NEURAL CONTEXTUAL BANDITS WITH UPPER CONFIDENCE BOUND-BASED EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the stochastic contextual bandit problem, where the reward is generated from an unknown bounded function with additive noise. We propose the NeuralUCB algorithm, which leverages the representation power of deep neural networks and uses a neural network-based random feature mapping to construct an upper confidence bound (UCB) of reward for efficient exploration. We prove that, under mild assumptions, NeuralUCB achieves $\tilde{O}(\sqrt{T})$ regret, where $T$ is the number of rounds. To the best of our knowledge, our algorithm is the first neural network-based contextual bandit algorithm with near-optimal regret guarantee. Preliminary experiment results on synthetic data corroborate our theory, and shed light on potential applications of our algorithm to real-world problems.

1 INTRODUCTION

The stochastic contextual bandit problem has been extensively studied in machine learning (Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2019): at round $t \in \{1, 2, \ldots, T\}$, an agent is presented with a set of $K$ actions, each of which is associated with a $d$-dimensional feature vector. After choosing an action, the agent will receive a stochastic reward generated from some unknown distribution conditioned on the chosen action’s feature vector. The goal of the agent is to maximize the expected cumulative rewards over $T$ rounds. Contextual bandit algorithms have been applied to many real-world applications, such as personalized recommendation, advertising and Web search (e.g., Agarwal et al., 2009; Li et al., 2010).

The most studied model in the literature is linear contextual bandits (Auer, 2002; Abe et al., 2003; Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011), which assumes that the expected reward at each round is a linear function of the feature vector. Linear bandit algorithms have achieved great success in both theory and practice, such as news article recommendation (Li et al., 2010). However, the linear-reward assumption often fails to hold exactly in practice, which motivates the study of nonlinear contextual bandits (e.g., Filippi et al., 2010; Srinivas et al., 2010; Bubeck et al., 2011; Valko et al., 2013). However, they still require fairly strong assumptions on the reward function. For instance, Filippi et al. (2010) makes a generalized linear model assumption on the reward, Bubeck et al. (2011) require it to have a Lipschitz continuous property in a proper metric space, and Valko et al. (2013) assume the reward function belongs to some Reproducing Kernel Hilbert Space (RKHS).

In order to overcome the above shortcomings, deep neural networks (DNNs) (Goodfellow et al., 2016) have been introduced to learn the underlying reward function in contextual bandit problem, thanks to their strong representation power. Given the fact that DNNs enable the agent to make use of nonlinear models with less domain knowledge, existing work (Riquelme et al., 2018; Zahavy and Mannor, 2019) focuses on the idea called neural-linear bandit. More precisely, they use the first $L - 1$ layers of a DNN as a feature map, which transforms contexts from the raw input space to a low-dimensional space, usually with better representation and less frequent update. Then they learn a linear exploration policy on top of the last hidden layer of the DNN with a more frequent update. These attempts have achieved great empirical success. However, none of these work provides a theoretical guarantee on the regret of the algorithms.

In this paper, we take the first step towards provable efficient contextual bandit algorithms based on deep neural networks. Specifically, we propose a new algorithm, NeuralUCB, which uses a deep neural network to learn the underlying reward function. At the core of the algorithm is an upper
For two sequences \( \nu \) and \( \tilde{\nu} \) achieves and analyzed some of the earliest algorithms. Auer (2002) proposed SupLinRel algorithm that \( \tilde{\nu} \). For the setting with finitely many arms, Abe et al. (2003) formalized the linear bandit setting \( \tilde{\nu} \). For the other setting \( \tilde{\nu} \), the authors also provided a lower bound of \( \tilde{\nu} \). For the other setting \( \tilde{\nu} \), a few authors \( \tilde{\nu} \) proposed algorithms that achieve \( \tilde{\nu} \) regret, whereas an \( \tilde{\nu} \) lower bound is given by \( \tilde{\nu} \).

While most algorithms above are based on the idea of upper confidence bounding, it is also possible to use proper randomization to achieve strong regret guarantees, such as Thompson sampling and reward perturbation (Thompson 1933; Chapelle and Li 2011; Agrawal and Goyal 2013; Russo and Van Roy 2014; 2016; Kveton et al., 2019).

To deal with nonlinearity, generalized linear bandit has been considered, which assumes that the reward function can be written as a composition of a linear function and a link function. In particular, Filippi et al. (2010) proposed a GLM-UCB algorithm, which attains \( \tilde{\nu} \) regret. Li et al. (2017) proposed SupCB-GLM for generalized contextual bandit problems and showed a \( \tilde{\nu} \) regret that matches the lower bound. Jun et al. (2017) studied how to scale up algorithms for GLM bandits.

A few authors have also explored more general nonlinear bandits without making strong modeling assumptions. One line of work is variants of expert learning algorithms (Auer et al., 2002), which typically have a time complexity linear in the number of experts (which in many cases can be exponential in the number of parameters). Another approach is to reduce a bandit problem to supervised

**Notation:** Scalars are denoted by lower case letters, vectors by lower case bold face letters, and matrices by upper case bold face letters. For a positive integer \( k \), \( [k] \) denotes \( \{1, \ldots, k\} \). For a vector \( \theta \in \mathbb{R}^d \), we denote its \( \ell_2 \) norm by \( ||\theta||_2 = \sqrt{\sum_{i=1}^{d} \theta_i^2} \) and its \( j \)-th coordinate by \( \theta_j \). For a matrix \( A \in \mathbb{R}^{d \times d} \), we denote its spectral norm, Frobenius norm, and \( (i, j) \)-th entry by \( \|A\|_2 \), \( \|A\|_F \), and \( [A]_{i,j} \), respectively. We denote a sequence of vectors by \( \{\theta_i\}_{i=1}^T \), and similarly for matrices. For two sequences \( \{a_n\} \) and \( \{b_n\} \), we use \( a_n = O(b_n) \) to denote that there exists some constant \( C > 0 \) such that \( a_n \leq Cb_n \); \( a_n = \Omega(b_n) \) to denote that there exists some constant \( C' > 0 \) such that \( a_n \geq C'b_n \). In addition, we use \( O(\cdot) \) to hide logarithmic factors. We say a random variable \( X \) is \( \nu \)-sub-Gaussian if \( \mathbb{E}\exp(\lambda(X - \mathbb{E}X)) \leq \exp(\lambda^2\nu^2/2) \) for any \( \lambda > 0 \).

## 2 Related Work

### 2.1 Contextual Bandits

There is a line of extensive work on linear bandits (e.g., Auer 2002; Abe et al., 2003; Dani et al., 2008; Rusmevichientong and Tsitsiklis 2010; Li et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011). For the setting with finitely many arms, Abe et al. (2003) formalized the linear bandit setting and analyzed some of the earliest algorithms. Auer (2002) proposed SupLinRel algorithm that achieves \( O(\sqrt{dT}) \) regret. Chu et al. (2011) obtained the same regret with SupLinUCB that is based on LinUCB (Li et al., 2010); the authors also provided a lower bound of \( \Omega(\sqrt{dT}) \). For the other setting with infinitely many arms, a few authors (Dani et al., 2008; Rusmevichientong and Tsitsiklis 2010; Abbasi-Yadkori et al., 2011) proposed algorithms that achieve \( O(\sqrt{dT}) \) regret, whereas an \( \Omega(\sqrt{dT}) \) lower bound is given by Dani et al. (2008).

While most algorithms above are based on the idea of upper confidence bounding, it is also possible to use proper randomization to achieve strong regret guarantees, such as Thompson sampling and reward perturbation (Thompson 1933; Chapelle and Li 2011; Agrawal and Goyal 2013; Russo and Van Roy 2014; 2016; Kveton et al., 2019).

To deal with nonlinearity, generalized linear bandit has been considered, which assumes that the reward function can be written as a composition of a linear function and a link function. In particular, Filippi et al. (2010) proposed a GLM-UCB algorithm, which attains \( O(\sqrt{dT}) \) regret. Li et al. (2017) proposed SupCB-GLM for generalized contextual bandit problems and showed a \( O(\sqrt{dT}) \) regret that matches the lower bound. Jun et al. (2017) studied how to scale up algorithms for GLM bandits.
learning, such as the epoch-greedy algorithm (Langford and Zhang, 2008) that has an $O(T^{2/3})$ regret. Later, Agarwal et al. (2014) develop an algorithm that yields a near-optimal regret bound, but relies on an optimization oracle that can be expensive. A third approach uses nonparametric modeling, such as Gaussian processes and kernels (Srinivas et al., 2010; Krause and Ong, 2011; Valko et al., 2013). More specifically, Srinivas et al. (2010) assumed that the reward function is generated from a Gaussian process with known mean and covariance functions. They proposed a GP-UCB algorithm which achieves $O(\sqrt{T/\gamma_T})$ regret, where $\gamma_T$ is the maximum information gain. Krause and Ong (2011) assumed the reward function is defined over the context-arm joint space and proposed a Contextual GP-UCB. Valko et al. (2013) assumed that the reward function lies in a RKHS with bounded RKHS norm. They proposed a SupKernelUCB algorithm and showed an $\tilde{O}(\sqrt{dT})$ regret, where $d$ is effective dimension of the kernel that can be seen as a generalized notion of the dimension of contexts. There is also work focusing on bandit problems in general metric space with Lipschitz continuous property on the context (Kleinberg et al., 2008; Bubeck et al., 2011).

### 2.2 Neural Networks

Different lines of research have been done to provide theoretical understandings of DNNs from different aspects. For example, to understand how the expressive power of DNNs are related to their architecture, Telgarsky (2015; 2016); Liang and Srikanth (2016); Yarotsky (2017; 2018); Hanin (2017) showed that deep neural networks can express more function classes than shallow networks. Lu et al. (2017); Hanin and Sellke (2017) suggested that the width of neural networks is crucial to improve the expressive power of neural networks.

For the optimization of DNNs, a series of work have been proposed to show that (stochastic) gradient descent can find the global minima of training loss (Li and Liang, 2018; Du et al., 2019b; Allen-Zhu et al., 2019; Du et al., 2019a; Zou et al., 2019; Zou and Ouyang, 2019). For the generalization of DNNs, a series of work (Daniely, 2017; Cao and Gu, 2019a,b; Arora et al., 2019) shows that by using (stochastic) gradient descent, the parameters of a DNN are located in a particular regime and the generalization bound of DNNs can be characterized by the best function in the corresponding neural tangent kernel space (Jacot et al., 2018).

### 3 Problem Setting

We consider the stochastic $K$-armed contextual bandit problem, where the total number of rounds $T$ is known. At round $t \in [T]$, the agent observes the $t$th context consisting of $K$ feature vectors: $\{x_{t,a} \in \mathbb{R}^d \mid a \in [K]\}$. The agent selects an action $a_t$ and receive a reward $r_{t,a_t}$. For simplicity, we denote $\{x^T_{1}\times_{i=1}^{K}\} = \{x_{1,1}, x_{1,2}, \ldots, x_{T,K}\}$.

Our goal is to maximize the following pseudo regret (or regret for short):

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} (r_{t,a^*_t} - r_{t,a_t})\right], \quad (3.1)$$

where $a^*_t$ is the optimal action at round $t$ that maximizes the expected reward, i.e., $a^*_t = \arg\max_{a \in [K]} \mathbb{E}[r_{t,a}]$.

This work makes the following assumption on reward generation: for any round $t$,

$$r_{t,a_t} = h(x_{t,a_t}) + \xi_t, \quad (3.2)$$

where $h$ is some unknown function satisfying $0 \leq h(x) \leq 1$ for any $x$, and $\xi_t$ is $\nu$-sub-Gaussian noise conditioned on $x_{t,a_t}, \ldots, x_{t-1,a_t-1}$. Note that the $\nu$-sub-Gaussian noise assumption for $\xi_t$ is a standard assumption in stochastic bandit literature (e.g., Abbasi-Yadkori et al., 2011); in particular, any bounded noise satisfies such an assumption. Our reward function class contains linear functions, generalized linear functions, Gaussian processes, and kernel functions with bounded RKHS norm over a bounded domain.

In order to learn the reward function $h$ in (3.2), we propose to use a fully connected deep neural networks with depth $L \geq 2$:

$$f(x; \theta) = \sqrt{m}W_L\sigma(W_{L-1}\sigma(\cdots\sigma(W_1x))), \quad (3.3)$$
We present in Algorithm 1 our algorithm, NeuralUCB. The key idea is to use a deep neural network of $\theta$ function by layer is the same (i.e., $p$ with $W$). For $W$, we initialize $W_l$ as $\left( \begin{array}{cc} W & 0 \\ 0 & \mathbf{W} \end{array} \right)$, where each entry of $W$ is generated independently from $N(0, 4/m)$. 

For $W_L$, we initialize $W_l$ as $(\mathbf{w}^T, -\mathbf{w}^T)$, where each entry of $\mathbf{w}$ is generated independently from $N(0, 2/m)$. At round $t$, Algorithm 1 observes the contexts for all actions, $\{x_{t,a}\}_{a=1}^K$. First, it computes the upper confidence bound, $U_{t,a}$, based on context $x_{t,a}$, the current neural network parameter $\theta_{t-1}$, and a scaling factor $\gamma_{t-1}$. It then chooses action $a_t$ with the largest $U_{t,a}$, and receives the corresponding reward $r_{t,a}$. At the end of round $t$, Algorithm updates $\theta_t$ by applying Algorithm 2 to (approximately) minimize $L(\theta)$ using gradient descent, and updates $\gamma_t$. We choose gradient descent in Algorithm 2 for the simplicity of analysis, although the training method can be replaced by more efficient algorithms like stochastic gradient descent with a more involved analysis (Allen-Zhu et al. 2019; Zou et al. 2019).

