and state-visitation counts are infeasible in high-dimensional or continuous tasks [Ladosz et al., 2022].

In contrast, a more principled approach to exploration is optimism in the face of uncertainty [Neu and Pike-Burke, 2020]. This means taking actions that maximize the upper bound return, directing agents to regions that are promising and have high uncertainty. However, this requires estimating an upper confidence bound (UCB) of the value function. In the Bayesian framework, the UCB is, under certain conditions, equivalent to the posterior standard deviation. This motivated techniques which estimate the variance of the model's predictive distribution [Burda et al., 2019, Zanger et al., 2026a, O'Donoghue et al., 2018, Oren et al., 2025], that is, the distribution of model predictions conditioned on the training data.

However, for wide neural networks, the variance of trained models obtained with gradient descent does not match the posterior variance of networks conditioned on the dataset [Lee et al., 2019], so the predictive variance does not equal the posterior variance. To further complicate the matter, for semi-gradient temporal difference (TD) learning, there exists no posterior interpretation even for linear value function estimators. Therefore, to design theoretically sound UCB estimators in semi-gradient TD learning, it is important to directly study the value function UCB. We utilize the theory of neural tangent kernels (NTKs) [Jacot et al., 2018, Lee et al., 2019] to handle the task of analyzing the behavior of neural networks. NTK theory states that wide neural networks can be well approximated by their linearization around initialization. However, current NTK theory only

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7403–7415.

applies to supervised learning problems with convex losses.

Thus, our contribution is threefold. We show that the Bayesian approach to UCB estimation breaks down in semi-gradient TD learning as a wide neural network cannot be interpreted as a Gaussian process posterior at convergence. We prove that NTK theory remains valid for regularized semi-gradient TD learning. With this assurance, we study neural networks in the NTK regime for semi-gradient TD learning and develop a neural network estimator for the value function UCB.

We begin with a survey of related work in Section 2. Then, we introduce the notation and background for our analysis in Section 3. In Section 4, we motivate the need for our estimator by highlighting the shortcomings of predictive variance estimation in semi-gradient TD learning. We then derive the value function UCB in Section 5. To design a neural network estimator for this UCB, we first show that NTK theory extends to regularized semi-gradient TD learning in Section 6. This serves as the theoretical basis for the estimator we propose in Section 7. In Section 8, we experimentally validate our main results. Finally, we comment on the limitations and future directions of our work in Section 9.

2 RELATED WORK

In this section, we summarize the existing approaches to uncertainty estimation in RL and their limitations. We also provide an overview of NTK theory and its current applications to RL.

Classical techniques for uncertainty estimation in deep RL include pseudo-counts [Ostrovski et al., 2017], bootstrap DQN [Osband et al., 2016], random network distillation (RND) [Burda et al., 2019], and randomized prior [Osband et al., 2018]. While intuitively motivated, the efficacy of these techniques is not well understood theoretically. For contextual bandit problems [Zhou et al., 2020, Kassraie and Krause, 2022], Markov Decision Processes (MDPs) with finite time horizons [Yang et al., 2020, Jin et al., 2020], and MDPs with finite states [Strehl et al., 2006, Jin et al., 2018], the UCB of the optimal value function can be estimated directly. However, these algorithms either assume a tabular model, or require a separate model for every time step, and thus they do not transfer to the large state-action spaces and time horizons typically studied in deep RL. In this work, we consider the value function UCB for a neural network trained with the commonly used semi-gradient TD algorithm.

It is generally difficult to analyze the dynamics of a deep neural network due to its nonlinearity. Notwithstanding, under certain conditions, a wide neural network behaves similarly to its linearization about initialization. This phenomenon is described by NTK theory [Jacot et al., 2018, Lee et al., 2019] and the closely related lazy training theory [Chizat et al., 2019]. Despite the usefulness of NTK theory, its application to reinforcement learning remains an under-researched topic. Agazzi and Lu [2022] show that the lazy training regime exists even for TD learning, where a nonlinear model converges exponentially to the true value function. Brandfonbrener and Bruna [2020] demonstrate that TD learning converges for a large class of nonlinear models, including ReLU networks. Similarly, Xu and Gu [2020] prove that Q-learning converges for ReLU networks when trained to minimize the mean square projected Bellman error. However, these results are not applicable to uncertainty estimation for the sake of exploration as they all assume that the agent has access to the MDP's stationary distribution.

More recently, Zanger et al. [2026a] propose Universal Value-Function Uncertainty (UVU), which estimates the variance of a value network trained by semi-gradient TD under the linearized network assumption. UVU faces three limitations. First, Zanger et al. [2026a] assume NTK theory to hold for semi-gradient TD learning, which is an open problem. Second, semi-gradient TD does not always converge on a finite dataset even for linear models [Sutton and Barto, 2018]. Third, the predictive variance of the value network is not guaranteed to relate to the UCB. We overcome all these limitations in this work, thereby establishing the theoretical basis for the empirical effectiveness of UVU and paving the way for exploration methods motivated by NTK theory.

3 PRELIMINARIES

In this section, we introduce the notation for the neural network and semi-gradient TD. We also summarize the tools and results needed for our analysis.

3.1 NEURAL NETWORKS

Let the feature space be $\mathcal{X}$. We assume that it is a bounded subset of $\mathbb{R}^d$. Without loss of generality, let $||x||_2 \leq 1$ for all $x \in \mathcal{X}$. An m-layer neural network of width n with parameters θ is defined recursively in terms of pre-activation layers:

$$\begin{aligned} \tilde{a}_1(x;\theta) &= \frac{W_1 x}{\sqrt{d}} + b_1, \\ \tilde{a}_i(x;\theta) &= \frac{W_i a_{i-1}(x;\theta)}{\sqrt{n}} + b_i, \quad i = 2, \dots, m, \end{aligned} \tag{1}$$

and post-activation layers:

$$a_i(x;\theta) = \sigma(\tilde{a}_i(x;\theta)), \quad i = 1, \dots m-1, \tag{2}$$

where the parameter vector θ is the collection of $W_1 \in \mathbb{R}^{n\times d}$, $W_i \in \mathbb{R}^{n\times n}$ for $i = 2, \dots, m-1$, $W_m = \mathbb{R}^{1\times n}$,

$b_i \in \mathbb{R}^n$ for $i = 1, \ldots, m-1$, and $b_m \in \mathbb{R}$. All parameters are initialized independently with the standard normal distribution, $\theta_0 \sim \mathcal{N}(0, I)$. The neural network is thus defined as

$$f(x;\theta) = \tilde{a}_m(x;\theta). \tag{3}$$

The activation function, σ, is assumed to be bounded at the origin and twice-differentiable with bounded first and second derivatives. In addition, we assume that the activation function is element-wise bounded. This simplifies the analysis by ensuring that the initial network predictions are subgaussian uniformly in n.

The network gradient with respect to (w.r.t.) the parameters, $\frac{\partial f(\cdot;\theta)}{\partial \theta}$, is denoted as $J(\cdot;\theta)$, or simply $J(\cdot)$ when there is no ambiguity. Similarly, $J(\cdot;\theta_t)$ is abbreviated as $J_t(\cdot)$.

This parameterization of the neural network is known as the NTK parameterization. It guarantees that the network is, around initialization with high probability w.r.t. the network width, Lipschitz-continuous and smooth, as well as bounded at initialization with high probability.

