Stochastic Bandits with ReLU Neural Networks

Kan Xu [1] **Hamsa Bastani** [2] **Surbhi Goel** [2] **Osbert Bastani** [2]

Abstract

We study the stochastic bandit problem with ReLU neural network structure. We show that a $\tilde{O}(\sqrt{T})$ regret guarantee is achievable by considering bandits with one-layer ReLU neural networks; to the best of our knowledge, our work is the first to achieve such a guarantee. In this specific setting, we propose an OFU-ReLU algorithm that can achieve this upper bound. The algorithm first explores randomly until it reaches a *linear* regime, and then implements a UCB-type linear bandit algorithm to balance exploration and exploitation. Our key insight is that we can exploit the piecewise linear structure of ReLU activations and convert the problem into a linear bandit in a transformed feature space, once we learn the parameters of ReLU relatively accurately during the exploration stage. To remove dependence on model parameters, we design an OFU-ReLU+ algorithm based on a batching strategy, which can provide the same theoretical guarantee.[1]

1. Introduction

The stochastic contextual bandit problem has been widely studied in the literature (Bubeck et al., 2012; Lattimore & Szepesvári, 2020), with broad applications in healthcare (Bastani & Bayati, 2020), personalized recommendation (Li et al., 2010), etc. The problem is important since real-world decision-makers oftentimes adaptively gather information about their environment to learn. Formally, the bandit algorithm actively selects a sequence of actions $\{x_t\}_{t\in[T]}$ with $x_t \in \mathcal{X}$ over some horizon $T \in \mathbb{N}$, and observes stochastic rewards $y_t = f_{\Theta^*}(x_t) + \xi_t$, where f_{Θ^*} is the true reward function (represented by a model with parameters Θ^*), and ξ_t is random noise. Thus, to achieve good performance or low regret, the decision-maker must maintain small decision error uniformly across all actions x_t over time to ensure that it generalizes to new, actively selected actions.

With the success of bandits in practice, there has been a great deal of recent interest in understanding the theoretical properties of bandit algorithms. For linear models, i.e., the expected reward f_{Θ^*} is linear in x_t, bandit algorithms have adapted techniques from statistics to address this challenge (Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011). Particularly, linear bandit algorithms build on linear regression, which provides parameter estimation bounds of the form $\|\hat{\theta} - \theta^*\|_2 \leq \epsilon$; then, we obtain uniform generalization bounds of the form $|f_{\hat{\theta}}(x) - f_{\theta^*}(x)| \leq L\epsilon, \forall x \in \mathcal{X}$, where L is a Lipschitz constant for f_θ. By adapting these techniques, linear bandit algorithms can achieve minimax rates of $\tilde{O}(\sqrt{T})$ (Dani et al., 2008; Abbasi-Yadkori et al., 2011) in terms of regret. However, the linear assumption oftentimes do not hold for complicated tasks (Valko et al., 2013), and has recently motivated the study of nonlinear contextual bandits, especially building upon neural networks with ReLU activations (see, e.g., Zhou et al., 2020; Zhang et al., 2020; Xu et al., 2020; Kassraie & Krause, 2022).

One of the key questions is what kinds of guarantees can be provided in such settings when the underlying true reward model has ReLU structures. The current existing literature of bandits based on ReLU neural networks mainly base their analyses on the theory of neural tangent kernel (NTK) (Jacot et al., 2018). Zhou et al. (2020); Gu et al. (2024) leverage NTK and upper confidence bound (UCB) techniques to achieve a $\tilde{O}(\gamma_T\sqrt{T})$ regret bound, where γ_T is the effective dimension or maximum information gain and is assumed to be T-independent in these literature. This assumption is strong because it intuitively assumes the function is linear on a low-dimensional subspace (more precisely, it says that the eigenvalues of the empirical covariance matrix in the kernel space vanish quickly, so the covariates "approximately" lie in a low-dimensional subspace); thus, it effectively converts the problem to linear bandit problem in a high-dimensional subspace. Indeed, for ReLU neural networks on a d-dimensional domain (i.e., $x \in \mathbb{R}^d$), (Kassraie & Krause, 2022) shows a best upper bound for the information gain known as $\gamma_T = \tilde{O}(T^{\frac{d-1}{d}})$, even for a one-layer neural network. Consequently, the re-