Algorithm 1 NeuralUCB

1: Input: Number of rounds $T$, regularization parameter $\lambda$, exploration parameter $\nu$, confidence parameter $\delta$, norm parameter $S$, step size $\eta$, number of gradient descent steps $J$, network width $m$, network depth $L$.
2: Initialization: Randomly initialize $\theta_0$ as described in the text
3: Initialize $Z_0 = \lambda \mathbf{I}$, $b_0 = 0$
4: for $t = 1, \ldots, T$ do
5:   Observe $\{x_{t,a}\}_{a=1}^K$
6:   for $a = 1, \ldots, K$ do
7:      Compute $U_{t,a} = f(x_{t,a}; \theta_{t-1}) + \gamma_{t-1} \sqrt{\nabla g(x_{t,a}; \theta_{t-1})^\top Z_{t-1}^{-1} g(x_{t,a}; \theta_{t-1}) / m}$
8:      Let $a_t = \arg\max_{a \in [K]} U_{t,a}$
9:   end for
10: Play $a_t$ and observe reward $r_{t,a_t}$
11: Let $\theta_t = \operatorname{TrainNN}(\lambda, \nu, J, m, \{x_{t,a}\}_{a=1}^K, \{r_{t,a}\}_{a=1}^K, \theta_0)$
12: Compute $Z_t = Z_{t-1} + g(x_{t,a_t}; \theta_t) g(x_{t,a_t}; \theta_t) / m$
13: Compute $\gamma_t = \sqrt{1 + C_1 m^{-1/6} \sqrt{\log mL^4 t^{7/6} \lambda^{-7/6}} + C_2 m^{-1/6} \sqrt{\log mL^3 t^{5/3} \lambda^{-1/6}} + 2 \log \delta + \sqrt{\lambda S}}$
14: end for

Comparison with Existing Algorithms Here, we compare NeuralUCB with other neural network based contextual bandit algorithms. Allesiardo et al. (2014) proposed NeuralBandit which consists of $K$ neural networks. It uses a committee of networks to compute the score of each action and chooses the action by $\epsilon$-greedy policy. In contrast, our NeuralUCB uses upper confidence bound based exploration, which is more effective than $\epsilon$-greedy. In addition, our algorithm only uses one neural network instead of $K$ neural networks, thus can be computationally more efficient.
Thus, (4.1) only requires to compute vector inner product. Based on \( d_{t,a} \), we update \( \det(Z_t) \) by matrix determinant lemma (Golub and Van Loan, 1996):

\[
\det(Z_t) = \det \left[ Z_{t-1} + g(x_{t,a}; \theta_t)g(x_{t,a}; \theta_t)^\top / m \right] \\
= \left[ 1 + g(x_{t,a}; \theta_t)^\top Z_{t-1}^{-1} g(x_{t,a}; \theta_t) / m \right] \det(Z_{t-1}).
\] (4.1)

Note that (4.1) only requires to compute vector inner product between \( g(x_{t,a}; \theta_t) \) and \( d_{t,a} \), which only requires \( O(p) \) time. Now we show how to compute \( d_{t,a} \) efficiently. By the definition of \( d_{t,a} \), we have

\[
g(x_{t,a}; \theta_t) = Z_{t-1}^{-1} d_{t,a} = \left( \lambda I + \sum_{i=0}^{t-1} g(x_{i,a}; \theta_t)g(x_{i,a}; \theta_t)^\top / m \right) d_{t,a}.
\]

Thus, \( d_{t,a} \) is the global minimizer of the following convex optimization problem:

\[
\min_{d \in \mathbb{R}^p} \left\| \left( \lambda I + \sum_{i=0}^{t-1} g(x_{i,a}; \theta_t)g(x_{i,a}; \theta_t)^\top / m \right) d - g(x_{t,a}; \theta_t) \right\|_2^2,
\]

and we can use (stochastic) gradient descent to find \( d_{t,a} \) efficiently, whose update formula can be written as follows:

\[
d^{(i+1)} = d^{(i)} - \eta \left( \lambda I + \sum_{i=0}^{t-1} g(x_{i,a}; \theta_t)g(x_{i,a}; \theta_t)^\top / m \right) \\
\left[ \left( \lambda I + \sum_{i=0}^{t-1} g(x_{i,a}; \theta_t)g(x_{i,a}; \theta_t)^\top / m \right) d^{(i)} - g(x_{t,a}; \theta_t) \right],
\]

which again only requires to compute vector inner products.
5 Regret Analysis

In this section, we present a regret analysis for Algorithm 1. Recall that \( \{x^i\}_{i=1}^{TK} \) is the collection of all \( \{x_{t,a}\} \). Since our regret analysis is built upon the recently proposed neural tangent kernel matrix (Jacot et al., 2018), we start with its formal definition.

**Definition 5.1** (Jacot et al., 2018; Cao and Gu, 2019a). For a set of contexts \( \{x^i\}_{i=1}^{TK} \), define

\[
H^{(1)}_{i,j} = \Sigma^{(1)}_{i,j} = \langle x^i, x^j \rangle, \quad A^{(l)}_{i,j} = \left( \Sigma^{(l)}_{i,j} \Sigma^{(l)}_{j,j} \right),
\]

\[
\Sigma^{(l+1)}_{i,j} = 2E_{(u,v)\sim N(0, A^{(l)}_{i,j})} \sigma(u)\sigma(v),
\]

\[
H^{(l+1)}_{i,j} = 2H^{(l)}_{i,j} E_{(u,v)\sim N(0, A^{(l)}_{i,j})} \sigma'(u)\sigma'(v) + \Sigma^{(l+1)}_{i,j}.
\]

Then, \( H = (H^{(L)} + \Sigma^{(L)})/2 \) is called the neural tangent kernel (NTK) matrix on the context set.

As shown in Definition 5.1, the Gram matrix of the NTK on the contexts \( \{x^i\}_{i=1}^{TK} \) for L-layer neural networks is defined in a recursive way from the input layer all the way to the output layer of the neural network. For more details about the derivation of Definition 5.1 please refer to Jacot et al. (2018).

**Assumption 5.2.** \( H \succeq \lambda_0 I \). Moreover, for any \( 1 \leq i \leq TK \), \( \|x^i\|_2 = 1 \) and \( |x^i| = |x^i|_{j+d/2} \).

The first part of the assumption says that the neural tangent kernel matrix is non-singular, a mild assumption commonly made in the related literature (Du et al., 2019a; Arora et al., 2019; Cao and Gu, 2019a). It can be satisfied as long as no two contexts in \( \{x^i\}_{i=1}^{TK} \) are parallel.

Next we define the effective dimension \( \tilde{d} \) of the neural tangent kernel matrix on contexts \( \{x^i\}_{i=1}^{TK} \).

**Definition 5.3.** The effective dimension \( \tilde{d} \) of the neural tangent kernel matrix on contexts \( \{x^i\}_{i=1}^{TK} \) is defined as

\[
\tilde{d} = \frac{\log \det(I + H/\lambda)}{\log(1 + TK/\lambda)}.
\]  

**Remark 5.4.** The notion of effective dimension was introduced by Valko et al. (2013) for analyzing kernel contextual bandits, which was defined by the eigenvalues of any kernel matrix restricted on the given contexts. We adapt a similar but different definition of Yang and Wang (2019), which was used for the analysis of kernel-based Q-learning. Suppose the effective dimension of the reproducing kernel Hilbert space induced by the given kernel is \( \tilde{d} \) and the feature mapping \( \psi \) induced by the given kernel satisfies \( \|\psi(x)\|_2 \leq 1 \) for any \( x \in \mathbb{R}^d \). Then, it is easy to verify that we always have \( \tilde{d} \leq \tilde{d} \); see Appendix A.1 for details. Intuitively, \( \tilde{d} \) measures how quickly the eigenvalues of \( H \) decay, and it only depends on \( T \) logarithmically in certain specific cases (Valko et al., 2013).

Now we are ready to present the main result, which provides the regret bound \( R_T \) of Algorithm 1.

**Theorem 5.5.** Let \( \tilde{d} \) be the effective dimension defined in Definition 5.3. Let \( h = [h(x^i)]_{i=1}^{TK} \in \mathbb{R}^{TK} \). Suppose there are constants \( C_1, C_2 > 0 \), such that for any \( \delta \in (0, 1) \), if

\[
m = \text{poly}(T, L, K, \lambda^{-1}, \lambda_0^{-1}, S^{-1}, \log(1/\delta)), \quad \eta = C_1(mTL + m\lambda)^{-1},
\]

\[
\lambda \geq \max\{1, S^{-2}\}, \quad S \geq 2h^{-1/2}H^{-1/2} \tilde{h}. \]

Then, with probability at least \( 1 - \delta \) over the random initialization of \( \Theta_0 \), the regret of Algorithm 1 satisfies

\[
R_T \leq 3\sqrt{T} \sqrt{\tilde{d} \log(1 + TK/\lambda)} + 2 \left[ \sqrt{\tilde{d} \log(1 + TK/\lambda)} + 2 - 2 \log \delta \right]
\]

\[
+ 2\sqrt{\lambda \ln S + C_2(1 - \lambda/(TL))^d} \sqrt{T/\lambda} + 1.
\]
Remark 5.6. It is worth noting that, simply applying results for linear bandit to our algorithm would lead to a linear dependence of $p$ or $\sqrt{p}$ in the regret. Such a bound is vacuous since in our setting $p$ would be very large compared with the number of rounds $T$ and the input context dimension $d$.

In contrast, our regret bound only depends on $d$, which is much smaller than $p$.

Remark 5.7. Our regret bound (5.2) has a term $(1 - \lambda/(TL))^J \sqrt{T/\lambda}$, which characterizes the optimization error of Algorithm 2. More specifically, it suffices to set $J = \log(\lambda S/\sqrt{T})/(\lambda/(TL)) = O(TL/\lambda)$ (independent of $m$) such that $(1 - \lambda/(TL))^J \sqrt{T/\lambda} \leq \sqrt{\lambda S}$, and therefore the optimization error is dominated by $\sqrt{\lambda S}$.

Remark 5.8. Treating $\nu$ and $\lambda$ as constants, and taking $S = \sqrt{h^T H^{-1} h}$ and $J = \log(\lambda S/\sqrt{T})/(\lambda/(TL))$, the regret bound (5.2) becomes $R_T = O\left(\sqrt{dT \max(d, h^T H^{-1} h)}\right)$. Specifically, suppose $h$ belongs to a RKHS $\mathcal{H}$ induced by NTK and it can be represented by $h(x) = \langle \psi(x), \beta^* \rangle$, where $\psi(x)$ is the feature mapping induced by NTK, for some $\beta^*$ in $\mathcal{H}$ (Valko et al. 2013). Then we can show that $\|h\|_\mathcal{H} = \|\beta^*\|_2 \geq \sqrt{h^T H^{-1} h}$; see Appendix A.2 for more details. Thus our regret bound can be further written as $R_T = O\left(\sqrt{dT \max(d, \|h\|_\mathcal{H})}\right)$. The high-probability result in Theorem 5.5 can be used to obtain a bound on the expected regret.

Corollary 5.9. Under the same conditions in Theorem 5.5 we have

$$E[R_T] \leq 3\sqrt{T} \left( \sqrt{d \log (1 + TK/\lambda)} + 2 \nu \sqrt{d \log (1 + TK/\lambda)} + 2 + 2 \log T \right) + 2 \sqrt{\lambda S} + C_1 (1 - \lambda/(TL))^J \sqrt{T/\lambda} + 2.$$

6 Proof of Main Results

This section provides the proof of Theorem 5.5. We first point out several technical challenges in this proof:

- We do not make parametric assumptions on the reward function as some previous work (Filippi et al. 2010; Chu et al. 2011; Abbasi-Yadkori et al. 2011). Furthermore, unlike the fixed feature mapping used in Valko et al. (2013), NeuralUCB uses neural network $f(x; \theta_i)$ and its gradient $g(x; \theta_i)$ as a dynamic feature mapping depending on $\theta_i$. These differences make the regret analysis of NeuralUCB more difficult.

- In practice, the neural network is often overparametrized, which implies $m$ (and thus $p$) is very large. Thus, we need to make sure the regret bound is independent of $m$.

The two challenges above are addressed by the following technical lemmas. Their proofs are gathered in the appendix.

Lemma 6.1. There exists some constant $\tilde{C} > 0$ such that for any $\delta \in (0, 1)$, if $m \geq CT^4K^4L^6 \log(T^2K^2L/\delta)/\lambda_0^4$, then with probability at least $1 - \delta$ over the random initialization of $\theta_0$, there is a $\theta^* \in \mathbb{R}^p$ such that

$$h(x^i) = g(x^i; \theta_0), \theta^* - \theta_0, \sqrt{m} \|\theta^* - \theta_0\|_2 \leq \sqrt{2h^T H^{-1} h},$$  \hfill (6.1)

for all $i \in [TK]$.