Theorem 3.1 [Lee et al., 2019, Ordoñez et al., 2026]**.** *For all $0 < p < 1$, $R > 0$, there exists some constant K, independent of R, such that if n is sufficiently large, then $f(x;\cdot)$ is K-Lipschitz and $\frac{K}{\sqrt{n}}$-smooth over $B(\theta_0, R)$ for all $x \in \mathcal{X}$ with probability at least $1-p$.*

In addition, although the initialization is random, and thus the initial network gradient is random, the neural tangent kernel induced by the initial network gradient is deterministic in the infinite width limit.

Theorem 3.2 [Jacot et al., 2018]**.** *The NTK, $\Theta(x, x') = J_0(x)^T J_0(x')$, is deterministic in the infinite width limit.*

For supervised learning, namely regression, NTK theory states that wide networks can be well approximated by their linearization. This simplifies analysis because the network's asymptotic behavior, that is, the behavior at convergence, can be derived analytically for linearized networks. The linearized network is defined as:

$$\bar{f}(\cdot;\bar{\theta}) = f(\cdot;\theta_0) + J_0(\cdot)(\bar{\theta} - \theta_0). \tag{4}$$

Let a dataset of M samples be denoted as (X, y), where $X \in \mathbb{R}^{M \times d}$ is the data matrix and $y \in \mathbb{R}^d$ is the target vector. The following theorem gives the asymptotic expression of a wide neural network trained under a L2-regularized MSE loss.

Theorem 3.3 [Ordoñez et al., 2026]**.** *In the NTK regime and under loss $L(\theta) = \frac{1}{2}||f(X) - y||^2 + \frac{\beta}{2}||\theta - \theta_0||^2$,*

$$\begin{aligned} f(x;\theta_\infty) = f(x;\theta_0) \\ + \Theta(x, X)\Theta_\beta(X, X)^{-1}(y - f_0(X;\theta_0)), \end{aligned} \tag{5}$$

where $\Theta_\beta(X, X) = \Theta(X, X) + \beta I$.

Theorem 3.3 provides a closed-form expression for wide neural networks at convergence, thereby enabling us to analyze the predictive variance in the supervised setting.

In the remainder, we condition all statements on the event that, for a training set X of M samples, $f(X;\theta_0)$ is $\sqrt{M}K$-bounded, and that, for all $x \in \mathcal{X}$, $f(x;\cdot)$ is K-Lipschitz and $\frac{K}{\sqrt{n}}$-smooth over $B(\theta_0, R)$.

3.2 SEMI-GRADIENT TD

The L2-regularized TD loss over a dataset of M samples, (X, r, X'), is defined as:

$$L(\theta) = \frac{1}{2}||f(X;\theta) - r - \gamma f(X';\theta)||^2 + \frac{\beta}{2}||\theta - \theta_0||^2, \tag{6}$$

where $X \in \mathbb{R}^{M \times d}$ is the data matrix of state vectors, $X' \in \mathbb{R}^{M \times d}$ is the data matrix of next-state vectors, $r \in \mathbb{R}^M$ is the reward vector, $\gamma \in (0, 1)$ is the discount factor, $\beta \in (0, \infty)$ is the regularization factor, and θ_0 are the initial parameters. Denote the TD error as $\delta(\theta) = f(X;\theta) - r - \gamma f(X';\theta)$ and let $\delta(\theta_t)$ be abbreviated as δ_t. The parameters evolve according to the ODE:

$$\dot{\theta}_t = -(J_t(X)^T \delta_t + \beta(\theta_t - \theta_0)). \tag{7}$$

Let the training gradient $J_t(X)^T \delta_t + \beta(\theta_t - \theta_0)$ be abbreviated as g_t. The TD error for the linearized network is denoted as $\bar{\delta}_t = \bar{\delta}(\bar{\theta}_t)$.

Throughout this work, we handle the asymmetries in semi-gradient TD with the observation that for any square matrix A, we have $x^T A x = x^T S(A) x$ and $A^{-1} S(A) A^{-T} = S(A^{-1})$, where $S(A)$ denotes the symmetric part of A, $S(A) = \frac{A + A^T}{2}$.

3.3 GAUSSIAN PROCESS

The initial neural network, in the infinite width limit, is a zero-mean Gaussian process (GP) [Lee et al., 2018]. Therefore, the network is still a GP at convergence in both the supervised and the semi-gradient TD setting [Lee et al., 2019, Zanger et al., 2026a]. Let $\mathcal{GP}(m, k)$ denote a GP with mean m and kernel k. The distribution of initial neural networks, known as the neural network Gaussian process (NNGP), is denoted as $\mathcal{GP}(0, \Sigma)$.

In supervised learning, namely regression, when the target function lies in the reproducing kernel Hilbert space (RKHS) associated with a GP kernel, the posterior standard deviation of the GP serves as a UCB for the error between the posterior mean and the target function.

Theorem 3.4 [Rasmussen and Williams, 2006]**.** *If y has additive and i.i.d. Gaussian noise $\mathcal{N}(0, \beta)$, the posterior of*

$\mathcal{GP}(0, k)$ conditioned on (X, y) satisfies

$$m_{post}(x) = k(x, X)(k(X, X) + \beta I)^{-1} y, \quad (8)$$

$$\sigma^2_{post}(x) = k(x, x) - k(x, X)K_\beta^{-1} k(X, x), \quad (9)$$

where $K_\beta = k(X, X) + \beta I$, m_{post} is the posterior mean, and σ^2_{post} is the posterior variance.

Theorem 3.5 [Valko et al., 2013]**.** *If the target function v, belongs to the RKHS associated with the GP kernel k, and if $y = v(X) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \beta I)$, then for all $0 < p < 1$, there exists some constant C_p such that $|v - m_{post}| \leq C_p \sqrt{\sigma^2_{post}}$ with probability at least $1 - p$ where $m_{post}, \sigma^2_{post}$ are as defined in Theorem 3.4.*

Similarly, in semi-gradient TD learning, when the value function belongs to the RKHS associated with the GP kernel, the posterior standard deviation acts as a UCB for the error between the value function and the posterior mean.

Theorem 3.6 [Lederer et al., 2026]**.** *If $r = v(X) - \gamma v(X') + \epsilon$, and $\epsilon \sim \mathcal{N}(0, \beta I)$, then the posterior mean and variance of $\mathcal{GP}(0, k)$ conditioned on (X, r, X') are given by*

$$m_{post}(x) = k_\triangle(x)^T (K_\triangle + \beta I)^{-1} r, \quad (10)$$

$$\sigma^2_{post}(x) = k(x, x) - k_\triangle(x)^T (K_\triangle + \beta I)^{-1} k_\triangle(x), \quad (11)$$

where $k_\triangle(x) = k(X, x) - \gamma k(X', x)$ and $K_\triangle = k(X, X) - \gamma k(X, X') - \gamma k(X', X) + \gamma^2 k(X', X')$.

Theorem 3.7 [Lederer et al., 2026]**.** *If the value function v, belongs to the RKHS associated with the GP kernel k, and if $r = v(X) - \gamma v(X') + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \beta I)$, then for all $0 < p < 1$, there exists some constant C_p such that $|v - m_{post}| \leq C_p \sqrt{\sigma^2_{post}}$ with probability at least $1 - p$ where $m_{post}, \sigma^2_{post}$ are as defined in Theorem 3.6.*

However, Section 4 shows that the predictive variance of a wide neural network trained with gradient descent on the semi-gradient TD loss does not match the posterior variance in Equation 11.