[1] Arizona State University, Arizona, USA [2] University of Pennsylvania, Pennsylvania, USA. Correspondence to: Kan Xu <kanxu1@asu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] Source code is available at https://github.com/kanxu526/ReLUBandit.

gret bound provided above becomes superlinear even for $d > 1$ without further restrictive assumptions. (Kassraie & Krause, 2022) improve upon this regret bound based on a variant of (Zhou et al., 2020) and obtain a sublinear bound of $\tilde{O}((\gamma_T T)^{1/2}) = \tilde{O}(T^{\frac{2d-1}{2d}})$, but is still far from the typical $\tilde{O}(\sqrt{T})$ guarantee.

Contribution. In contrast, we want to shed light on the achievable regret guarantee for a nonlinear bandit problem with ReLU neural network structure by estimating the model f_{Θ^*} directly (without making an effective dimension assumption using NTK techniques). Due to the complexity of the problem, we consider the setting of one-layer ReLU neural network as the true reward function. We design two bandit algorithms that exploit the piecewise linear structure of ReLU activations; *to the best of our knowledge, we provide the first $\tilde{O}(\sqrt{T})$ regret bound for bandit learning with ReLU neural networks.*

Our first bandit algorithm OFU-ReLU is designed based on the following insight. Let our true reward function have the following ReLU structure with k neurons (see, e.g., Du & Lee, 2018; Zhang et al., 2019)

$$f_{\Theta^*}(x) = \sum_{i\in[k]} \theta_i^{*\top} x \cdot \mathbb{1}(\theta_i^{*\top} x \geq 0),$$

where $\Theta^* = [\theta_1^*, \cdots, \theta_k^*]^\top$. Intuitively, once we learn a sufficiently good estimate $\tilde{\theta}_i$ that is close to its corresponding true neuron parameter θ_i^*, we can "freeze" the contribution of the indicator function by $\mathbb{1}(\tilde{\theta}_i^\top x \geq 0)$. The problem then becomes linear, and we can use a UCB-type linear bandit algorithm to achieve good performance. This insight motivates our two-phase bandit algorithm, where we first randomly explore to estimate the parameters, and then run linear bandits once we reach the linear regime to obtain at a minimax optimal rate.

We want to note that even though this strategy may seem straightforward, such a two-stage design with a phase transition from exploration to bandit learning has been shown inevitable for specific nonlinear bandit problems (Rajaraman et al., 2023), and might be a more general phenomenon. In addition, applying UCB-type algorithm (i.e., Eluder UCB (Russo & Van Roy, 2013)) directly on the reward has been demonstrated suboptimal for certain family of nonlinear bandit problem, posing an interesting theoretical challenge (Rajaraman et al., 2023). *In contrast, we show that instead of applying UCB to the ReLU structure directly, we will be able to reduce it to a linear bandit problem and this will make UCB optimal again.*

Our second bandit algorithm OFU-ReLU+ is designed to eliminate the assumed knowledge in OFU-ReLU of the minimum gap ν_* between the optimal action x^* and any neuron θ_i^* (we prove such a gap always exists for our ReLU structure). As long as the estimate $\tilde{\theta}_i$ is within $\nu_*/2$ of the true neuron θ_i^*, the indicator estimate is guaranteed to be consistent with the true indicator value at the optimal action $x = x^*$, i.e., $\mathbb{1}(\tilde{\theta}_i^\top x^* \geq 0) = \mathbb{1}(\theta_i^{*\top} x^* \geq 0)$. In other words, unless our estimate $\tilde{\theta}$ has less than a ν_*-dependent error, the optimal action x^* would lie near a nonlinear regime of a neuron, which might fail the following linear bandit learning. Yet, in practice we do not know ν_*; thus, we design a batching strategy that first makes a guess on ν_* and keep cutting this guess every batch so that our estimate $\tilde{\theta}_i$ will be accurate enough after a constant number of batches. *Different from the previous batching strategies (see, e.g., Golrezaei et al., 2019; Luo et al., 2022), our exploration and OFUL phase both use samples from all previous batches without discarding data from previous batches.*

Finally, we provide a parameter estimation error bound for one-layer ReLU neural networks that can ensure theoretical guarantee for each neuron independently through a novel proof strategy, which might be of separate interest.