Lemma 6.1 suggests that with high probability, the reward function restricted on $\{x^i\}_{i=1}^{TK}$ can be regarded as a linear function of $g(x^i; \theta_0)$ parameterized by $\theta^* - \theta_0$, where $\theta^*$ lies in a ball centered at $\theta_0$. Note that here $\theta^*$ is not a ground truth parameter for the reward function. Instead, it is introduced only for the sake of analysis. Equipped with Lemma 6.1, we can utilize existing results on linear bandits (Abbasi-Yadkori et al. 2011) to show that with high probability, $\theta^*$ lies in the sequence of confidence sets.

Lemma 6.2. There exist constants $\{\tilde{C}_i\}_{i=1}^2 > 0$ such that for any $\delta \in (0, 1)$, if $\eta \leq \tilde{C}_1 (TM + m\lambda)^{-1}$ and $m \geq \tilde{C}_2 \max\{T^7\lambda^{-7} L^{41} (\log m)^9, \lambda^{-1/2} L^{-3/2} \log (TKL^2/\delta)^{3/2}\}$, then with probability at least $1 - \delta$ over the random initialization of $\theta_0$, we have $\|\theta_t - \theta_0\|_2 \leq 2 \sqrt{T/(m\lambda)}$ and $\|\theta^* - \theta_t\|_2 \leq \gamma_1(t)/\sqrt{m}$ for all $t \in [T]$. 
Lemma 6.3. Denote \( a_t^* = \arg\max_{a \in [K]} h(x_{t,a}) \). There exist constants \( \{ \bar{C}_i \}_{i=1}^3 > 0 \) such that for any \( \delta \in (0,1) \), if \( \eta \leq \bar{C}_1 (TmL + m\lambda)^{-1}, m \geq \bar{C}_2 \max \{ T^7\lambda^{-7}L^{21}(\log m)^3, \lambda^{-3/2} (\log(TKL^2/\delta))^{3/2} \} \), then with probability at least \( 1 - \delta \) over the random initialization of \( \theta_0 \), we have

\[
\begin{align*}
& h(x_{t,a_t^*}) - h(x_{t,a_t}) \\
\leq & ~ 2\gamma_{t-1} \min \left\{ \|g(x_{t,a_t}; \theta_{t-1})/\sqrt{m}\|_{Z_{t-1}}^2, 1 \right\} \\
& + \bar{C}_3 \left( Sm^{-1/6} \sqrt{\log m} T^{7/6} \lambda^{-1/6} L^{7/2} + m^{-1/6} \sqrt{\log m} T^{5/3} \lambda^{-2/3} L^3 \right),
\end{align*}
\]

Lemma 6.3 gives upper bounds for \( h(x_{t,a_t^*}) - h(x_{t,a_t}) \), which helps us to bound the regret \( R_T \). It is worth noting that \( \gamma_t \) has a term \( \log \det Z_t \). A trivial upper bound of \( \log \det Z_t \) would result in a dependence on the neural network width \( m \), since the dimension of \( Z_t \) is \( p = md + m^2(L - 2) + m \). The next lemma establishes an upper bound which is independent of \( m \), and is only related to effective dimension \( \bar{d} \). The dependence on \( \bar{d} \) is similar to Lemma 4 of Valko et al. (2013), but the proof is different as we use a different notion of effective dimension.

Lemma 6.4. There exist constants \( \{ \bar{C}_i \}_{i=1}^3 > 0 \) such that for any \( \delta \in (0,1) \), if \( \eta \leq \bar{C}_1 (TmL + m\lambda)^{-1}, m \geq \bar{C}_2 \max \{ T^7\lambda^{-7}L^{21}(\log m)^3, T^6K^6L^6(\log(TKL^2/\delta))^{3/2} \} \), then with probability at least \( 1 - \delta \) over the random initialization of \( \theta_0 \), we have

\[
\begin{align*}
\sqrt{\sum_{t=1}^T \gamma_{t-1}^2 \min \left\{ \|g(x_{t,a_t}; \theta_{t-1})/\sqrt{m}\|_{Z_{t-1}}^2, 1 \right\}} \\
\leq & ~ \sqrt{\bar{d} \log(1 + TK/\lambda) + 1 + \bar{C}_3 m^{-1/6} \sqrt{\log m} T^{4/3} \lambda^{-1/6} \left[ \sqrt{1 + \bar{C}_3 m^{-1/6} \sqrt{\log m} T^{4/3} \lambda^{-1/6}} - 2 \log \delta + \sqrt{\lambda} S \right]} \\
& + \bar{C}_3 \left[ (1 - \eta m\lambda)^{\theta} \sqrt{T/\lambda} + m^{-1/6} \sqrt{\log m} T^{7/2} \lambda^{-5/3} (1 + \sqrt{T/\lambda}) \right].
\end{align*}
\]

With above lemmas, we begin to prove Theorem 5.5

Proof of Theorem 5.5. The total regret \( R_T \) can be bounded as follows:

\[
\begin{align*}
R_T = & ~ \sum_{t=1}^T \left[ h(x_{t,a_t^*}) - h(x_{t,a_t}) \right] \\
\leq & ~ 2\sum_{t=1}^T \gamma_{t-1} \min \left\{ \|g(x_{t,a_t}; \theta_{t-1})/\sqrt{m}\|_{Z_{t-1}}^2, 1 \right\} \\
& + \bar{C}_1 \left( Sm^{-1/6} \sqrt{\log m} T^{13/6} \lambda^{-1/6} L^{7/2} + m^{-1/6} \sqrt{\log m} T^{8/3} \lambda^{-2/3} L^3 \right),
\end{align*}
\]

\[
\leq 2 \sum_{t=1}^T \gamma_{t-1} \min \left\{ \|g(x_{t,a_t}; \theta_{t-1})/\sqrt{m}\|_{Z_{t-1}}^2, 1 \right\} \\
& + \bar{C}_1 \left( Sm^{-1/6} \sqrt{\log m} T^{13/6} \lambda^{-1/6} L^{7/2} + m^{-1/6} \sqrt{\log m} T^{8/3} \lambda^{-2/3} L^3 \right)
\]

\[
\leq 3\sqrt{T} \sqrt{\bar{d} \log(1 + TK/\lambda)} + 2 \left[ \sqrt{\bar{d} \log(1 + TK/\lambda)} + 2 - 2 \log \delta \right] \\
& + 2\sqrt{\lambda} S + C_2 (1 - \eta m\lambda)^{\theta} \sqrt{T/\lambda} + 1,
\]

where \( \bar{C}_1, C_2 > 0 \) are constants, the first inequality holds due to Lemma 6.3, the second inequality holds Cauchy-Schwarz inequality, the third inequality holds due to Lemma 6.4, and choosing large enough \( m \). This completes our proof. \( \square \)
Meanwhile, it is worth noting that NeuralUCB and Neural $\epsilon$-Greedy achieve a sublinear regret of actions $h$ though the network width $m$ is not as large as our theory suggests. For $\gamma_1$ in NeuralUCB, we choose $\gamma_1 = \gamma$ in the experiment. We do a grid search over $\{0.01, 0.1, 1, 10\}$ for $\gamma$. For NeuralUCB, we choose $\nu = 1, \lambda = 1, \delta = 0.1$, and we do a grid search over $\{0.01, 0.1, 1, 10\}$ for hyper-parameter $S$ to choose the best $S$ for comparison. For Neural UCBB0, we choose $\nu = 1, \lambda = 1, \delta = 0.1$, and we do a grid search over $\{0.01, 0.1, 1, 10\}$ for hyper-parameter $S$ to choose the best $S$ for comparison. For Neural $\epsilon$-Greedy and Neural $\epsilon$-Greedy0, we do a grid search for $\epsilon$ over $\{0.001, 0.01, 0.1, 0.2\}$ and choose the best $\epsilon$ for comparison. For all the algorithms, we repeat the experiment for 10 runs and report the averaged results for comparison. For TrainNN, we choose the step size $\eta = 0.1$. To accelerate the training process, we update the parameter $\theta_t$ by TrainNN every 50 rounds. We use stochastic gradient descent with batch size 50 and set $J = t$ at $t$-th round of NeuralUCB and Neural $\epsilon$-Greedy.

We plot the cumulative regret of compared algorithms in Figure 1 for reward function $h \in \{h_1, h_2, h_3\}$. We can see that due to the nonlinearity of reward function $h$, LinUCB fails to learn the true reward function and hence achieve an almost linear regret, as expected. In contrast, thanks to the neural network representation and efficient exploration, NeuralUCB achieves a sublinear regret which is much lower than that of LinUCB. The performance of Neural $\epsilon$-Greedy is in-between. This suggests that while Neural $\epsilon$-greedy can capture the nonlinearity of the underlying reward function, $\epsilon$-Greedy based exploration is not as effective as UCB based exploration. This confirms the effectiveness of NeuralUCB for contextual bandit problems with any bounded (nonlinear) reward function. Meanwhile, it is worth noting that NeuralUCB and Neural $\epsilon$-Greedy outperform NeuralUCB0 and Neural $\epsilon$-Greedy0, which suggests that using deep neural networks to predict the reward function is better than using a fixed feature mapping associated with NTK. Another observation is that although the network width $m$ in the experiment is not as large as our theory suggests, NeuralUCB still
achieves a sublinear regret for nonlinear reward functions. We leave it as a future work to investigate the impact of $m$ on regret.

8 CONCLUSIONS AND FUTURE WORK

In this work, we proposed a new algorithm NeuralUCB for stochastic contextual bandit problems based on neural networks. We show that for arbitrary bounded reward function, our algorithm achieves $\tilde{O}(d^{1/2}T)$ regret bound. Our preliminary experiment results on synthetic data corroborate our theoretical findings. In the future, we are interested in a systematic empirical evaluation of NeuralUCB on real world datasets, and compare it with the state-of-the-art neural network based contextual bandit algorithms (without provable guarantee in regret) (Riquelme et al. [2018]; Zahavy and Mannor [2019]). Another interesting direction is provably efficient exploration with neural network using other strategies like Thompson sampling.

REFERENCES


A Proof of Additional Results in Section 5

A.1 Verification of Remark 5.4

Suppose there exists a mapping $\psi : \mathbb{R}^d \to \mathbb{R}^\hat{d}$ satisfying $\|\psi(x)\|_2 \leq 1$ which maps any context $x \in \mathbb{R}^d$ to the Hilbert space $\mathcal{H}$ associated with the Gram matrix $\mathbf{H} \in \mathbb{R}^{TK \times TK}$ over contexts $\{x^i\}_{i=1}^{TK}$. Then $\mathbf{H} = \Psi^T \Psi$, where $\Psi = [\psi(x^1), \ldots, \psi(x^{TK})] \in \mathbb{R}^{d \times TK}$. Thus, we can bound the effective dimension $\hat{d}$ as follows

$$\hat{d} = \frac{\log \det(\mathbf{I} + \mathbf{H}/\lambda)}{\log(1 + TK/\lambda)} = \frac{\log \det(\mathbf{I} + \Psi^T \Psi / \lambda)}{\log(1 + TK/\lambda)} \leq \hat{d} \cdot \frac{\log \|\mathbf{I} + \Psi^T \Psi / \lambda\|_2}{\log(1 + TK/\lambda)},$$

where the second equality holds due to the fact that $\det(\mathbf{I} + \mathbf{A}^T \mathbf{A} / \lambda) = \det(\mathbf{I} + \mathbf{A} \mathbf{A}^T / \lambda)$ holds for any matrix $\mathbf{A}$, and the inequality holds since $\det \mathbf{A} \leq \|\mathbf{A}\|_2^2$ for any $\mathbf{A} \in \mathbb{R}^{d \times \hat{d}}$. Clearly, $\hat{d} \leq \hat{d}$ as long as $\|\mathbf{I} + \Psi^T \Psi / \lambda\|_2 \leq 1 + TK/\lambda$.

We can bound the effective dimension $\hat{d}$ as follows

$$\hat{d} = \frac{\log \det(\mathbf{I} + \mathbf{H}/\lambda)}{\log(1 + TK/\lambda)} = \frac{\log \det(\mathbf{I} + \Psi^T \Psi / \lambda)}{\log(1 + TK/\lambda)} \leq \frac{\log \|\mathbf{I} + \Psi^T \Psi / \lambda\|_2}{\log(1 + TK/\lambda)}.$$ 

where the first inequality is due to triangle inequality and the fact $\lambda \geq 1$, the second inequality holds due to the definition of $\Psi$ and triangle inequality, and the last inequality is by $\|\psi(x^i)\|_2 \leq 1$ for any $1 \leq i \leq TK$.