4 PREDICTIVE AND POSTERIOR VARIANCES

In this section, we discuss the discrepancy between the predictive variance and the posterior variance in supervised learning. Using RND as an example, we summarize existing results which demonstrate that this discrepancy can be remedied for supervised learning through regularization and an appropriate choice of prior. However, for semi-gradient TD learning, we show there exist no modifications to the neural network such that it behaves, at convergence, as a sample from the posterior of a Gaussian process conditioned on the dataset. This motivates the need to analyze the UCB for semi-gradient TD directly, which is given in Section 5.

4.1 PREDICTIVE VARIANCE IN SUPERVISED LEARNING

Even though an initial neural network is a GP and the network at convergence is also GP, the GP at convergence is not the posterior of the initial GP. Thus, Theorem 3.5 does not guarantee the predictive standard deviation of the asymptotic network to be a UCB.

Theorem 4.1 [Lee et al., 2019]**.** *In the NTK regime and under unregularized MSE loss,*

$$\mathbb{V}[f(x; \theta_\infty)] = \Sigma(x, x) - 2\Theta(x, X)\Theta(X, X)^{-1}\Sigma(X, x) + \Theta(x, X)\Lambda\Theta(X, x), \quad (12)$$

where $\Lambda = \Theta(X, X)^{-1}\Sigma(X, X)\Theta(X, X)^{-1}$.

Compared to Equation 9, Theorem 4.1 shows that the discrepancy between the predictive and posterior variances is caused by the difference between the NNGP kernel and the NTK as well as the lack of a noise term. The difference between the NNGP kernel and the NTK [Jacot et al., 2018] means that the predictive variance of a neural network cannot be interpreted as the posterior variance of any GP. Furthermore, without a noise term in the kernel, posterior variance alone is insufficient for estimating the UCB. This can be seen by considering the residual between the target function and the posterior of a noiseless GP conditioned on a noisy dataset [Lyu et al., 2020]:

$$f^*(x) - m_{\text{post}}(x) = f_\perp(x) + k(x, X)k(X, X)^{-1}\epsilon, \quad (13)$$

where k is the GP kernel, m_{post} is the posterior mean of the noiseless GP, ϵ is the target noise vector, f^* is the target function belonging to the RKHS associated with k, and $f_\perp$ is a function that satisfies $|f_\perp(x)| \leq ||f^*||\sigma_{\text{post}}(x)$ with σ_{post} being the posterior variance. Since the posterior variance of a noiseless GP vanishes for all x where $k(x, \cdot) = a^T k(X, \cdot)$ for some $a \in \mathbb{R}^M$, it cannot control the component of the residual caused by data noise. However, He et al. [2020] showed that in supervised learning, these issues can be addressed through a combination of regularization, data noise, and an untrainable correction term, so that the initial network is instead a neural tangent kernel Gaussian process (NTKGP), $\mathcal{GP}(0, \Theta)$. These modifications cause the predictive standard deviation to coincide with the posterior standard deviation and thus is guaranteed to act as a UCB.

Theorem 4.2 [He et al., 2020, Ordoñez et al., 2026]**.** *In the NTK regime and under the loss $L(\theta) = \frac{1}{2}||\tilde{f}(X; \theta) - y - \epsilon||^2 + \frac{\beta}{2}||\theta - \theta_0||^2$,*

$$\mathbb{V}[\tilde{f}(X; \theta_\infty)] = \Theta(x, x) - \Theta(x, X)(\Theta(X, X) + \beta I)^{-1}\Theta(X, x), \quad (14)$$

where $\tilde{f}(x; \theta) = f(x; \theta) + \frac{\partial f(x)}{\partial \theta^{<m}}|_{\theta_0} \tilde{\theta}^{<m}$ is the corrected network with an NTKGP prior, $\tilde{\theta}$ is a set of parameters

independently initialized from θ_0, and $\theta^{<m}, \tilde{\theta}^{<m}$ denote all the parameters before layer m.

4.2 RANDOM NETWORK DISTILLATION

RND is a popular technique in deep RL to estimate the reward function uncertainty in infinite state spaces. It estimates the uncertainty about the reward function using the approximation error of a neural network trained on a set of synthetic rewards generated by another independently initialized network. Since estimating the transition reward is a supervised learning problem, RND can be viewed as a predictive variance estimator for a reward network. Indeed, RND is equivalent, in the sense of distribution, to the sample variance estimator of a size 2 ensemble.

Theorem 4.3 [Zanger et al., 2026b]**.** *Let $\theta_0, \tilde{\theta}, \psi_0^1, \psi_0^2$ be i.i.d. initial parameters. In the NTK regime, under losses $L_{RND}(\theta) = \frac{1}{2}||f(X;\theta) - f(X;\tilde{\theta})||^2$ and $L_{Ensemble}(\psi) = \frac{1}{2}||f(X;\psi) - y||^2$, the RND estimate $\frac{1}{2}(f(\cdot;\theta_\infty) - f(\cdot;\tilde{\theta}))^2$ and the ensemble sample variance estimate $\frac{1}{2}(f(\cdot;\psi_\infty^1) - f(\cdot;\psi_\infty^2))^2$ are identically distributed.*

Thus, RND has the same biasness and efficiency as the sample variance estimator over an ensemble of size 2. For RND to have a posterior interpretation, the target network needs to be from the NTKGP. Furthermore, the loss needs to be regularized and the training target requires additive noise.

Theorem 4.4. *In the NTK regime and under loss $L(\theta) = \frac{1}{2}||f(X;\theta) - f(X;\theta_0) - \tilde{f}(X) - \epsilon||^2 + \frac{\beta}{2}||\theta - \theta_0||^2$, if $\tilde{f} \sim \mathcal{GP}(0, \Theta)$ and $\epsilon \sim \mathcal{N}(0, \beta I)$, then*

$$\begin{aligned}\mathbb{E}[h(x)] = \Theta(x,x) \\ - \Theta(x,X)(\Theta(X,X) + \beta I)^{-1}\Theta(X,x),\end{aligned} \tag{15}$$

where $h(x) = (f(x;\theta_\infty) - f(x;\theta_0) - \tilde{f}(x))^2$.

Proof. See Appendix A. □

Theorem 4.4 is closely related to the results of Zanger et al. [2026b], with the only difference being the inclusion of regularization and data noise.

4.3 POSTERIOR VARIANCE IN SEMI-GRADIENT TD LEARNING

For supervised learning problems, predictive variance estimators such as RND can become theoretically sound UCB estimators through modifications that align the predictive variance with the posterior variance. However, this Bayesian approach breaks down for semi-gradient TD learning where there exists no posterior interpretation for an asymptotic network even on a noiseless dataset.

Theorem 4.5 [Zanger et al., 2026a]**.** *Assuming $f_0 \sim \mathcal{GP}(0, \Theta)$, $\beta = 0$, and that semi-gradient TD converges, the expectation over random initialization of a converged linearized neural network is*

$$\mathbb{E}[\bar{f}(x;\bar{\theta}_\infty)] = \Theta(x,X)\triangle(X)^{-1}r \tag{16}$$

and the variance is

$$\begin{aligned}\mathbb{V}[\bar{f}(x;\bar{\theta}_\infty)] = \Theta(x,x) \\ - 2\Theta(x,X)\triangle(X)^{-1}\triangle(x) \\ + \Theta(x,X)\triangle(X)^{-1}\Psi\triangle(X)^{-1}\Theta(X,x),\end{aligned} \tag{17}$$

where $\triangle(x) = \Theta(X,x) - \gamma\Theta(X',x)$ and $\Psi = \triangle(X) - \gamma\triangle(X')$.