1.1. Other Related Work

Early work on stochastic bandits focus on linear rewards (Dani et al., 2008; Chu et al., 2011; Abbasi-Yadkori et al., 2011) and typically use an UCB algorithm. Later, it has been extended to kernel based models, where the reward functions belong to the reproducing kernel Hilbert space (RKHS) (Valko et al., 2013). Along this line, the recent ReLU bandit literature are build upon the kernelized algorithm using NTK and achieve an effective-dimension-dependent bound (Zhou et al., 2020; Xu et al., 2020; Kassraie & Krause, 2022; Gu et al., 2024); (Salgia, 2023) generalize to smooth activations and provide a bound $\tilde{O}(T^{\frac{2d+2s-3}{2d+4s-4}})$ depending on smoothness s. Dong et al. (2021) study the optimization scheme of nonlinear bandit, and provide a $\tilde{O}(T^{3/4})$ and $\tilde{O}(T^{7/8})$ local and global regret for two-layer neural network bandit. (Rajaraman et al., 2023) focus specifically on the ridge function family (which does not include ReLU but only ReLU with single neuron), and design an explore-then-commit strategy. Finally, there are also many other works that study neural network bandits with other activations, such as quadratic activations (Xu et al., 2021a), polynomial functional form (Huang et al., 2021), etc.

2. Problem Formulation

Notation. Let $[n] = \{1, \cdots, n\}$. We let $S^{d-1}(r) \subseteq \mathbb{R}^d$ denote the $(d-1)$-sphere in d dimensions with radius r, and let $S^{d-1} = S^{d-1}(1)$. Define $A^{d-1}(r)$ to be the area of the $(d-1)$-sphere with radius r (i.e., $A^{d-1}(r) = |S^{d-1}(r)|$). We use $\vee$ to represent the maximum value. Define p to be the distribution of the covariates, i.e., $x \sim p$.

ReLU Neural Network. Consider a function family $f_\Theta : \mathcal{X} \to \mathcal{Y}$ (where $\mathcal{X} = S^{d-1} \subseteq \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and $\Theta \in \mathit{\Theta}$ for a given domain $\mathit{\Theta} \subset \mathbb{R}^{k \times d}$) consisting of neural networks with ReLU activations (see, e.g., Du & Lee, 2018; Zhang et al., 2019)

$$f_\Theta(x) = \sum_{i \in [k]} g(\theta_i^\top x), \tag{1}$$

where θ_i is the i^{th} component of the parameter $\Theta = [\theta_1, \theta_2, \cdots, \theta_k]^\top$ (which we call a *neuron*) and $g(z) = z \cdot \mathbb{1}(z \geq 0)$. Let Θ^* be the ground-truth parameters.

We are provided with n data points $Z = \{(x_i, y_i)\}_{i \in [n]}$ such that $y_i = f_{\Theta^*}(x_i) + \xi_i$, where ξ_i is i.i.d. σ-subgaussian random noise with mean 0. Assume the covariates follow certain distribution $x \sim p$. Then, the population mean squared loss is defined as

$$L_{S,p}(\Theta) = \mathbb{E}_p\left[(f_\Theta(x) - f_{\Theta^*}(x))^2\right].$$

We slightly abuse our notation and use p to denote the joint distribution of each training example (x_i, y_i); the corresponding empirical mean squared loss is

$$\hat{L}_S(\Theta; Z) = \frac{1}{n} \sum_{i \in [n]} (f_\Theta(x_i) - y_i)^2. \tag{2}$$

For some neural network f_{Θ^*}, we obtain an estimate $\hat{\Theta}$ for the parameter Θ^* by taking the minimizer of $\hat{L}_S(\Theta; Z)$, i.e., $\hat{\Theta} = \arg\min_{\Theta \in \mathit{\Theta}} \hat{L}_S(\Theta; Z)$.