A.2 Verification of Remark 5.8

Let $K(\cdot, \cdot)$ be the NTK kernel, then for any $i, j \in [TK]$, we have $\mathbf{H}_{i,j} = K(x^i, x^j)$. We first prove that $\|h\|_{\mathcal{H}} = \|\beta^*\|_2$. Since $h \in \mathcal{H}$, we know that

$$h(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) = \sum_{i=1}^{n} \alpha_i \langle \psi(x_i), \psi(x) \rangle = \left\langle \sum_{i=1}^{n} \alpha_i \psi(x_i), \psi(x) \right\rangle,$$

for some $\{x_i\}_{i=1}^{n}, \{\alpha_i\}_{i=1}^{n}$. Thus, we have $\beta^* = \sum_{i=1}^{n} \alpha_i \psi(x_i)$. Therefore, we have

$$\|h\|_{\mathcal{H}}^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)$$

$$= \sum_{i,j=1}^{n} \alpha_i \alpha_j \langle \psi(x_i), \psi(x_j) \rangle$$

$$= \left\langle \sum_{i=1}^{n} \alpha_i \psi(x_i), \sum_{j=1}^{n} \alpha_j \psi(x_j) \right\rangle$$

$$= \langle \beta^*, \beta^* \rangle$$

$$= \|\beta^*\|_2^2,$$

which immediately suggests that $\|h\|_{\mathcal{H}} = \|\beta^*\|_2$. Next we prove that $\|h\|_{\mathcal{H}} \geq \sqrt{h^T \mathbf{H}^{-1} h}$. First, $h$ can be decomposed as $h = h_H + h_{\perp}$, where $h_H(x) = \sum_{i=1}^{TK} \tilde{\alpha}_i K(x, x^i)$ is the projection of $h$ to the function space spanned by $\{K(x, x^i)\}_{i=1}^{TK}$ and $h_{\perp}$ is the orthogonal part. By definition we have $h(x^i) = h_H(x^i)$ for $i \in [TK]$, thus

$$h = [h(x^1), \ldots, h(x^{TK})]^T$$

$$= [h_H(x^1), \ldots, h_H(x^{TK})]^T$$

$$= \left[ \sum_{i=1}^{TK} \tilde{\alpha}_i K(x^1, x^i), \ldots, \sum_{i=1}^{TK} \tilde{\alpha}_i K(x^{TK}, x^i) \right]^T$$

$$= \mathbf{H} \tilde{\alpha},$$

which implies that $\tilde{\alpha} = \mathbf{H}^{-1} h$. Thus, we have

$$\|h\|_{\mathcal{H}} \geq \|h_H\|_{\mathcal{H}} = \sqrt{\tilde{\alpha}^T \mathbf{H} \tilde{\alpha}} = \sqrt{h^T \mathbf{H}^{-1} \mathbf{H} \mathbf{H}^{-1} h} = \sqrt{h^T \mathbf{H}^{-1} h}.$$
A.3 Proof of Corollary 5.9

Proof of Corollary 5.9 Notice that $R_T \leq T$ since $0 \leq h(x) \leq 1$. Thus, with the fact that with probability at least $1 - \delta$, (5.2) holds, we can bound $\mathbb{E}[R_T]$ as

$$
\mathbb{E}[R_T] \leq (1 - \delta) \left( 3\sqrt{T} \sqrt{d \log (1 + TK/\lambda)} + 2\sqrt{d \log (1 + TK/\lambda)} + 2 - 2 \log \delta 
+ 2\sqrt{\lambda} \delta + C_2 (1 - \eta m \lambda)^{1/2} \sqrt{T/\lambda} \right) + \delta T.
$$

(A.1)

Taking $\delta = 1/T$ completes the proof.

B Proof of Lemmas in Section 6

B.1 Proof of Lemma 6.1

We start with the following lemma:

Lemma B.1. Let $G = [g(x^1; \theta_0), \ldots, g(x^{TK}; \theta_0)]/\sqrt{m} \in \mathbb{R}^{p \times (TK)}$. We denote the neural tangent kernel matrix $H$ the same as Definition 5.1. For any $\delta \in (0, 1)$, if

$$
m = \Omega \left( \frac{L^0 \log(TKL/\delta)}{\epsilon^4} \right),
$$

then with probability at least $1 - \delta$ over the random initialization of $\theta_0$, we have

$$
\|G^\top G - H\|_F \leq TK\epsilon.
$$

We begin to prove Lemma 6.1.

Proof of Lemma 6.1 By Assumption 5.2 we know that $\lambda_0 > 0$. Then with probability at least $1 - \delta$, we have

$$
G^\top G \succeq H - \|G^\top G - H\|_F I \succeq H - \lambda_0 I/2 \succeq H/2 > 0,
$$

(B.1)

where the second inequality holds due to Lemma 6.1 with the choice of $m$, the third and fourth inequality holds due to $H \succeq \lambda_0 I > 0$. Thus, suppose the singular value decomposition of $G$ is $G = PAQ^\top$, $P \in \mathbb{R}^{p \times TK}$, $A \in \mathbb{R}^{TK \times TK}$, $Q \in \mathbb{R}^{TK \times TK}$, we have $A > 0$ and $\theta^* = \theta_0 + PA^{-1}Q^\top/\sqrt{m}$ satisfies (6.1). To validate that $\theta^*$ satisfies (6.1), first we have

$$
G^\top \sqrt{m}(\theta^* - \theta_0) = QAP^\top PA^{-1}Q^\top/\sqrt{m} h = h,
$$

which suggests that for any $i$, $(g(x^i; \theta_0), \theta^* - \theta_0) = h(x^i)$. We also have

$$
m\|\theta^* - \theta_0\|_2^2 = h^\top QA^{-2}Q^\top h = h^\top (G^\top G)^{-1} h \leq 2h^\top H^{-1} h,
$$

where the last inequality holds due to (B.1). Thus, our proof finishes.

B.2 Proof of Lemma 6.2

In this section we prove Lemma 6.2. For the simplicity, we denote $\bar{Z}_t, \bar{b}_t, \bar{\gamma}_t$ as follows:

$$
\bar{Z}_t = \lambda I + \sum_{i=1}^{t} g(x_{i,a_i}; \theta_0)g(x_{i,a_i}; \theta_0)^\top/m,

\bar{b}_t = \sum_{i=1}^{t} r_{i,a_i}g(x_{i,a_i}; \theta_0)/\sqrt{m},

\bar{\gamma}_t = \nu \sqrt{\log \frac{\det \bar{Z}_t}{\det \lambda I} - 2 \log \delta + \sqrt{\lambda} S}.
$$

We need the following lemmas. The first lemma shows that the network parameter at each round $\theta_t$ can be well approximated by $\theta_0 + Z_t^{-1}b_t/\sqrt{m}$.
Lemma B.2. There exist constants \( \{\tilde{C}_i\}_{i=1}^5 > 0 \) such that for any \( \delta > 0 \), if \( \eta, m \) satisfy that for all \( t \in [T] \),

\[
2 \sqrt{t/(m \lambda)} \geq \tilde{C}_1 m^{-3/2} L^{-3/2} \log(TKL^2/\delta)^{3/2},
\]

\[
2 \sqrt{t/(m \lambda)} \leq \tilde{C}_2 \min \left\{ L^{-6} \log m, m(\lambda \eta)^2 L^{-6} t^{-1} (\log m)^{-1} \right\}^{3/8},
\]

\[
\eta \leq \tilde{C}_3 (m \lambda + t m L)^{-1},
\]

\[
m^{1/6} \geq \tilde{C}_4 \sqrt{\log m L^{7/2} t^{7/6} \lambda^{-7/6}(1 + \sqrt{T / \lambda}),}
\]

then with probability at least \( 1 - \delta \) over the random initialization of \( \theta_0 \), for any \( t \in [T] \), we have that \( \| \theta_t - \theta_0 \|_2 \leq 2 \sqrt{t/(m \lambda)} \) and

\[
\| \theta_t - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m} \|_2 \\
\leq (1 - \eta m \lambda)^t \sqrt{t/(m \lambda)} + \tilde{C}_5 m^{-2/3} \sqrt{\log m L^{7/2} t^{5/3} \lambda^{-5/6}(1 + \sqrt{T / \lambda}).}
\]

Next lemma shows the error bounds for \( \tilde{Z}_t \) and \( Z_t \).

Lemma B.3. There exist constants \( \{\tilde{C}_i\}_{i=1}^5 > 0 \) such that for any \( \delta > 0 \), if \( m \) satisfies that

\[
\tilde{C}_1 m^{-3/2} L^{-3/2} \log(TKL^2/\delta)^{3/2} \leq 2 \sqrt{t/(m \lambda)} \leq \tilde{C}_2 L^{-6} \log m)^{-3/2}, \forall t \in [T],
\]

then with probability at least \( 1 - \delta \) over the random initialization of \( \theta_0 \), we have the following inequalities for any \( t \in [T] \):

\[
\| Z_t \|_F \leq \tilde{C}_3 t L,
\]

\[
\| Z_t - \tilde{Z}_t \|_F \leq \tilde{C}_4 m^{-1/6} \sqrt{\log m L^{4} t^{7/6} \lambda^{-1/6}},
\]

\[
\left| \frac{\log \det(Z_t)}{\det(\lambda)} - \frac{\log \det(\tilde{Z}_t)}{\det(\lambda)} \right| \leq \tilde{C}_5 m^{-1/6} \sqrt{\log m L^{4} t^{5/3} \lambda^{-1/6}}.
\]

With above lemmas, we prove Lemma 6.2 as follows.

Proof of Lemma 6.2. By Lemma B.2 we know that \( \| \theta_t - \theta_0 \|_2 \leq 2 \sqrt{t/(m \lambda)} \). By Lemma 6.1 with probability at least \( 1 - \delta \), there exists \( \theta^* \) such that for any \( 1 \leq t \leq T \),

\[
h(x_{t,a}) = \langle g(x_{t,a}; \theta_0) / \sqrt{m}, \sqrt{m}(\theta^* - \theta_0) \rangle,
\]

\[
\sqrt{m} \| \theta^* - \theta_0 \|_2 \leq \sqrt{2h^T H^{-1} h} \leq S,
\]

where the second inequality holds due to the choice of \( S \). Thus, by Theorem 2 in [Abbasi-Yadkori et al. 2011], with probability at least \( 1 - 2\delta \), for any \( 1 \leq t \leq T \), \( \theta^* \) satisfies that

\[
\| \sqrt{m}(\theta^* - \theta_0) - \tilde{Z}_t^{-1} \tilde{b}_t \|_{Z_t} \leq \tilde{\gamma}_t.
\]

(B.2)

We now prove that \( \| \theta^* - \theta_t \|_{Z_t} \leq \gamma_t / \sqrt{m} \). From the triangle inequality,

\[
\| \theta^* - \theta_t \|_{Z_t} \leq \underbrace{\| \theta^* - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m} \|_{Z_t}}_{I_1} + \underbrace{\| \theta_t - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m} \|_{Z_t}}_{I_2}.
\]

(B.3)

We bound \( I_1 \) and \( I_2 \) separately. For \( I_1 \), we have

\[
I_1^2 = (\theta^* - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m})^T Z_t (\theta^* - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m})
\]

\[
= (\theta^* - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m})^T Z_t (\theta^* - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m})
\]

\[
+ (\theta^* - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m})^T (Z_t - \tilde{Z}_t) (\theta^* - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m})
\]

\[
\leq (\theta^* - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m})^T Z_t (\theta^* - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m})
\]

\[
+ \frac{\| \tilde{Z}_t - Z_t \|_2}{\lambda} (\theta^* - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m})^T Z_t (\theta^* - \theta_0 - \tilde{Z}_t^{-1} \tilde{b}_t / \sqrt{m})
\]

\[
\leq (1 + \| Z_t - \tilde{Z}_t \|_2 / \lambda) \tilde{\gamma}_t^2 / m,
\]

(B.4)
where the first inequality holds due to the fact that $x^\top Ax \leq x^\top Bx: \|A\|_2/\lambda_{\text{min}}(B)$ for some $B > 0$ and the fact that $\lambda_{\text{min}}(Z_i) \geq \lambda$, the second inequality holds due to (B.2). By Lemma B.3, we have
\[
\|Z_t - Z_t\|_2 \leq \|Z_t - Z_t\|_F \leq C_1 m^{-1/6} \sqrt{\log m} L^{7/6} \lambda^{-1/6},
\] (B.5)
and
\[
\tilde{\gamma}_t = \nu \sqrt{\log \det \frac{Z_t}{\det \lambda I}} - 2 \log \delta + \sqrt{\lambda S} \\
\leq \nu \sqrt{\log \det \frac{Z_t}{\det \lambda I}} + C_2 m^{-1/6} \sqrt{\log m} L^{45/3} \lambda^{-1/6} - 2 \log \delta + \sqrt{\lambda S},
\] (B.6)
where $C_1, C_2 > 0$ are two constants. Substituting (B.5) and (B.6) into (B.4), we have
\[
I_1 \leq \sqrt{1 + \|Z_t - Z_t\|_2/\lambda \tilde{\gamma}_t/\sqrt{m}} \\
\leq \sqrt{1 + C_1 m^{-1/6} \sqrt{\log m} L^{7/6} \lambda^{-7/6}/\sqrt{m}} \\
\cdot \left( \nu \sqrt{\log \det \frac{Z_t}{\det \lambda I}} + C_2 m^{-1/6} \sqrt{\log m} L^{45/3} \lambda^{-1/6} - 2 \log \delta + \sqrt{\lambda S} \right).
\] (B.7)
For $I_2$, we have
\[
I_2 \leq C_3 L \|\theta_t - \theta_0 - \bar{Z}_t^{-1} \bar{b}_t/\sqrt{m}\|_2 \\
\leq C_4 \left[ (1 - \eta m \lambda)^J \sqrt{t/(m \lambda)} + m^{-2/3} \sqrt{\log m} L^{7/2} L^{5/3} \lambda^{-5/3} (1 + \sqrt{t/\lambda}) \right],
\] (B.8)
where $C_3, C_4 > 0$ are two constants, the first inequality holds due to $\|Z_t\|_2 \leq \|Z_t\|_F \leq C_3 m L$ by Lemma B.3, the second inequality holds due to Lemma B.2. Substituting (B.7) and (B.8) into (B.3), we have $\|\theta^* - \theta_t\|_{Z_t} \leq \gamma_t/\sqrt{m}$. Our proof is thus completed.