Contrasting Equation 16 against Equation 10 and Equation 17 against Equation 11 with $\beta = 0$, there is a structural mismatch between $\Theta(X,x)$ and $k_\triangle(x) = k(X,x) - \gamma k(X',x)$. If the full gradient is used instead of the semi-gradient and if the neural network has an NTKGP prior, then this mismatch would vanish and the predictive variance would coincide with the posterior variance analogous to the supervised setting [Sasnauskas et al., 2024]. However, since this is not the case for semi-gradient TD, there is no modification to the neural network prior such that the predictive variance matches the posterior variance. Therefore, the predictive standard deviation is not guaranteed to serve as a UCB. Nevertheless, it is still possible to derive and estimate the UCB directly for semi-gradient TD learning.

5 UCB FOR L2-REGULARIZED SEMI-GRADIENT TD LEARNING

In this section, we consider the value function UCB estimated by a linear model trained via regularized semi-gradient TD. It suffices to consider a linear model because in Section 6 we show that a wide neural network can be well approximated by its linearization. Since the analysis is not restricted to linearized networks, we assume that the value function, v, belongs in some RKHS with a known feature map ϕ, such that $v(\cdot) = \phi(\cdot)^T\theta^*$ for some θ^*. Furthermore, we assume that the reward vector can be decomposed into $r = v(X) - \gamma v(X') + \epsilon$, where the noise vector, ϵ, is element-wise independent, zero-mean, and either element-wise normally distributed with variance ζ^2 or almost surely bounded in $[-\frac{\zeta}{2\sqrt{2}}, \frac{\zeta}{2\sqrt{2}}]$. We consider a linear model, $\phi(\cdot)^T\theta$, corresponding to the asymptotic solution of semi-gradient TD. Analogously to Equation 7, the parameters of a linear model in regularized semi-gradient TD evovles according to:

$$\dot{\theta} = -\phi(X)^T(\phi(X)\theta - r - \gamma\phi(X')\theta) - \beta\theta. \tag{18}$$

Thus, the asymptotic solution, which must be the equilibrium of Equation 18, is $\theta = Z_\beta^{-1}\phi(X)^T r$, where $Z_\beta = \phi(X)^T\phi(X) - \gamma\phi(X)^T\phi(X') + \beta I$. For the asymptotic

solution to exist and to simplify the analysis, we assume that β is sufficiently large such that $\lambda_{\min}(S(Z_\beta)) \geq \frac{\beta}{2}$ and $S(Z_\beta) \succeq \phi(X)^T\phi(X)$.

Theorem 5.1. *For any $0 < p < 1$, $|\phi(x)^T\theta^* - \phi(x)^T\theta| \leq K(1 + \sqrt{\ln \frac{1}{p}})u(x)$ holds uniformly in x with probability at least $1 - p$, where $u(x) = \sqrt{\phi(x)^T\phi(x) - \phi(x)^T\phi(X)^T W_\beta^{-1}(\phi(X) - \gamma\phi(X'))\phi(x)}$, $W_\beta = \phi(X)\phi(X)^T - \gamma\phi(X')\phi(X)^T + \beta I$, and K is some constant dependent on $\theta^*, \phi(X), \beta, \zeta$.*

Proof.

$$\begin{aligned}\phi(x)^T\theta^* - \phi(x)^T\theta &= \phi(x)^T Z_\beta^{-1}(Z_\beta\theta^* - \phi(X)^T r) \\ &= \phi(x)^T Z_\beta^{-1}(\beta\theta^* - \phi(X)^T\epsilon).\end{aligned}$$

So

$$\begin{aligned}|\phi(x)^T\theta^* - \phi(x)^T\theta| \leq \beta||\phi(x)^T Z_\beta^{-1}||\ ||\theta^*|| \\ + ||\phi(x)^T Z_\beta^{-1}||\ ||\phi(X)^T\epsilon||.\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}||\phi(x)^T Z_\beta^{-1}||^2 &= \phi(x)^T Z_\beta^{-1} Z_\beta^{-T}\phi(x) \\ &\leq \frac{\phi(x)^T Z_\beta^{-1} S(Z_\beta) Z_\beta^{-T}\phi(x)}{\lambda_{\min}(S(Z_\beta))} \\ &\leq \frac{2\phi(x)^T S(Z_\beta^{-1})\phi(x)}{\beta} \\ &= \frac{2\phi(x)^T Z_\beta^{-1}\phi(x)}{\beta}.\end{aligned}$$

By the Woodbury identity,

$$Z_\beta^{-1} = \frac{I - \phi(X)^T W_\beta^{-1}(\phi(X) - \gamma\phi(X'))}{\beta}.$$

So $\frac{2\phi(x)^T Z_\beta^{-1}\phi(x)}{\beta} = \frac{2u(x)^2}{\beta^2}$, which means $||\phi(x)^T Z_\beta^{-1}|| \leq \frac{\sqrt{2}}{\beta}u(x)$. Hence,

$$\beta||\phi(x)^T Z_\beta^{-1}||\ ||\theta^*|| \leq \sqrt{2}||\theta^*||u(x).$$

Because $||\phi(X)\epsilon||$ is convex and $||\phi(X)||_2$-Lipschitz w.r.t. ϵ, where ϵ is zero-mean and either Gaussian or almost surely bounded, the event $||\phi(X)\epsilon|| \leq \zeta||\phi(X)||_{HS} + \zeta||\phi(X)||_2\sqrt{-2\ln p}$ has probability at least $1 - p$ [Wainwright, 2019]. Conditioning on this event,

$$\begin{aligned}|\phi(x)^T\theta^* - \phi(x)^T\theta| \leq \sqrt{2}||\theta^*||u(x) \\ + \frac{\sqrt{2}\zeta||\phi(X)||_{HS}}{\beta}u(x) \\ + \frac{2\zeta||\phi(X)||_2}{\beta}\sqrt{\ln\frac{1}{p}}u(x).\end{aligned}$$

□

The consequence of Theorem 5.1 is that any estimator for $u(x)$, or its upper bound, serves as a UCB estimator for semi-gradient TD. Based on this, we devise a neural network that estimates an upper bound of $u(x)^2$ in Section 7.

6 NTK THEORY FOR L2-REGULARIZED SEMI-GRADIENT TD LEARNING

We seek to design a neural UCB estimator conformal to Theorem 5.1. To that end, in this section we show that wide neural networks can be well approximated by its linearization for regularized semi-gradient TD by extending the approach of Ordoñez et al. [2026]. We begin by establishing a series of technical lemmas. These lead to the main results in Theorem 6.1 and Theorem 6.2.

Lemma 6.1 (Local Boundedness of Loss and Spectral Abscissa)**.** *$||\delta(\theta)||, \lambda_{max}(S(J(X;\theta)^T J(X';\theta))) = \mathcal{O}(1)$ uniformly in $\theta \in B(\theta_0, R)$ for any $R > 0$.*

Proof. See Appendix B.1. □

By examining the dynamics of the training gradient's magnitude, it can be shown that for a wide neural network, if there is sufficient regularization such that semi-gradient TD converges, then the convergence occurs at exponential rates. This is given by the following lemma.

Lemma 6.2 (Exponential Decay of Training Gradient)**.** *For any $R > 0$, there exists constant C, independent of R, such that for sufficiently large β, n dependent on R, $||g_t|| \leq Ce^{-\frac{\beta}{2}t}$ for $0 \leq t < H$ where $H = \inf\{t|\theta_t \notin B(\theta_0, R)\}$.*

Proof. See Appendix B.2. □

In the following lemma, we show that with sufficient regularization, the parameters of a wide neural network stays bounded about its initialization.