Assumptions. We provide a statistical guarantee for parameter estimation of ReLU neural network under the following assumptions.

Assumption 2.1. $\|\theta_i^*\|_2 = 1$ holds for all neurons $i \in [k]$.

Note that under the above assumption, the domain of Θ, i.e., $\mathit{\Theta}$, is included in $\{\Theta \in \mathbb{R}^{k \times d} \mid \|\Theta\|_{2,1} = k\}$, where $\|\Theta\|_{2,1} \coloneqq \sum_{i \in [k]} \|\theta_i\|_2$ is the $\ell_{2,1}$-norm.

Assumption 2.2. There exists a constant $\alpha_0 > 0$ such that

$$\min_{j, j' \in [k], j \neq j'} \|\theta_j^* \pm \theta_{j'}^*\|_2 \geq \alpha_0.$$

Collectively, our assumptions limit the structure of the neural network. As evidenced by lower bounds in (Dong et al., 2021), some restrictions on the structure are necessary to obtain regret $\tilde{\mathcal{O}}(T^\alpha)$ for $\alpha < 1$.

Assumption 2.1 is critical for our analysis, since neurons become hard to estimate when they are small. Similarly, Assumption 2.2 is necessary since two neurons that are close together are hard to distinguish. The assumption that the second layer consists of weights ones is not critical—most of our proofs can be extended to the case where the second layer has values in $\{\pm 1\}$.

ReLU Bandit. We consider the following bandit problem. At each step t, we choose an action $x_t \in \mathcal{X}$, and observe reward

$$y_t = f_{\Theta^*}(x_t) + \xi_t, \tag{3}$$

where ξ_t is i.i.d. σ-subgaussian random noise with mean 0. Here, the true parameters Θ^* are unknown and will be learned in an online manner. Our goal is to minimize the cumulative regret R_T over a time horizon T:

$$R_T = \sum_{t=1}^{T} r_t, \quad r_t = f_{\Theta^*}(x^*) - f_{\Theta^*}(x_t)$$

where $x^* = \arg\max_{x \in \mathcal{X}} f_{\Theta^*}(x)$ is the optimal action and r_t is the per period regret.

3. Parameter Estimation for ReLU Neural Networks

We provide an estimation error bound on the parameters of ReLU neural networks, which is important for the convergence of our bandit algorithm. Since the population loss function is symmetric regarding the neurons, meaning that any column-permutation of Θ^* achieves zero loss, our bound shows the existence of some mapping $\sigma : [k] \to [k]$ from the ground truth neurons θ_i^* to the estimated neurons $\hat{\theta}_{\sigma(i)}$ such that $\hat{\theta}_{\sigma(i)} \approx \theta_i^*$.

First, we show the following proposition, which states that a small generalization error on a special subset of the unit sphere $X_i(\epsilon) \subset \mathcal{X} = S^{d-1}$ for $\epsilon \in \mathbb{R}_{>0}$, i.e.,

$$X_i(\epsilon) = \{x \in \mathcal{X} \mid |\theta_i^{*\top} x| \leq \epsilon\},$$

implies a small parameter estimation error of the corresponding neuron θ_i^* up to a sign flip.

Proposition 3.1. *Suppose*

$$L_{X_i(\epsilon)}(\Theta) \coloneqq \int_{X_i(\epsilon)} |f_\Theta(x) - f_{\Theta^*}(x)| dx \leq \eta$$

for some $\eta \in \mathbb{R}_{>0}$. Then, there exists a bijection $\sigma : [k] \to [k]$ such that

$$\min\{\|\theta_{\sigma(i)} - \theta_i^*\|_2, \|\theta_{\sigma(i)} + \theta_i^*\|_2\} \leq h(\eta, \epsilon),$$

where $h(\eta, \epsilon) \coloneqq \frac{k\epsilon^3 |S^{d-3}|/2}{\epsilon^2(1 - d\epsilon^2/2)|S^{d-2}|/8 - \eta - 6kd\epsilon^3|S^{d-2}|}$.