B.3 Proof of Lemma 6.3

The proof starts with three lemmas that bound the error terms of the function value and gradient of neural networks.

**Lemma B.4** (Lemma 4.1, Cao and Gu (2019a)). There exist constants $\{C_i\}_{i=1}^3 > 0$ such that for any $\delta > 0$, if $\tau$ satisfies that
\[
\bar{C}_1 m^{-3/2} L^{-3/2} \log(TKL^2/\delta)^{3/2} \leq \tau \leq \bar{C}_2 L^{-6} \log m \right)^{3/2},
\] then with probability at least $1 - \delta$ over the random initialization of $\theta_0$, for all $\tilde{\theta}, \hat{\theta}$ satisfying $\|\tilde{\theta} - \theta_0\|_2 \leq \tau, \|\hat{\theta} - \theta_0\|_2 \leq \tau$ and $j \in [TK]$ we have
\[
\left| f(x^j; \tilde{\theta}) - f(x^j; \hat{\theta}) - \frac{g(x^j; \tilde{\theta}) - g(x^j; \hat{\theta})}{\sqrt{m} \log m} \right| \leq \bar{C}_3 \tau^{4/3} L^3 \sqrt{m} \log m.
\]

**Lemma B.5** (Theorem 5, Allen-Zhu et al. (2019)). There exist constants $\{C_i\}_{i=1}^3 > 0$ such that for any $\delta \in (0, 1)$, if $\tau$ satisfies that
\[
\bar{C}_1 m^{-3/2} L^{-3/2} \log(TKL^2/\delta)^{3/2} \leq \tau \leq \bar{C}_2 L^{-6} \log m \right)^{3/2},
\] then with probability at least $1 - \delta$ over the random initialization of $\theta_0$, for all $\|\theta - \theta_0\|_2 \leq \tau$ and $j \in [TK]$ we have
\[
\|g(x^j; \theta) - g(x^j; \theta_0)\|_2 \leq \bar{C}_3 \sqrt{\log m} \tau^{-1/3} L^3 \|g(x^j; \theta_0)\|_2.
\]

**Lemma B.6** (Lemma B.3, Cao and Gu (2019a)). There exist constants $\{C_i\}_{i=1}^3 > 0$ such that for any $\delta > 0$, if $\tau$ satisfies that
\[
\bar{C}_1 m^{-3/2} L^{-3/2} \log(TKL^2/\delta)^{3/2} \leq \tau \leq \bar{C}_2 L^{-6} \log m \right)^{3/2},
\] then with probability at least $1 - \delta$ over the random initialization of $\theta_0$, for any $\|\theta - \theta_0\|_2 \leq \tau$ and $j \in [TK]$ we have $\|g(x^j; \theta)\|_F \leq \bar{C}_3 \sqrt{m} L.$
Proof of Lemma 6.2: We follow the regret bound analysis in [Abbasi-Yadkori et al., 2011; Valko et al., 2013]. Denote \(a_t^* = \arg\max_{a \in [K]} h(x_t, a)\) and \(C_t = \{\hat{\theta} : \|\hat{\theta} - \theta_t\|_2 \leq \gamma_t/\sqrt{m}\}\). By Lemma 6.2 for all \(1 \leq t \leq T\), we have \(\|\theta_t - \theta_0\|_2 \leq 2\sqrt{t/(m\lambda)}\) and \(\theta^* \in C_t\). By the choice of \(m\), Lemma B.4, B.5 and B.6 hold. Thus, \(h(x_t, a_t^*) - h(x_t, a_t)\) can be bounded as follows:

\[
\begin{align*}
&h(x_t, a_t^*) - h(x_t, a_t) \\
&= \langle g(x_t, a_t^*; \theta_0), \theta^* - \theta_0 \rangle - \langle g(x_t, a_t; \theta_0), \theta^* - \theta_0 \rangle \\
&\leq \langle g(x_t, a_t^*; \theta_{t-1}), \theta^* - \theta_0 \rangle - \langle g(x_t, a_t; \theta_{t-1}), \theta^* - \theta_0 \rangle \\
&\quad + \|\theta^* - \theta_0\|_2 \langle g(x_t, a_t; \theta_{t-1}) - g(x_t, a_t^*; \theta_0), \theta^* - \theta_0 \rangle \\
&\leq \langle g(x_t, a_t^*; \theta_{t-1}), \theta^* - \theta_0 \rangle - \langle g(x_t, a_t; \theta_{t-1}), \theta^* - \theta_0 \rangle \\
&\quad + C_1 \sqrt{\ln |H^{-1}Hm|^{-1/6}} \sqrt{\log mt^{1/6} \lambda^{-1/6} L^{7/2}} \\
&\leq \max_{\theta \in C_{t-1}} \langle g(x_t, a_t^*; \theta_{t-1}), \theta - \theta_0 \rangle - \langle g(x_t, a_t; \theta_{t-1}), \theta^* - \theta_0 \rangle \\
&\quad + C_1 \sqrt{\ln |H^{-1}Hm|^{-1/6}} \sqrt{\log mt^{1/6} \lambda^{-1/6} L^{7/2}},
\end{align*}
\]

where the first inequality holds due to triangle inequality, the second inequality holds due to Lemma 6.1, Lemma B.5 and B.6, the third inequality holds due to \(\theta^* \in C_{t-1}\).

Denote

\[
\tilde{U}_{t,a} = \langle g(x_t, a^*; \theta_{t-1}), \theta_{t-1} - \theta_0 \rangle + \gamma_{t-1} \sqrt{\ln |H^{-1}Hm|^{-1/6}} \sqrt{\log mt^{1/6} \lambda^{-1/6} L^{7/2}},
\]

then we have \(\tilde{U}_{t,a} = \max_{\theta \in C_{t-1}} \langle g(x_t, a^*; \theta_{t-1}), \theta - \theta_0 \rangle\) with the fact that

\[
(a, b) + c\sqrt{a^T A^{-1} a} = \max_{\|x-b\|_A \leq c} \langle a, x \rangle.
\]

We also have

\[
|U_{t,a} - \tilde{U}_{t,a}| \leq C_2 m^{-1/6} \sqrt{\log mt^{2/3} \lambda^{-2/3} L^3},
\]

where \(C_2 > 0\) is a constant, the inequality holds due to Lemma B.4, with the fact \(\|\theta_{t-1} - \theta_0\|_2 \leq 2\sqrt{t/(m\lambda)}\) and the fact \(f(x^*; \theta_0) = 0\). Since \(\theta^* \in C_{t-1}\), then (B.9) can be bounded as

\[
\max_{\theta \in C_{t-1}} \langle g(x_t, a^*; \theta_{t-1}), \theta - \theta_0 \rangle - \langle g(x_t, a^*; \theta_{t-1}), \theta^* - \theta_0 \rangle \\
= \tilde{U}_{t,a^*} - \langle g(x_t, a^*; \theta_{t-1}), \theta^* - \theta_0 \rangle \\
\leq U_{t,a^*} - \langle g(x_t, a^*; \theta_{t-1}), \theta^* - \theta_0 \rangle + C_2 m^{-1/6} \sqrt{\log mt^{2/3} \lambda^{-2/3} L^3} \\
\leq U_{t,a^*} - \langle g(x_t, a^*; \theta_{t-1}), \theta^* - \theta_0 \rangle + C_2 m^{-1/6} \sqrt{\log mt^{2/3} \lambda^{-2/3} L^3} \\
\leq \tilde{U}_{t,a^*} - \langle g(x_t, a^*; \theta_{t-1}), \theta^* - \theta_0 \rangle + 2C_2 m^{-1/6} \sqrt{\log mt^{2/3} \lambda^{-2/3} L^3},
\]

where the first inequality holds due to (B.10), the second inequality holds since \(a_t = \arg\max_{a} U_{t,a}\), the third inequality holds due to (B.10). Furthermore,

\[
\begin{align*}
\tilde{U}_{t,a_t} - \langle g(x_t, a_t; \theta_0), \theta^* - \theta_0 \rangle \\
&= \max_{\theta \in C_{t-1}} \langle g(x_t, a_t; \theta_{t-1}), \theta - \theta_0 \rangle - \langle g(x_t, a_t; \theta_{t-1}), \theta^* - \theta_0 \rangle \\
&= \max_{\theta \in C_{t-1}} \langle g(x_t, a_t; \theta_{t-1}), \theta - \theta_{t-1} \rangle - \langle g(x_t, a_t; \theta_{t-1}), \theta^* - \theta_{t-1} \rangle \\
&\leq \max_{\theta \in C_{t-1}} \|\theta - \theta_{t-1}\|_{Z_{t-1}} \|g(x_t, a_t; \theta_{t-1})\|_{Z_{t-1}^{-1}} + \|\theta^* - \theta_{t-1}\|_{Z_{t-1}} \|g(x_t, a_t; \theta_{t-1})\|_{Z_{t-1}^{-1}} \\
&\leq 2\gamma_{t-1} \|g(x_t, a_t; \theta_{t-1})\|_{\sqrt{m}} \|Z^{-1}\|_{Z^{-1} Z_{t-1}},
\end{align*}
\]
where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to Lemma 6.4. Combining (B.3), (B.11) and (B.12), we have

\[ h(x_{t,a_t}) - h(x_{t,a_t}) \]

\[ \leq 2\gamma_t-1||g(x_{t,a_t};\theta_{t-1})/\sqrt{m}||_{z_{t-1}} + C_1\sqrt{h^T H^{-1} h} m^{-1/6} \sqrt{\log mt^{1/6} \lambda^{-1/6} L^7} / 3 \\
+ 2C_2 m^{-1/6} \sqrt{\log mt^{2/3} \lambda^{-2/3} L^3} \]

\[ \leq \min \left\{ 2\gamma_t-1||g(x_{t,a_t};\theta_{t-1})/\sqrt{m}||_{z_{t-1}}, 1 \right\} + C_1\sqrt{h^T H^{-1} h} m^{-1/6} \sqrt{\log mt^{1/6} \lambda^{-1/6} L^7} / 3 \\
+ 2C_2 m^{-1/6} \sqrt{\log mt^{2/3} \lambda^{-2/3} L^3} \]

\[ \leq 2\gamma_t-1 \min \left\{ ||g(x_{t,a_t};\theta_{t-1})/\sqrt{m}||_{z_{t-1}}, 1 \right\} + C_1\sqrt{h^T H^{-1} h} m^{-1/6} \sqrt{\log mt^{1/6} \lambda^{-1/6} L^7} / 3 \\
+ 2C_2 m^{-1/6} \sqrt{\log mt^{2/3} \lambda^{-2/3} L^3}, \quad \text{B.13} \]

where the second inequality holds due to the fact that \( h(x_{t,a_t}) - h(x_{t,a_t}) \leq 1 \), the third inequality holds due to the fact that \( \min\{a, b, 1\} \leq \min\{a, 1\} + b \), the fourth inequality holds due to the fact \( \gamma_t-1 \geq \sqrt{\lambda S} \geq 1 \). Finally, by the fact that \( \sqrt{2hH^{-1}h} \leq S \), we finish the proof. \( \square \)

**B.4 Proof of Lemma 6.4**

In this section we prove Lemma 6.4, we need the following lemma from [Abbasi-Yadkori et al. (2011)].

**Lemma B.7** (Lemma 11, [Abbasi-Yadkori et al. (2011)].) We have the following inequality:

\[ \sum_{t=1}^{T} \min \left\{ ||g(x_{t,a_t};\theta_{t-1})/\sqrt{m}||_{z_{t-1}}, 1 \right\} \leq 2 \log \frac{\det Z_T}{\det \lambda} \]

**Proof of Lemma 6.4** First by the definition of \( \gamma_t \), we know that \( \gamma_t \) is a monotonic function w.r.t. \( \det Z_t \). By the definition of \( Z_t \), we know that \( Z_T \geq Z_t \), which implies that \( \det Z_t \leq \det Z_T \). Thus, \( \gamma_t \leq \gamma_T \). Second, by Lemma B.7 we know that

\[ \sum_{t=1}^{T} \min \left\{ ||g(x_{t,a_t};\theta_{t-1})/\sqrt{m}||_{z_{t-1}}, 1 \right\} \leq 2 \log \frac{\det Z_T}{\det \lambda} \]

\[ \leq 2 \log \frac{\det Z_T}{\det \lambda} + C_1 m^{-1/6} \sqrt{\log m L^4 T^{5/3} \lambda^{-1/6}}, \quad \text{B.14} \]

where the second inequality holds due to Lemma B.3. Next we are going to bound \( \log \det Z_T \).