Lemma 6.3 (Boundedness of Network Parameters)**.** *For any $R > 0$ and sufficiently large β, n dependent on R, $\sup_{t\geq 0}||\theta_t - \theta_0|| \leq R$.*

Proof. See Appendix B.3. □

In the remainder, we only consider sufficiently large β, n such that $\theta_t \in B(\theta_0, R)$ for all $t \geq 0$.

Lemma 6.4 (Dynamics of Network Gradient)**.** *For sufficiently large β, n, $||J_t(x) - J_0(x)|| = \mathcal{O}(\frac{1}{\sqrt{n}})$ uniformly in x.*

Proof. See Appendix B.4. □

The following lemma demonstrates that with sufficient regularizaiton, the parameters of a wide network remains close

to the parameters of its linearization at any point during training.

Lemma 6.5 (Dynamics of Network Parameters)**.** *For sufficiently large β, n, $\sup_{t\geq 0} ||\theta_t - \bar{\theta}_t|| = \mathcal{O}(\frac{1}{\beta\sqrt{n}})$.*

Proof. See Appendix B.5. □

Based on the lemmas, it can be shown that the discrepancy between a neural network and its linearization, at any point during training, vanishes as the width goes to infinity for semi-gradient TD learning.

Theorem 6.1. $\sup_{t\geq 0} |f(x;\theta_t) - \bar{f}(x;\bar{\theta}_t)| = \mathcal{O}(\frac{1}{\beta\sqrt{n}})$ *uniformly in x.*

Proof. For all $x \in \mathcal{X}$,

$$\begin{aligned}|f(x,\theta_t) - \bar{f}(x;\bar{\theta}_t)| \leq\ & |f(x,\theta_t) - \bar{f}(x;\theta_t)| \\ & + |\bar{f}(x,\bar{\theta}_t) - \bar{f}(x;\theta_t)|.\end{aligned}$$

$$\begin{aligned}|f(x,\theta_t) - \bar{f}(x;\theta_t)| &\leq |\int_0^t \frac{d}{d\tau}(f(x,\theta_\tau) - \bar{f}(x;\theta_\tau))d\tau| \\ &\leq \int_0^t ||J_\tau(x) - J_0(x)||\ ||\dot{\theta}_\tau|| d\tau \\ &= \mathcal{O}(\frac{1}{\beta\sqrt{n}})\end{aligned}$$

by Lemma 6.2 and Lemma 6.4. Similarly,

$$|\bar{f}(x,\bar{\theta}_t) - \bar{f}(x;\theta_t)| = ||J_0(x)||\ ||\bar{\theta}_t - \theta_t|| = \mathcal{O}(\frac{1}{\beta\sqrt{n}})$$

by Lemma 6.5. Therefore,

$$|f(x,\theta_t) - \bar{f}(x;\bar{\theta}_t)| = \mathcal{O}(\frac{1}{\beta\sqrt{n}}) + \mathcal{O}(\frac{1}{\beta\sqrt{n}}) = \mathcal{O}(\frac{1}{\beta\sqrt{n}}).$$

□

Theorem 6.1 serves as our theoretical justification for approximating a neural network by its linearization for semi-gradient TD learning when designing our UCB estimator in Section 7. Analogously to Theorem 3.3, the asymptotic behavior of a wide neural network can be analytically derived by considering its linearization. The results are given by the following theorem.

Theorem 6.2. *In the NTK regime and for sufficiently large β,*

$$f(x;\theta_\infty) = f_0(x) - J_0(x)J_0(X)^T W_\beta^{-1}\delta_0, \qquad (19)$$

where $W_\beta = J_0(X)J_0(X)^T - \gamma J_0(X')J_0(X)^T + \beta I$.

Proof. By Theorem 6.1, it suffices to consider the asymptotic behavior of the linearized network.

$$\begin{aligned}\frac{d}{dt}(\bar{\theta}_t - \theta_0) &= -J_0(X)^T\bar{\delta}_t - \beta(\bar{\theta}_t - \theta_0) \\ &= -J_0(X)^T\delta_0 - Z_\beta(\bar{\theta}_t - \theta_0),\end{aligned}$$

where $Z_\beta = J_0(X)^T J_0(X) - \gamma J_0(X)^T J_0(X') + \beta I$. So if semi-gradient TD converges, then we must have,

$$\bar{\theta}_\infty - \theta_0 = -Z_\beta^{-1}J_0(X)^T\delta_0 = -J_0(X)^T W_\beta^{-1}\delta_0$$

by the push-through identity. □

7 A STOCHASTIC NEURAL UCB ESTIMATOR

When the value function, v, belongs to the RKHS associated with the NTK, the UCB can be estimated by a wide neural network. Similar to RND and UVU, our estimator consists of a trainable network f and a target network $\tilde{f}$. We require $f_0 = f(\cdot;\theta_0) \sim \mathcal{GP}(0,\Sigma)$ and $\tilde{f} \sim \mathcal{GP}(0,\Theta)$. The trainable network is trained by semi-gradient TD with synthetic reward $r = \tilde{f}(X) - \gamma\tilde{f}(X') + f_0(X) - \gamma f_0(X')$. Our estimator is defined as

$$U^2 = \max(0, (\tilde{f} - \frac{f_\infty - f_0}{2})^2 - \frac{(f_\infty - f_0)^2}{4}), \qquad (20)$$

where $f_\infty = f(\cdot;\theta_\infty)$.

Theorem 7.1. *For all $x \in \mathcal{X}$, $\sqrt{\mathbb{E}[U^2(x)]} \geq u(x)$ in the NTK regime, where u is defined in Theorem 5.1.*

Proof. By Theorem 6.2,

$$f_\infty(x) - f_0(x) = \Theta(x,X)\triangle_\beta^{-1}\tilde{r},$$

where $\triangle_\beta = \Theta(X,X) - \gamma\Theta(X',X) + \beta I$ and $\tilde{r} = \tilde{f}(X) - \gamma\tilde{f}(X')$. Let $\tilde{U}^2 = (\tilde{f} - \frac{f_\infty - f_0}{2})^2$, then

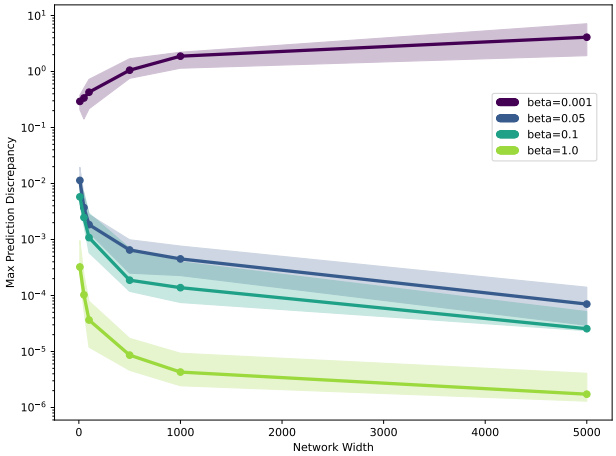
$$\tilde{U}^2(x) = \tilde{f}(x)^2 - \Theta(x,X)\triangle_\beta^{-1}\tilde{r}\tilde{f}(x) + \frac{(f_\infty(x) - f_0(x))^2}{4}.$$