Denote \( G = [g(x_1;\theta_0)/\sqrt{m}, \ldots, g(x^{TK};\theta_0)/\sqrt{m}] \in \mathbb{R}^{p \times (TK)} \), then we have

\[ \log \frac{\det Z_T}{\det \lambda} = \log \det \left( I + \sum_{t=1}^{T} \frac{g(x_{t,a};\theta_0)g(x_{t,a};\theta_0)^T}{(m\lambda)} \right) \]

\[ \leq \log \det \left( I + \sum_{i=1}^{TK} \frac{g(x_i;\theta_0)g(x_i;\theta_0)^T}{(m\lambda)} \right) \]

\[ \leq \log \det \left( I + GG^T / \lambda \right) \]

\[ = \log \det \left( I + G^T G / \lambda \right), \quad \text{B.15} \]
where the inequality holds naively, the third equality holds since for any matrix $A \in \mathbb{R}^{p \times TK}$, we have $\det(I + AA^\top) = \det(I + A^\top A)$. We can be further bound (B.15) by the follows:

$$\log \det \left( I + \frac{G^\top G}{\lambda} \right) = \log \det \left( I + \frac{H}{\lambda} + (G^\top G - H)/\lambda \right)$$

$$\leq \log \det \left( I + \frac{H}{\lambda} \right) + \langle (I + \frac{H}{\lambda})^{-1}, (G^\top G - H)/\lambda \rangle$$

$$\leq \log \det \left( I + \frac{H}{\lambda} \right) + ||(I + \frac{H}{\lambda})^{-1}||_F ||G^\top G - H||_F/\lambda$$

$$\leq \log \det \left( I + \frac{H}{\lambda} \right) + \sqrt{TK} ||G^\top G - H||_F$$

$$\leq \log \det \left( I + \frac{H}{\lambda} \right) + 1$$

$$\leq \tilde{d} \log (1 + TK/\lambda) + 1,$$

(B.16)

where the first inequality holds due to the concavity of $\log \det(\cdot)$, the second inequality holds due to the fact that $\langle A, B \rangle \leq ||A||_F ||B||_F$, the third inequality holds due to the facts that $I + H/\lambda \succeq I$, $\lambda \geq 1$ and $||A||_F \leq \sqrt{TK} ||A||_2$ for any $A \in \mathbb{R}^{TK \times TK}$, the fourth inequality holds by Lemma B.1 with the choice of $m$, the fifth inequality holds by the definition of effective dimension in Definition 5.3 and the last inequality holds due to the choice of $\lambda$. We now bound $\gamma_T$, which is

$$\gamma_T = \sqrt{1 + C_1 m^{-1/6} \sqrt{\log mL^4 T^{7/6} \lambda^{-7/6}}}$$

$$\cdot \left[ \nu \log \frac{\det Z_T}{\det \lambda I} + C_2 m^{-1/6} \sqrt{\log mL^4 T^{5/3} \lambda^{-1/6} - 2 \log \delta + \sqrt{\lambda S}} \right]$$

$$+ C_3 \left[ (1 - \eta m \lambda)^T \sqrt{T/(m \lambda)} + m^{-2/3} \sqrt{\log mL^7/2 T^{5/3} \lambda^{-5/3} (1 + \sqrt{T/\lambda})} \right]$$

$$\leq \sqrt{1 + C_1 m^{-1/6} \sqrt{\log mL^4 T^{7/6} \lambda^{-7/6}}}$$

$$\cdot \left[ \nu \log \frac{\det Z_T}{\det \lambda I} + C_2 m^{-1/6} \sqrt{\log mL^4 T^{5/3} \lambda^{-1/6} - 2 \log \delta + \sqrt{\lambda S}} \right]$$

$$+ C_3 \left[ (1 - \eta m \lambda)^T \sqrt{T/(m \lambda)} + m^{-2/3} \sqrt{\log mL^7/2 T^{5/3} \lambda^{-5/3} (1 + \sqrt{T/\lambda})} \right], \quad (B.17)$$
where the inequality holds due to Lemma B.3. Finally, substituting (B.15), (B.16) into (B.14), using (B.14) and (B.17), we have

\[
\sqrt{T} \sum_{t=1}^{T} \gamma_t^2 \min \left\{ \|g(x_t, \theta_{t-1})/\sqrt{m}\|_{Z_{t-1}}^2, 1 \right\} 
\leq \gamma_T \sqrt{T} \sum_{t=1}^{T} \min \left\{ \|g(x_t, \theta_{t-1})/\sqrt{m}\|_{Z_{t-1}}^2, 1 \right\} 
\leq \sqrt{\log \frac{\det Z_T}{\det \lambda I}} + C_1 m^{-1/6} \sqrt{\log m L^4 T^{5/3} \lambda^{-1/6}} \left[ \sqrt{1 + C_1 m^{-1/6} \sqrt{\log m L^4 T^{7/6} \lambda^{-7/6}}} \right. 
\quad \left. \left( \nu \sqrt{\log \frac{\det Z_T}{\det \lambda I}} + C_2 m^{-1/6} \sqrt{\log m L^4 T^{5/3} \lambda^{-1/6}} - 2 \log \delta + \sqrt{\lambda S} \right) 
\quad + C_3 \left( 1 - \eta_m \lambda \right)^t \sqrt{T/(m \lambda)} + m^{-3/2} \sqrt{\log m L^7/2 T^{5/3} \lambda^{-5/3}} (1 + \sqrt{T/\lambda}) \right] 
\leq \sqrt{d \log (1 + TK/\lambda)} + 1 + C_1 m^{-1/6} \sqrt{\log m L^4 T^{5/3} \lambda^{-1/6}} \left[ \sqrt{1 + C_1 m^{-1/6} \sqrt{\log m L^4 T^{7/6} \lambda^{-7/6}}} \right. 
\quad \left. \left( \nu \sqrt{d \log (1 + TK/\lambda)} + 1 + C_2 m^{-1/6} \sqrt{\log m L^4 T^{5/3} \lambda^{-1/6}} - 2 \log \delta + \sqrt{\lambda S} \right) 
\quad + C_3 \left( 1 - \eta_m \lambda \right)^t \sqrt{T/(m \lambda)} + m^{-3/2} \sqrt{\log m L^7/2 T^{5/3} \lambda^{-5/3}} (1 + \sqrt{T/\lambda}) \right].
\]

\[\Box\]

C Proofs of Technical Lemmas in Appendix B

C.1 Proof of Lemma B.1

In this section we prove Lemma B.1, we need the following lemma from Arora et al. (2019):

Lemma C.1 (Theorem 3.1, Arora et al. (2019)). Fix \( \epsilon > 0 \) and \( \delta \in (0, 1) \). Suppose that

\[ m = \Omega \left( \frac{L^6 \log(L) \delta}{\epsilon^4} \right), \]

then for any \( i, j \in [TK] \), with probability at least \( 1 - \delta \) over random initialization of \( \theta_0 \), we have

\[ |\langle g(x^i; \theta_0), g(x^j; \theta_0) \rangle/m - H_{i,j}| \leq \epsilon. \]  

(C.1)

Proof of Lemma B.1 Taking union bound over \( i, j \in [TK] \), we have that if

\[ m = \Omega \left( \frac{L^6 \log(T^2 K^2 L) \delta}{\epsilon^4} \right), \]

then with probability at least \( 1 - \delta, \) (C.1) holds for all \( (i, j) \in [TK] \times [TK] \). Therefore, we have

\[ ||G^T G - H||_F = \sqrt{\sum_{i=1}^{TK} \sum_{j=1}^{TK} |\langle g(x^i; \theta_0), g(x^j; \theta_0) \rangle/m - H_{i,j}|^2} \leq TK \epsilon. \]

\[\Box\]
C.2 PROOF OF LEMMA B.2

In this section we prove Lemma B.2. During the proof, for the simplicity, we omit the subscript \( t \) by default. We define the following quantities:

\[
\begin{align*}
J^{(j)} &= \left( g(x_{1,a_1}; \theta^{(j)}), \ldots, g(x_{t,a_t}; \theta^{(j)}) \right) \in \mathbb{R}^{(md + m^2(L-2) + m) \times t}, \\
H^{(j)} &= [J^{(j)}]^T J^{(j)} \in \mathbb{R}^{t \times t}, \\
f^{(j)} &= (f(x_{1,a_1}; \theta^{(j)}), \ldots, f(x_{t,a_t}; \theta^{(j)}) \in \mathbb{R}^{t \times 1}, \\
v &= (r_{1,a_1}, \ldots, r_{t,a_t}) \in \mathbb{R}^{t \times 1}.
\end{align*}
\]

Then the update rule of \( \theta^{(j)} \) can be written as follows:

\[
\theta^{(j+1)} = \theta^{(j)} - \eta [J^{(j)}] (f^{(j)} - v) + m\lambda(\theta^{(j)} - \theta^{(0)})].
\] (C.2)

We also define the following auxiliary sequence \( \{\tilde{\theta}^{(k)}\} \) during the proof:

\[
\tilde{\theta}^{(0)} = \theta^{(0)}, \quad \tilde{\theta}^{(j+1)} = \theta^{(j)} - \eta [J^{(0)}] (J^{(j)} \tilde{\theta}^{(0)} - v) + m\lambda(\tilde{\theta}^{(j)} - \tilde{\theta}^{(0)})] \quad \text{for any } j \in [J].
\]

Next lemma provides perturbation bounds for \( J^{(j)}, H^{(j)} \) and \( \|f^{(j+1)} - f^{(j)} - [J^{(j)}]^T (\theta^{(j+1)} - \theta^{(j)})\|_2 \).

**Lemma C.2.** There exist constants \( \{\bar{C}_i\}_{i=1}^6 > 0 \) such that for any \( \delta > 0 \), if \( \tau \) satisfies that

\[
\bar{C}_1 m^{-3/2} L^{-3/2} \log(T KL^2 / \delta)^{3/2} \leq \tau \leq \bar{C}_2 L^{-6} [\log m]^{-3/2},
\]

then with probability at least \( 1 - \delta \) over the random initialization of \( \theta^{(0)} \), if for any \( j \in [J] \), \( \|\theta^{(j)} - \theta^{(0)}\|_2 \leq \tau \), we have the following inequalities for any \( j \in [J] \):

\[
\begin{align*}
\|J^{(j)}\|_F &\leq \bar{C}_4 \sqrt{tmL}, \quad \text{(C.3)} \\
\|J^{(j)} - J^{(0)}\|_F &\leq \bar{C}_5 \sqrt{tm \log m}\tau^{1/3} L^{7/2}, \quad \text{(C.4)} \\
\|f^{(j+1)} - f^{(j)} - [J^{(j)}]^T (\theta^{(j+1)} - \theta^{(j)})\|_2 &\leq \bar{C}_6 \tau^{4/3} L^{3} \sqrt{tm \log m}, \quad \text{(C.5)} \\
\|v\|_2 &\leq \sqrt{\tau}. \quad \text{(C.6)}
\end{align*}
\]

Next lemma gives an upper bound for \( \|f^{(j)} - v\|_2 \).

**Lemma C.3.** There exist constants \( \{\bar{C}_i\}_{i=1}^4 > 0 \) such that for any \( \delta > 0 \), if \( \tau, \eta \) satisfy that

\[
\begin{align*}
\bar{C}_1 m^{-3/2} L^{-3/2} \log(T KL^2 / \delta)^{3/2} &\leq \tau \leq \bar{C}_2 L^{-6} [\log m]^{-3/2}, \\
\eta &\leq \bar{C}_3 (m\lambda + tmL)^{-1}, \\
\tau^{8/3} &\leq \bar{C}_4 m(\lambda\eta)^2 L^{-6} t^{-1} (\log m)^{-1},
\end{align*}
\]

then with probability at least \( 1 - \delta \) over the random initialization of \( \theta^{(0)} \), if for any \( j \in [J] \), \( \|\theta^{(j)} - \theta^{(0)}\|_2 \leq \tau \), we have that for any \( j \in [J] \), \( \|f^{(j)} - v\|_2 \leq 2\sqrt{\tau} \).

Next lemma gives an upper bound of the distance between auxiliary sequence \( \|\tilde{\theta}^{(j)} - \tilde{\theta}^{(0)}\|_2 \).

**Lemma C.4.** There exist constants \( \{\bar{C}_i\}_{i=1}^1 > 0 \) such that for any \( \delta \in (0, 1) \), if \( \tau, \eta \) satisfy that

\[
\begin{align*}
\bar{C}_1 m^{-3/2} L^{-3/2} \log(T KL^2 / \delta)^{3/2} &\leq \tau \leq \bar{C}_2 L^{-6} [\log m]^{-3/2}, \\
\eta &\leq \bar{C}_3 (tmL + m\lambda)^{-1},
\end{align*}
\]

then with probability at least \( 1 - \delta \), we have that for any \( j \in [J] \),

\[
\begin{align*}
\|\tilde{\theta}^{(j)} - \theta^{(0)}\|_2 &\leq \sqrt{t/(m\lambda)}, \\
\|\tilde{\theta}^{(j)} - \theta^{(0)} - (Z)^{-1} \mathbf{b} / \sqrt{m}\|_2 &\leq (1 - \eta m\lambda)^J \sqrt{t/(m\lambda)}
\end{align*}
\]

With above lemmas, we prove Lemma B.2 as follows.
Proof of Lemma B.2  
Set $\tau = 2\sqrt{t/(m\lambda)}$. First we assume that $\|\theta^{(j)} - \theta^{(0)}\|_2 \leq \tau$ for all $0 \leq j \leq J$. Then with this assumption and the choice of $m, \tau$, we have that Lemma C.2, C.3 and C.4 hold. Then we have
\[
\|\theta^{(j+1)} - \tilde{\theta}^{(j+1)}\|_2 = \|\theta^{(j)} - \tilde{\theta}^{(j)} - \eta (J^{(j)} - J^{(0)}) (f^{(j)} - v) - \eta m \lambda (\theta^{(j)} - \tilde{\theta}^{(j)}) \|_2
\]
\[
= \|\eta (J^{(0)} (f^{(j)} - [J^{(0)}]^{\top} (\tilde{\theta}^{(j)} - \theta^{(0)})) - \eta m \lambda (\theta^{(j)} - \tilde{\theta}^{(j)}) \|_2
\]
\[
= \bigg\| (1 - \eta m \lambda) (\theta^{(j)} - \tilde{\theta}^{(j)}) - \eta (J^{(j)} - J^{(0)}) (f^{(j)} - v) \bigg\|_2
\]
\[
\leq \| \bigg[ I - \eta (m \lambda I + H^{(0)}) \bigg] (\theta^{(j)} - \tilde{\theta}^{(j)}) \bigg\|_2 + \eta \| (J^{(j)} - J^{(0)}) (f^{(j)} - v) \bigg\|_2
\]
\[
\leq (1 - \eta m \lambda) \| \tilde{\theta}^{(j)} - \theta^{(j)} \|_2 + \eta \| (J^{(j)} - J^{(0)}) (f^{(j)} - v) \|_2
\]
where the first inequality holds due to triangle inequality, the second inequality holds due to the fact that $\eta H^{(0)} = \eta [J^{(0)}]^{\top} J^{(0)} \leq I$ with (C.3) in Lemma C.2 for some $C_1 > 0$. We now bound $I_1, I_2$ separately. For $I_1$, we have
\[
I_1 \leq \eta \| J^{(j)} - J^{(0)} \|_2 \| f^{(j)} - v \|_2 \leq \eta C_2 \sqrt{m \log m^{1/3} L^7/2},
\]
where $C_2 > 0$ is a constant, the first inequality holds due to matrix spectral norm and the second inequality holds due to (C.4) in Lemma C.2 and Lemma C.3. For $I_2$, we have
\[
I_2 \leq \eta \| J^{(0)} \|_2 \| f^{(j)} - [J^{(0)}]^{\top} (\theta^{(j)} - \theta^{(0)}) \|_2 \leq \eta C_3 m L^{7/2} \tau^{1/3} \sqrt{\log m},
\]
where $C_3 > 0$, the first inequality holds due to matrix spectral norm, the second inequality holds due to (C.3) and (C.5) in Lemma C.2 and the fact that $f^{(0)} = 0$ by random initialization over $\theta^{(0)}$. Substituting (C.8) and (C.9) into (C.7), we have
\[
\| \theta^{(j+1)} - \tilde{\theta}^{(j+1)} \|_2 \leq (1 - \eta m \lambda) \| \theta^{(j)} - \tilde{\theta}^{(j)} \|_2 + C_4 \eta \| m \log m^{1/3} L^7/2 + \eta m L^{7/2} \tau^{4/3} \sqrt{\log m},
\]
where $C_4 > 0$ is a constant. Expanding (C.10) for $k$ times, we have
\[
\| \theta^{(j+1)} - \tilde{\theta}^{(j+1)} \|_2 \leq C_4 \eta \| m \log m^{1/3} L^7/2 + \eta m L^{7/2} \tau^{4/3} \sqrt{\log m}
\]
\[
= C_5 m^{-2/3} \sqrt{\log m^{1/2} 2^{1/3} \lambda^{-5/3}} (1 + \sqrt{t/\lambda})
\]
\[
\leq \tau/2,
\]
where $C_5 > 0$ is a constant, the equality holds by the definition of $\tau$, the last inequality holds due to the choice of $m$, where
\[
m^{1/6} \geq C_6 \sqrt{\log m L^{7/2} t^{7/6} \lambda^{-7/6}} (1 + \sqrt{t/\lambda}),
\]
and $C_6 > 0$ is a constant. Thus, for any $j \in [J]$, we have
\[
\| \theta^{(j)} - \theta^{(0)} \|_2 \leq \| \tilde{\theta}^{(j)} - \theta^{(0)} \|_2 + \| \theta^{(j)} - \tilde{\theta}^{(j)} \|_2 \leq \sqrt{t/(m \lambda)} + \tau/2 = \tau,
\]
where the first inequality holds due to triangle inequality, the second inequality holds due to Lemma C.4, (C.13) suggests that our assumption $\| \theta^{(j)} - \theta^{(0)} \|_2 \leq \tau$ holds for any $u$. Note that we have the following inequality by Lemma C.4
\[
\| \tilde{\theta}^{(j)} - \theta^{(0)} - (Z)^{-1} b / \sqrt{m} \|_2 \leq (1 - \eta m \lambda)^j \sqrt{t/(m \lambda)}.
\]
Using (C.11) and (C.13), we have
\[
\| \theta^{(j)} - \theta^{(0)} - (Z)^{-1} b / \sqrt{m} \|_2 \leq (1 - \eta m \lambda)^j \sqrt{t/(m \lambda)} + C_5 m^{-2/3} \sqrt{\log m^{1/2} 2^{1/3} \lambda^{-5/3}} (1 + \sqrt{t/\lambda}).
\]
This completes the proof. \qed
C.3 PROOF OF LEMMA B.3

In this section we prove Lemma B.3.

Proof of Lemma B.3

Set $\tau = 2\sqrt{t/(m\lambda)}$. By Lemma 6.2, we have that $\|\theta_t - \theta_0\|_2 \leq \tau$ for $i \in [t]$. \|Z_t\|_F can be bounded as follows.

\[
\|Z_t\|_F = \left\| \sum_{i=1}^{t} g(x_{i,a_i}; \theta_0)g(x_{i,a_i}; \theta_0)^\top / m \right\|_F \leq \sum_{i=1}^{t} \| g(x_{i,a_i}; \theta_0) \|_2^2 / m \leq C_0 t L,
\]

where $C_0 > 0$ is a constant, the first inequality holds due to the fact that $\|aa^\top\|_F = \|a\|_2^2$, the second inequality holds due to Lemma B.6 with the fact that $\|\theta_t - \theta_0\|_2 \leq \tau$. We bound $\|Z_t - Z_t\|_2$ as follows. We have

\[
\|Z_t - Z_t\|_F = \left\| \sum_{i=1}^{t} \left( g(x_{i,a_i}; \theta_0)g(x_{i,a_i}; \theta_0)^\top - g(x_{i,a_i}; \theta_t)g(x_{i,a_i}; \theta_t)^\top \right) / m \right\|_F \leq \sum_{i=1}^{t} \left( \| g(x_{i,a_i}; \theta_0) \|_2 + \| g(x_{i,a_i}; \theta_t) \|_2 \right) \left\| g(x_{i,a_i}; \theta_0) - g(x_{i,a_i}; \theta_t) \right\|_2 / m,
\]

where the first inequality holds due to triangle inequality, the second inequality holds the fact that $\|aa^\top - bb^\top\|_F \leq (\|a\|_2 + \|b\|_2)\|a - b\|_2$ for any vectors $a, b$. To bound (C.14), we have

\[
\| g(x_{i,a_i}; \theta_0) \|_2, \| g(x_{i,a_i}; \theta_t) \|_2 \leq C_1 \sqrt{mL},
\]

where $C_1 > 0$ is a constant, the inequality holds due to Lemma B.6 with the fact that $\|\theta_t - \theta_0\|_2 \leq \tau$. We also have

\[
\| g(x_{i,a_i}; \theta_0) - g(x_{i,a_i}; \theta_t) \|_2 \leq C_2 \sqrt{\log m \tau^{1/3} L^3} \| g(x; \theta_0) \|_2 \leq C_3 \sqrt{m \log m \tau^{1/3} L^{7/2}},
\]

where $C_2, C_3 > 0$ are constants, the first inequality holds due to Lemma B.3 with the fact that $\|\theta_t - \theta_0\|_2 \leq \tau$, the second inequality holds due to Lemma B.6. Substituting (C.15) and (C.16) into (C.14), we have

\[
\|Z_t - Z_t\|_F \leq C_4 t \sqrt{\log m \tau^{1/3} L^4},
\]

where $C_4 > 0$ is a constant. We now bound $\log \det Z_t - \log \det Z_t$. It is easy to verify that $Z_t = \lambda I + J^\top J$, $Z_t = \lambda I + JJ^\top$, where

\[
\bar{J} = \left( g(x_{1,a_1}; \theta_0), \ldots, g(x_{t,a_t}; \theta_0) \right) / \sqrt{m},
\]

\[
J = \left( g(x_{1,a_1}; \theta_0), \ldots, g(x_{t,a_t}; \theta_{t-1}) \right) / \sqrt{m}.
\]

We have the following inequalities:

\[
\log \frac{\det(Z_t)}{\det(\lambda I)} - \log \frac{\det(Z_t)}{\det(\lambda I)} = \log \det(I + \bar{J}^\top J / \lambda) - \log \det(I + J J^\top / \lambda)
\]

\[
= \log \det(I + [J J^\top / \lambda)^{-1}, [J J^\top / \lambda)\bar{J} - [J J^\top / \lambda)J]
\]

\[
\leq \| (I + [J J^\top / \lambda)^{-1}, [J J^\top / \lambda)\bar{J} - [J J^\top / \lambda)J \|_F \leq \sqrt{t} \| (I + [J J^\top / \lambda)^{-1}, [J J^\top / \lambda)\bar{J} - [J J^\top / \lambda)J \|_F \leq \sqrt{t} \| [J J^\top / \lambda)^{-1}, [J J^\top / \lambda)\bar{J} - [J J^\top / \lambda)J \|_F, \]
\]
where the second equality holds due to the fact that $\det(I + AA^T) = \det(I + A^TA)$, the first inequality holds due to the fact that $\log \det$ function is convex, the second inequality hold due to the fact that $\langle A, B \rangle \leq \|A\|_F \|B\|_F$, the third inequality holds since $I + \|J\|/\lambda$ is a $t$-dimension matrix, the fourth inequality holds since $I + \|J\|/\lambda \geq I$. We have

$$
\|\|J\| - [J]^T J\|_F \leq \frac{t}{1 - \lambda} \max_{1 \leq i,j \leq t} \left| \langle g(x_{i,a}; \theta_0), g(x_{j,a}; \theta_0) \rangle - \langle g(x_{i,a}; \theta_j), g(x_{j,a}; \theta_j) \rangle \right| /m
$$

$$
\leq \frac{t}{1 - \lambda} \max_{1 \leq i,j \leq t} \left\{ \left\| g(x_{i,a}; \theta_0) - g(x_{i,a}; \theta_j) \right\|_2, \left\| g(x_{j,a}; \theta_0) - g(x_{j,a}; \theta_j) \right\|_2 \right\} /m
$$

$$
\leq \|C\| \sqrt{\log m \tau^1/3} L^4.
$$

where $C_5 > 0$ is a constant, the first inequality holds due to the fact that $\|A\|_F \leq t \max_{1 \leq i,j \leq t} |\langle A, B \rangle|$ for any $A, B \in \mathbb{R}^{t \times t}$, the second inequality holds due to the fact $\|a^T a' - b^T b'\|_2 \leq \|a - b\|_2 \|b'\|_2 + \|a' - b\|_2$, the third inequality holds due to [C.15] and [C.16]. Substituting (C.18) into (C.17), we have

$$
\log \frac{\det(Z_t)}{\det(\lambda I)} - \log \frac{\det(\lambda I)}{\det(Z_t)} \leq C_5 t^3/2 \sqrt{\log m \tau^{1/3} L^4}.
$$

Using the same method, we also have

$$
\log \frac{\det(Z_t)}{\det(\lambda I)} - \log \frac{\det(\lambda I)}{\det(Z_t)} \leq C_5 t^3/2 \sqrt{\log m \tau^{1/3} L^4}.
$$

This completes our proof.

\[\square\]

D Proofs of Lemmas in Appendix C

D.1 Proof of Lemma C.2

In this section we give the proof of Lemma C.2.