Because $\mathbb{E}[\tilde{f}(x)^2] = \Theta(x,x)$, $\mathbb{E}[\tilde{r}\tilde{f}(x)] = \Theta(X,x) - \gamma\Theta(X',x)$, and $\frac{1}{4}(f_\infty(x) - f_0(x))^2$ is non-negative, therefore, $\mathbb{E}[\tilde{U}^2(x)] \geq u(x)^2$. Thus,

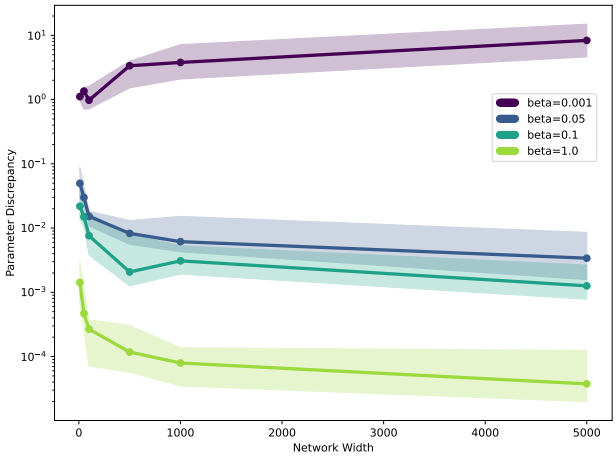
$$u(x)^2 \leq \mathbb{E}[U^2(x)] \leq \mathbb{E}[\tilde{U}^2(x)].$$

□

Theorem 7.1 shows that U^2 serves as a value function UCB in expectation over random initialization.



(a) Median (20 trials) max absolute difference between a neural network's predictions and its linearization's predictions over the test set with 95% confidence interval (shaded).



(b) Median (20 trials) Euclidean distance between a neural network's parameters and its linearization's parameters over the test set with 95% confidence interval (shaded).

Figure 1: Comparison between a neural network and its linearization for different network widths and regularization factors in semi-gradient TD.

8 NUMERICAL ANALYSIS

In this section, we experimentally demonstrate that wide neural networks can be well approximated by their linearizations about initialization and that Equation 20 is indeed an UCB estimator in expectation for semi-gradient TD.

For our experiments, we use a 2D feature space and a training set, (X, r, X'), of size 500 as well as a test set, $\tilde{X}$, of size 500. Each state vector in X is independently sampled from $\mathcal{N}(0, \frac{1}{9}I)$. X' is constructed as $X' = X + \frac{1}{3}Z$, where Z is elementwise i.i.d. standard normal. Each state vector in the test set is sampled uniformly from $[-1, 1]^2$. The target value function is defined as $v(x) = ||x||^2$ and the reward vector is defined as $r = v(X) - 0.9v(X')$. In all our experiments, we use two layer neural networks with tanh activation. To construct neural networks with the NTKGP prior, we use the method of He et al. [2020].

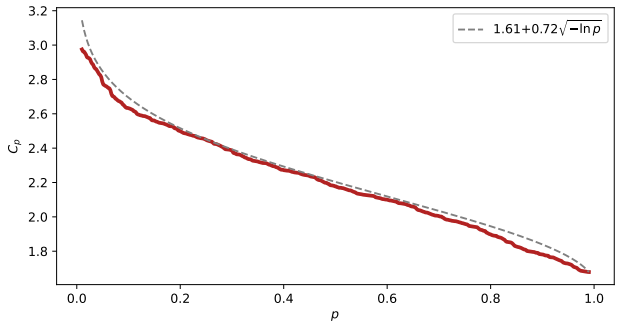


Figure 2: The relationship between empirical probability, p, and the smallest scaling factor, C_p, such that $C_p\sqrt{U^2}$ is a value function UCB over the test set with empirical probability at least $1 - p$.

To demonstrate that wide neural networks can be well approximated by their linearizations, we empirically validate Theorem 6.1 and Lemma 6.5. For each set of initial parameters θ_0, we define a neural network $f(\cdot; \theta_0)$ and a linearized network $\bar{f}(\cdot; \theta_0)$. These models are then trained on (X, r, X') until convergence. The maximum absolute difference between $f(\cdot; \theta_\infty)$ and $\bar{f}(\cdot; \bar{\theta}_\infty)$ over $\tilde{X}$, and the Euclidean distance between θ_∞ and $\bar{\theta}_\infty$ are then measured. The experiment is conducted for network widths $n = 10, 100, 500, 1000, 5000$ and regularization factors $\beta = 0.001, 0.05, 0.1, 1$. Figure 1a presents the difference in predictions w.r.t. the network width and the regularization factor. Similarly, Figure 1b presents the difference in parameters w.r.t. the network width and the regularization factor. The experiment is repeated for 20 trials.

The results of this experiment match Theorem 6.1 and Lemma 6.5. The discrepancies between a neural network and its linearization, in terms of their predictions and parameters, are inversely proportional to the network width and the regularization factor, if they are sufficiently large.

To empirically validate Theorem 5.1 and Theorem 7.1, we train a neural network, f, of width 5000 to approximate the value function using semi-gradient TD over training set (X, r, X') of 500 samples with regularization factor 0.1, where the reward is perturbed by $\mathcal{U}([-\sqrt{0.1}, \sqrt{0.1}])$ noise. This is repeated 20 times to compute the smallest scaling factor, C_p, such that $|v(x) - f_\infty(x)| \leq C_p\sqrt{U^2(x)}$ for all x in the test set with empirical probability at least $1 - p$, where U^2 is our UCB estimator given by Equation 20. Since Theorem 7.1 holds in expectation over random initialization, the predictions of U^2 are averaged over 20 runs. Figure 2 shows the relationship between p and C_p. The results agree with Theorem 5.1 which states that $C_p = \mathcal{O}(1 + \sqrt{\ln \frac{1}{p}})$.

9 CONCLUSION

This paper addresses the problem of UCB estimation in semi-gradient TD learning. We show that, in contrast to supervised learning, a neural network cannot be interpreted as a Gaussian process posterior at convergence in semi-gradient TD learning. Therefore, to design a neural network estimator for the value function UCB, we directly study the approximation error in semi-gradient TD learning. To simplify our analysis, we extend NTK theory to regularized semi-gradient TD learning. This permits us to approximate wide neural networks by their linearizations. Based on this UCB, we develop a theoretically sound neural network estimator for the value function UCB in semi-gradient TD learning.

Our work is not without limitations. We consider wide neural networks to avoid analyzing the asymptotic behavior of nonlinear models. However, in practice, neural networks are often not wide enough relative to the input dimensionality to be in the NTK regime. Furthermore, although regularization is necessary both for UCB estimation and the convergence of semi-gradient TD, it is unclear whether the regularization factor needed for convergence lies in the same regime as the noise variance. This is a nontrivial problem as it depends on the network architecture and the dataset.

For follow-up research, it is worth examining the efficiency of our UCB estimator. More concretely, the efficiency of our UCB estimator depends on the efficiency in estimating the NTKGP prior. In this work, we use the NTKGP prior estimator by He et al. [2020] as a canonical example. However, its efficiency remains unknown. It is also valuable to experimentally validate our theoretical results on benchmark problems. In general, we emphasize that the application of NTK theory to exploration remains an under-explored topic.

Acknowledgements

This project has received funding from the EU Horizon 2020 programme *Epistemic AI* under grant number 964505.

References

Andrea Agazzi and Jianfeng Lu. Temporal-difference learning with nonlinear function approximation: lazy training and mean field regimes. In *Mathematical and Scientific Machine Learning*, 2022.