Proof of Lemma C.2. For any $j \in [J]$, the following inequalities hold. We first have

$$
\|J^{(j)}\|_F \leq \sqrt{t} \max_{i \in [t]} \left\| g(x_{i,a}; \theta^{(j)}) \right\|_2 \leq C_1 \sqrt{tmL}, \tag{D.1}
$$

where $C_1 > 0$ is a constant, the first inequality holds due to the definition of $J^{(j)}$, the second inequality holds due to Lemma B.6. We also have

$$
\|J^{(j)} - J^{(0)}\|_F \leq C_2 \sqrt{\log m \tau^{1/3} L^3} \|J^{(0)}\|_F \leq C_3 \sqrt{tm \log m \tau^{1/3} L^{7/2}}, \tag{D.2}
$$

where $C_2, C_3 > 0$ are constants, the first inequality holds due to Lemma B.5 with the assumption that $\|\theta^{(j)} - \theta^{(0)}\|_2 \leq \tau$, the second inequality holds due to (D.1). We also have

$$
\|f^{(j+1)} - f^{(j)} - [J^{(j)}]^T (\theta^{(j+1)} - \theta^{(j)})\|_2 \leq \max_{i \in [t]} \sqrt{t} \left| f(x_{i,a}; \theta^{(j+1)}) - f(x_{i,a}; \theta^{(j)}) - g(x_{i,a}; \theta^{(j)}) \right| g(x_{i,a}; \theta^{(j)} - \theta^{(j)})
$$

$$
\leq C_4 \tau^{4/3} L^3 \sqrt{tm \log m},
$$

where $C_4 > 0$ is a constant, the first inequality holds due to the the fact that $\|x\|_2 \leq \sqrt{t} \max_{i \in [t]} |x_i|$ for any $x \in \mathbb{R}^t$, the second inequality holds due to Lemma B.4 with the assumption that $\|\theta^{(j)} - \theta^{(0)}\|_2 \leq \tau_1, |\theta^{(j+1)} - \theta^{(0)}|_2 \leq \tau$. For $\|v\|_2$, we have $\|v\|_2 \leq \sqrt{t} \max_{i \leq t} |r(x_{i,a})| \leq \sqrt{t}$. This completes our proof.

\[\square\]
D.2  PROOF OF LEMMA C.3

Proof of Lemma C.3 Recall that the loss function \( L \) is defined as

\[
L(\theta) = \frac{1}{2} \| f(\theta) - v \|_2^2 + \frac{m\lambda}{2} \| \theta - \theta^{(0)} \|_2^2.
\]

We define \( J(\theta) \) and \( f(\theta) \) as follows:

\[
J(\theta) = \left( g(x_{1,a_1}; \theta), \ldots, g(x_{t,a_t}; \theta) \right) \in \mathbb{R}^{(md + m^2(L-2) + m)t},
\]

\[
f(\theta) = (f(x_1,a_1; \theta), \ldots, f(x_{t,a_t}; \theta))^\top \in \mathbb{R}^{t \times 1}.
\]

Suppose \( \| \theta - \theta^{(0)} \|_2 \leq \tau \). Then by the fact that \( \| \cdot \|_2^2 / 2 \) is 1-strongly convex and 1-smooth, we have the following inequalities:

\[
L(\theta') - L(\theta) \leq \langle \nabla L(\theta), \theta' - \theta \rangle + \frac{1}{2} \| \nabla L(\theta) \| \| \theta' - \theta \|_2^2
\]

where \( \epsilon = f(\theta') - f(\theta) - J(\theta)^\top (\theta' - \theta) \). \( I_1 \) can be bounded as follows:

\[
I_1 \leq \| f(\theta) - v \|_2^2 \| \epsilon \|_2^2 + \| J(\theta) \|_2^2 \| \theta' - \theta \|_2^2 + \frac{m\lambda}{2} \| \theta' - \theta \|_2^2
\]

\[
\leq C_1 \left( (m\lambda + tmL) \| \theta' - \theta \|_2^2 + \| f(\theta) - v \|_2^2 + \| \epsilon \|_2^2 \right),
\]

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to the fact that \( \| J(\theta) \|_2 \leq C_2 \sqrt{tmL} \) with \( \| \theta - \theta^{(0)} \|_2 \leq \tau \) by (C.3) in Lemma C.2

Substituting (D.4) into (D.3), we have

\[
L(\theta') - L(\theta) \leq \langle \nabla L(\theta), \theta' - \theta \rangle + \frac{C_1}{2} \left( (m\lambda + tmL) \| \theta' - \theta \|_2^2 + \| f(\theta) - v \|_2^2 + \| \epsilon \|_2^2 \right).
\]

Taking \( \theta' = \theta - \eta \nabla L(\theta) \), then by (D.5), we have

\[
L(\theta - \eta \nabla L(\theta)) - L(\theta) \leq -\eta \| \nabla L(\theta) \|_2^2 (1 - C_1 (m\lambda + tmL) \eta) + \| f(\theta) - v \|_2^2 + \| \epsilon \|_2^2.
\]

By the 1-strongly convexity of \( \| \cdot \|_2^2 \), we further have

\[
L(\theta') - L(\theta) \geq \langle f(\theta) - v, f(\theta') - f(\theta) \rangle + m\lambda \langle \theta - \theta^{(0)}, \theta' - \theta \rangle + \frac{m\lambda}{2} \| \theta' - \theta \|_2^2
\]

\[
= \langle f(\theta) - v, J(\theta)^\top (\theta' - \theta) + \epsilon \rangle + m\lambda \langle \theta - \theta^{(0)}, \theta' - \theta \rangle + \frac{m\lambda}{2} \| \theta' - \theta \|_2^2
\]

\[
= \langle \nabla L(\theta), \theta' - \theta \rangle + \frac{m\lambda}{2} \| \theta' - \theta \|_2^2 + \langle f(\theta) - v, \epsilon \rangle
\]

\[
\geq \langle \nabla L(\theta), \theta' - \theta \rangle + \frac{m\lambda}{2} \| \theta' - \theta \|_2^2 - \| f(\theta) - v \|_2 \| \epsilon \|_2
\]

\[
\geq -\frac{\| \nabla L(\theta) \|_2^2}{2m\lambda} - \| f(\theta) - v \|_2 \| \epsilon \|_2,
\]
where the second inequality holds due to Cauchy-Schwarz inequality, the last inequality holds due
to the fact that $\langle a, x \rangle + c\|x\|^2 \geq -\|a\|^2/(4c)$ for any vectors $a$, $x$ and $c > 0$. Substituting (D.7)
into (D.6), we have

\[
L(\theta' - \eta \nabla L(\theta)) - L(\theta)
\]

\[
\leq 2m\lambda\eta(1 - C_1(m\lambda + tm\lambda\eta)) [L(\theta') - L(\theta) + \|f(\theta) - v\|_2^2] + \|f(\theta) - v\|_2^2 + \|e\|^2_2
\]

\[
\leq m\lambda\eta[L(\theta') - L(\theta)] + \|f(\theta) - v\|_2^2 + \|e\|^2_2
\]

\[
\leq m\lambda\eta[L(\theta') - L(\theta)] + \|f(\theta) - v\|_2^2 + 8 + 4\|e\|^2_2 + m\lambda\eta\|f(\theta) - v\|_2^2 + 4\|e\|^2_2/(m\lambda\eta) + \|e\|^2_2
\]

\[
\leq m\lambda\eta[L(\theta') - L(\theta)]/2 + \|e\|^2_2(1 + 4m\lambda\eta + 4/(m\lambda\eta)),
\]

(D.8)

where the second inequality holds due to the fact of $\eta$, third inequality holds due to Young’s inequality, fourth inequality holds due to the fact that $\|f(\theta) - v\|^2_2 \leq 2L(\theta)$. Now taking $\theta = \theta^{(j)}$
and $\theta' = \bar{\theta}^{(0)}$, rearranging (D.8), with the fact that $\theta^{(j+1)} = \theta^{(j)} - \eta \nabla L(\theta^{(j)})$, we have

\[
L(\theta^{(j+1)}) - L(\theta^{(0)})
\]

\[
\leq (1 - m\lambda\eta/2)[L(\theta^{(j)}) - L(\theta^{(0)})] + m\lambda\eta/2L(\theta^{(0)}) + \|e\|^2_2(1 + 4m\lambda\eta + 4/(m\lambda\eta))
\]

\[
\leq (1 - m\lambda\eta/2)[L(\theta^{(j)}) - L(\theta^{(0)})] + m\lambda\eta/2 \cdot t + m\lambda\eta/2 \cdot t
\]

\[
\leq (1 - m\lambda\eta/2)[L(\theta^{(j)}) - L(\theta^{(0)})] + m\lambda\eta t,
\]

(D.9)

where the second inequality holds due to the fact that $L(\theta^{(0)}) = \|f(\theta^{(0)}) - v\|^2_2 = \|v\|^2_2 \leq t$, and

\[
(1 + 4m\lambda\eta + 4/(m\lambda\eta))\|e\|^2_2 \leq 5/(m\lambda\eta) \cdot C_22^{8/3}L^6tm\log m \leq tm\lambda\eta/2,
\]

(D.10)

where the first inequality holds due to (C.5) in Lemma C.2 the second inequality holds due to the
choice of $\tau$. Expanding (D.9) $t$ times, we have

\[
L(\theta^{(j+1)}) - L(\theta^{(0)}) \leq \frac{m\lambda\eta t}{m\lambda\eta/2} = 2t,
\]

which implies that $\|f^{(j+1)} - v\|^2_2 \leq 2\sqrt{t}$. This completes our proof. \qed

D.3 Proof of Lemma C.4

In this section we prove Lemma C.4.

Proof ofLemma C.4 It is worth noting that $\bar{\theta}^{(j)}$ is the sequence generated by applying gradient
descent on the following problem:

\[
\min_{\theta} \tilde{L}(\theta) = \frac{1}{2}\|J(0)^\top(\theta - \theta^{(0)}) - v\|^2_2 + \frac{m\lambda}{2}\|\theta - \theta^{(0)}\|^2_2.
\]

Then $\|\theta^{(0)} - \bar{\theta}^{(j)}\|_2$ can be bounded as

\[
\frac{m\lambda}{2}\|\theta^{(0)} - \bar{\theta}^{(j)}\|^2_2 \leq \frac{1}{2}\|J(0)^\top(\theta^{(j)} - \theta^{(0)}) - v\|^2_2 + \frac{m\lambda}{2}\|\bar{\theta}^{(j)} - \theta^{(0)}\|^2_2
\]

\[
\leq \frac{1}{2}\|J(0)^\top(\theta^{(0)} - \theta^{(0)}) - v\|^2_2 + \frac{m\lambda}{2}\|\tilde{\theta}^{(0)} - \theta^{(0)}\|^2_2
\]

\[
\leq T/2,
\]

where the first inequality holds trivially, the second inequality holds due to the monotonic decreasing
property brought by gradient descent, the third inequality holds due to (C.6) in Lemma C.2. It is easy to verify that $\tilde{L}$ is a $\lambda$-strongly convex function and $C_1(tmL + m\lambda)$-smooth function, since

\[
\nabla^2 \tilde{L} \preceq (\|J(0)\|^2_2 + m\lambda)I \preceq C_1(tmL + m\lambda),
\]

where the first inequality holds due to the definition of $\tilde{L}$, the second inequality holds due to (C.3)
in Lemma C.2. Since we choose $\eta \leq C_2(tmL + m\lambda)^{-1}$ for some small enough $C_2 > 0$, then by
standard result of gradient descent on ridge linear regression, $\bar{\theta}^{(j)}$ converges to $\theta^{(0)} + (Z)^{-1}b/\sqrt{m}$
with the convergence rate

\[
\|\bar{\theta}^{(j)} - \theta^{(0)} - (Z)^{-1}b/\sqrt{m}\|^2_2 \leq (1 - \eta m\lambda)^j\|f^{(0)} - v\|^2_2/\sqrt{m} \lambda \leq (1 - \eta m\lambda)^j\sqrt{t/(m\lambda)}.
\]

\qed
E  A Variant of NeuralUCB

In this section, we present a variant of NeuralUCB called NeuralUCB\textsubscript{0}. Compared with Algorithm\textsubscript{1}, the main differences between NeuralUCB and NeuralUCB\textsubscript{0} are as follows: NeuralUCB uses gradient descent to train a deep neural network to learn the reward function \( h(x) \) based on observed contexts and rewards. In contrast, NeuralUCB\textsubscript{0} uses matrix inversions to obtain parameters in closed forms. At each round, NeuralUCB uses the current DNN parameters (\( \theta_t \)) to compute an upper confidence bound. In contrast, NeuralUCB\textsubscript{0} computes the UCB using the initial parameters (\( \theta_0 \)).

Algorithm 3 NeuralUCB\textsubscript{0}

1: **Input:** number of rounds \( T \), regularization parameter \( \lambda \), exploration parameter \( \nu \), confidence parameter \( \delta \), norm parameter \( S \), network width \( m \), network depth \( L \)
2: **Initialization:** Generate each entry of \( W_l \) independently from \( N(0, 2/m) \) for \( 1 \leq l \leq L - 1 \), and each entry of \( W_L \) independently from \( N(0, 1/m) \). Define \( \phi(x) = g(x; \theta_0) / \sqrt{m} \), where \( \theta_0 = [\text{vec}(W_1)^\top, \ldots, \text{vec}(W_L)^\top]^\top \in \mathbb{R}^p \)
3: \( Z_0 = \lambda I \), \( b_0 = 0 \)
4: for \( t = 1, \ldots, T \) do
5: Observe \( \{x_{t,a}\}_{a=1}^K \) and compute
6: Play \( a_t \) and receive reward \( r_{t,a_t} \)
7: Compute \( Z_t = Z_{t-1} + \phi(x_{t,a_t})\phi(x_{t,a_t})^\top \in \mathbb{R}^{p \times p} \), \( b_t = b_{t-1} + r_{t,a_t} \phi(x_{t,a_t}) \in \mathbb{R}^p \)
8: Compute \( \theta_t = Z_t^{-1}b_t + \theta_0 \in \mathbb{R}^p \)
9: Construct \( C_t \) as
10: end for

\[ (a_t, \tilde{\theta}_{t,a_t}) = \arg\max_{a \in [K], \theta \in C_{t-1}} \langle \phi(x_{t,a}), \theta - \theta_0 \rangle \] (E.1)

\[ C_t = \{ \theta : \| \theta_t - \theta \|_{Z_t} \leq \gamma_t \}, \quad \text{where} \quad \gamma_t = \nu \sqrt{\log \frac{\det Z_t}{\det \lambda I} - 2 \log \delta + \sqrt{\lambda S}} \] (E.2)