David Brandfonbrener and Joan Bruna. Geometric insights into the convergence of nonlinear TD learning. In *International Conference on Learning Representations*, 2020.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

Bobby He, Balaji Lakshminarayanan, and Yee Whye Teh. Bayesian deep ensembles via the neural tangent kernel. In *Advances in Neural Information Processing Systems*, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2020.

Parnian Kassraie and Andreas Krause. Neural contextual bandits without regret. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European Conference on Machine Learning*, 2006.

Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85, 2022.

Armin Lederer, Anuj Srivastava, Marco Bagatella, and Andreas Krause. Policy search via bayesian optimization with temporal difference gaussian processes. In *International Conference on Machine Learning*, 2026.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

Yueming Lyu, Yuan Yuan, and Ivor W. Tsang. Efficient batch black-box optimization with deterministic regret bounds. *arXiv preprint arXiv:1905.10041*, 2020.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Brendan O'Donoghue, Ian Osband, Rémi Munos, and Vlad Mnih. The uncertainty Bellman equation and exploration. In *International Conference on Machine Learning*, 2018.

Sergio Calvo Ordoñez, Jonathan Plenk, Richard Bergna, Alvaro Cartea, José Miguel Hernández-Lobato, Konstantina Palla, and Kamil Ciosek. A Gaussian process view on observation noise and initialization in wide neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2026.

Yaniv Oren, Viliam Vadocz, Matthijs T. J. Spaan, and Wendelin Böhmer. Epistemic Monte Carlo tree search. In *International Conference on Learning Representations*, 2025.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*, 2016.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2018.

Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *International Conference on Machine Learning*, 2017.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Paulius Sasnauskas, Filippo Valdettaro, and Aldo A. Faisal. Wide neural network training dynamics for reinforcement learning, 2024.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2, 2010.

Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC model-free reinforcement learning. In *International Conference on Machine Learning*, 2006.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2018.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, 2013.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Pan Xu and Quanquan Gu. A finite-time analysis of Q-learning with neural network function approximation. In *International Conference on Machine Learning*, 2020.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I. Jordan. On function approximation in reinforcement learning: optimism in the face of large state spaces. In *Advances in Neural Information Processing Systems*, 2020.

Moritz A. Zanger, Max Weltevrede, Yaniv Oren, Pascal R. Van der Vaart, Caroline Horsch, Wendelin Böhmer, and Matthijs T. J. Spaan. Universal value-function uncertainties. In *International Conference on Learning Representations*, 2026a.

Moritz A. Zanger, Yijun Wu, Pascal R. van der Vaart, Wendelin Böhmer, and Matthijs T. J. Spaan. On the equivalence of random network distillation, deep ensembles, and Bayesian inference. In *Conference on Uncertainty in Artificial Intelligence*, 2026b.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In *International Conference on Machine Learning*, 2020.

An NTK Theory Approach to UCB Estimation in Semi-gradient TD Learning (Supplementary Material)

Yijun Wu[1] **Pascal R. van der Vaart**[1] **Moritz A. Zanger**[1] **Matthijs T. J. Spaan**[1]

[1]Delft University of Technology, Delft, 2628 XE, The Netherlands

A PROOF OF THEOREM 4.4

By Theorem 3.3,

$$f(x;\theta_\infty) - f(x;\theta_0) - \tilde{f}(x) = \Theta(x,X)(\Theta(X,X)+\beta I)^{-1}(\tilde{f}(X)+\epsilon) - \tilde{f}(x).$$

Since $\tilde{f} \sim \mathcal{G}P(0,\Theta)$ and $\epsilon \sim \mathcal{N}(0,\beta I)$, it follows that $\mathbb{E}[\tilde{f}(X)\tilde{f}(X)^T] = \Theta(X,X)$, $\mathbb{E}[\tilde{f}(X)\epsilon^T] = O$, and $\mathbb{E}[\epsilon\epsilon^T] = \beta I$. Therefore,

$$\begin{aligned}\mathbb{E}[(f(x;\theta_\infty) - f(x;\theta_0) - \tilde{f}(x))^2] &= \mathbb{E}[\tilde{f}(x)^2] - 2\Theta(x,X)(\Theta(X,X)+\beta I)^{-1}\mathbb{E}[\tilde{f}(X)\tilde{f}(x) + \epsilon\tilde{f}(x)] \\ &\quad + \Theta(x,X)(\Theta(X,X)+\beta I)^{-1}\mathbb{E}[\tilde{f}(X)\tilde{f}(X)^T + \epsilon\epsilon^T](\Theta(X,X)+\beta I)^{-1}\Theta(X,x) \\ &\quad + \Theta(x,X)(\Theta(X,X)+\beta I)^{-1}\mathbb{E}[\tilde{f}(X)\epsilon^T + \epsilon\tilde{f}(X)^T](\Theta(X,X)+\beta I)^{-1}\Theta(X,x) \\ &= \Theta(x,x) - \Theta(x,X)(\Theta(X,X)+\beta I)^{-1}\Theta(X,x).\end{aligned}$$

B PROOF OF TECHNICAL LEMMAS

B.1 PROOF OF LEMMA 6.1

For any $\theta \in B(\theta_0, R)$:

$$\begin{aligned}||\delta(\theta)|| &\le ||\delta(\theta_0)|| + ||\delta(\theta) - \delta(\theta_0)|| \\ &\le ||f(X;\theta_0)|| + ||r|| + \gamma||f(X';\theta_0)|| + ||f(X;\theta) - f(X;\theta_0)|| + \gamma||f(X';\theta) - f(X';\theta_0)|| \\ &\le ||r|| + (1+\gamma)\sqrt{M}K(1+R).\end{aligned}$$

Likewise,

$$\lambda_{\max}(S(J(X;\theta)^T J(X';\theta))) \le \frac{||J(X;\theta)^T J(X';\theta)||}{2} + \frac{||J(X';\theta)^T J(X;\theta)||}{2} \le ||J(X;\theta)||\,||J(X';\theta)|| \le MK^2.$$

B.2 PROOF OF LEMMA 6.2

$$\frac{d}{dt}||g_t||^2 = 2g_t^T\frac{\partial g_t}{\partial\theta}\dot{\theta}_t = -2g_t^T\frac{\partial g_t}{\partial\theta}g_t = -2g_t^T S(\frac{\partial g_t}{\partial\theta})g_t.$$

Since

$$\frac{\partial g}{\partial\theta} = \frac{\partial^2 f(X)}{\partial\theta^2}^T\delta + J(X)^T J(X) - \gamma J(X)^T J(X') + \beta I,$$

it follows that

$$\frac{d}{dt}||g_t||^2 = -2g_t^T \frac{\partial^2 f_t(X)}{\partial \theta^2}^T \delta_t g_t - 2g_t^T J_t(X)^T J_t(X) g_t - 2g_t^T(\beta I - \gamma S(J_t(X)^T J_t(X')))g_t$$
$$\leq -2\lambda_{\min}(\frac{\partial^2 f_t(X)}{\partial \theta^2}^T \delta_t)||g_t||^2 - 2\lambda_{\min}(\beta I - \gamma S(J_t(X)^T J_t(X')))||g_t||^2.$$

By Theorem 3.1, $||\frac{\partial^2 f(X)}{\partial \theta^2}|| \leq \sqrt{\frac{M}{n}}K$, so

$$\lambda_{\min}(\frac{\partial^2 f_t(X)}{\partial \theta^2}^T \delta_t) \geq -||\frac{\partial^2 f_t(X)}{\partial \theta^2}^T \delta_t||_2 \geq -\sqrt{\frac{M}{n}}K\delta_{\sup}^R,$$

where $\delta_{\sup}^R = \sup_{\theta \in B(\theta_0, R)} ||\delta(\theta)||$. Similarly,

$$\lambda_{\min}(\beta I - \gamma S(J_t(X)^T J_t(X'))) \geq \beta - \gamma \lambda_{\sup}^R,$$

where $\lambda_{\sup}^R = \sup_{\theta \in B(\theta_0, R)} \lambda_{\max}(S(J(X;\theta)^T J(X';\theta)))$. In the remainder, we only consider sufficiently large β, n such that $\beta - \gamma\lambda_{\sup}^R - \sqrt{\frac{M}{n}}K\delta_{\sup}^R \geq \frac{\beta}{2}$. Note that β does not need to go to infinity when n goes to infinity by Lemma 6.1. Under this setting,

$$\frac{d}{dt}||g_t||^2 \leq -\beta||g_t||^2.$$

By Gronwall's inequality,

$$||g_t|| \leq ||g_0||e^{-\frac{\beta}{2}t}.$$

Since $||g_0|| \leq ||J_0(X)||\,||\delta_0|| = \mathcal{O}(1)$ by Theorem 3.1, there exists a constant C independent of R such that $||g_0|| \leq C$ for sufficiently large β, n.

B.3 PROOF OF LEMMA 6.3

$$||\theta_t - \theta_0|| = ||\int_0^t \dot{\theta}_\tau d\tau|| \leq \int_0^\infty ||g_\tau|| d\tau.$$

By Lemma 6.2, there exists constant C such that, for sufficiently large β and n, $||g_\tau|| \leq Ce^{-\frac{\beta}{2}t}$, which means,

$$||\theta_t - \theta_0|| \leq \frac{2C}{\beta}.$$

The claim thus holds for $\beta \geq \frac{2C}{R}$.

B.4 PROOF OF LEMMA 6.4

Pick any $R > 0$, then, for sufficiently large β, n, $||J_t(x) - J_0(x)|| \leq \text{Lip}(J(x;\cdot))||\theta_t - \theta_0|| \leq \frac{KR}{\sqrt{n}} = \mathcal{O}(\frac{1}{\sqrt{n}})$ by Theorem 3.1 and Lemma 6.3.

B.5 PROOF OF LEMMA 6.5

$$\begin{aligned}\dot{\theta}_t &= -(J_t(X) - J_0(X))^T\delta(\theta_t) - J_0(X)^T\delta(\theta_t) - \beta(\theta_t - \theta_0)\\ &= -(J_t(X) - J_0(X))^T\delta(\theta_t) - J_0(X)^T(\delta(\theta_t) - \bar{\delta}(\theta_t)) - J_0(X)^T\bar{\delta}(\theta_t) - \beta(\theta_t - \theta_0)\\ &= -h_t - (J_0(X)^T\bar{\delta}(\theta_t) + \beta(\theta_t - \theta_0)),\end{aligned}$$

where $h_t = (J_t(X) - J_0(X))^T \delta(\theta_t) + J_0(X)^T(\delta(\theta_t) - \bar{\delta}(\theta_t))$. Then,

$$\begin{aligned}\frac{d}{dt}(\theta_t - \bar{\theta}_t) &= -h_t - J_0(X)^T(\bar{\delta}(\theta_t) - \bar{\delta}(\bar{\theta}_t)) - \beta(\theta_t - \bar{\theta}_t) \\ &= -h_t - J_0(X)^T J_0(X)(\theta_t - \bar{\theta}_t) + \gamma J_0(X)^T J_0(X')(\theta_t - \bar{\theta}_t) - \beta(\theta_t - \bar{\theta}_t).\end{aligned}$$

By the Laplace transform,

$$\begin{aligned}\theta_t - \bar{\theta}_t &= -h_t * e^{-(J_0(X)^T J_0(X) - \gamma J_0(X)^T J_0(X') + \beta I)t}, \\ ||\theta_t - \bar{\theta}_t|| &\leq \int_0^t ||h_\tau||\, ||e^{-(J_0(X)^T J_0(X) - \gamma J_0(X)^T J_0(X') + \beta I)(t-\tau)}|| d\tau.\end{aligned}$$

$||e^{-(J_0(X)^T J_0(X) - \gamma J_0(X)^T J_0(X') + \beta I)(t-\tau)}|| \leq e^{-\frac{\beta}{2}(t-\tau)}$ because $\lambda_{\min} S(J_0(X)^T J_0(X) - \gamma J_0(X)^T J_0(X') + \beta I) \geq \beta - \gamma\lambda^R_{\sup} \geq \frac{\beta}{2}$. To control h_τ, observe that $||h_t|| \leq ||J_t(X) - J_0(X)||\, ||\delta(\theta_t)|| + ||J_0(X)||\, ||\delta(\theta_t) - \bar{\delta}(\theta_t)||$.
$||J_t(X) - J_0(X)||\, ||\delta(\theta_t)||$ is controlled by

$$||J_t(X) - J_0(X)||\, ||\delta(\theta_t)|| \leq \sqrt{\frac{M}{n}} K ||\theta_t - \theta_0|| \delta^R_{\sup} \leq \sqrt{\frac{M}{n}} K R \delta^R_{\sup} = \mathcal{O}(\frac{1}{\sqrt{n}}),$$

which follows from Lemma 6.1 and Lemma 6.4. To control $||J_0(X)||\, ||\delta(\theta_t) - \bar{\delta}(\theta_t)||$, observe that

$$\begin{aligned}||\delta(\theta_t) - \bar{\delta}(\theta_t)|| &\leq ||f(X;\theta_t) - \bar{f}(X;\theta_t)|| + \gamma ||f(X';\theta_t) - \bar{f}(X';\theta_t)|| \\ &\leq || \int_0^t \frac{d}{d\tau}(f(X;\theta_\tau) - \bar{f}(X;\theta_\tau)) d\tau|| + \gamma || \int_0^t \frac{d}{d\tau}(f(X';\theta_\tau) - \bar{f}(X';\theta_\tau)) d\tau|| \\ &\leq \int_0^t ||J_\tau(X) - J_0(X)||\, ||\dot{\theta}_\tau|| d\tau + \gamma \int_0^t ||J_\tau(X') - J_0(X')||\, ||\dot{\theta}_\tau|| d\tau \\ &= \mathcal{O}(\frac{1}{\sqrt{n}}) \int_0^t e^{-\frac{\beta}{2}\tau} d\tau \\ &= \mathcal{O}(\frac{1}{\beta\sqrt{n}})\end{aligned}$$

by Lemma 6.2 and Lemma 6.4. Hence $||J_0(X)||\, ||\delta(\theta_t) - \bar{\delta}(\theta_t)|| = \mathcal{O}(\frac{1}{\beta\sqrt{n}})$. Overall,

$$||h_\tau|| = \mathcal{O}(\frac{1}{\sqrt{n}}) + \mathcal{O}(\frac{1}{\beta\sqrt{n}}) = \mathcal{O}(\frac{1}{\sqrt{n}}).$$

Therefore,

$$||\theta_t - \bar{\theta}_t|| = \mathcal{O}(\frac{1}{\sqrt{n}}) \int_0^t e^{-\frac{\beta}{2}(t-\tau)} d\tau = \mathcal{O}(\frac{1}{\beta\sqrt{n}}).$$