We provide a detailed proof with illustrations in Appendix A. Intuitively, X_i contains those $x \in \mathcal{X}$ close to the boundary where the neuron θ_i^* of f_{Θ^*} is nonlinear (i.e., $\theta_i^{*\top} x = 0$). Then, this proposition claims that if the loss on certain X_i is small, then the corresponding neuron θ_i^* in the ground-truth neural network f_{Θ^*} can be identified up to a sign flip by a corresponding neuron $\theta_{\sigma(i)}$ in the approximate neural network f_Θ, i.e., $\theta_{\sigma(i)} \approx \theta_i^*$ or $\theta_{\sigma(i)} \approx -\theta_i^*$. Moreover, if the loss on all X_i's with $i \in [k]$ are small, then Θ is close to Θ^* via the mapping σ up to sign flips.

Our main result combines Proposition 3.1 with a standard generalization error bound for ReLU neural networks to obtain the following parameter estimation error bounds.

Theorem 3.2. *Suppose the distribution p satisfies*

$$\frac{1}{|S^{d-1}|}\int_{\mathcal{X}} |f_\Theta(x) - f_{\Theta^*}(x)|dx \le \mathbb{E}_p\left[|f_\Theta(x) - f_{\Theta^*}(x)|\right].$$

Then, there exists a bijection $\sigma : [k] \to [k]$ such that

$$\min\{\|\hat{\theta}_{\sigma(i)} - \theta_i^*\|_2, \|\hat{\theta}_{\sigma(i)} + \theta_i^*\|_2\} \le 727\pi^{-\frac{1}{4}} k d^{\frac{1}{4}} (2\zeta)^{\frac{1}{4}}$$

holds for all $i \in [k]$ with at least a probability $1 - \delta$, where $\zeta = \tilde{\mathcal{O}}(\sqrt{k^5 d/n})$.

We give a proof and the expression of ζ in Appendix B.

One potential limit of our proof strategy is that we may not correctly identify the sign of the ground truth neurons—i.e., our guarantee has the form $\hat{\theta}_{\sigma(i)} \approx \pm\theta_i^*$. However, we show in the next section that this caveat does not affect our bandit algorithm and analysis. Particularly, we show the difference between the estimated neural network $f_{\hat{\Theta}}$ and the true model f_{Θ^*} can be captured by a linear structure under sign misspecification. Thus, we can still run linear bandit algorithm to learn f_{θ^*} in an online manner.

4. Algorithms for ReLU Bandits

Now, we describe our bandit algorithm and provide corresponding regret analysis. We begin with a simple case where we know the gap between the optimal action x^* and the nearest neuron, and then provide a solution when this knowledge of gap is not assumed.

4.1. Algorithm Design

We first provide intuition for our design choices. The challenge of running a ReLU bandit algorithm is the nonlinearity of the ReLU neural network, which is due to the indicator function in the ReLU activations—i.e., in

$$\theta_i^\top x \cdot \mathbb{1}(\theta_i^\top x \ge 0),$$

the first occurrence of θ_i is the linear contribution and the second occurrence is the contribution via the indicator function. The key of our design to tackle the nonlinearity is to first learn the indicator contribution, and then use a typical linear bandit algorithm to keep updating the model given the indicator function fixed. Particularly, once we have a sufficiently good estimate $\tilde{\theta}_i \approx \theta_i^*$, then our estimate of the indicator is exact:

$$\mathbb{1}(\tilde{\theta}_i^\top x^* \ge 0) = \mathbb{1}(\theta_i^{*\top} x^* \ge 0), \quad \forall i \in [k]. \tag{4}$$

Next, we can fix the value of the indicator function using $\tilde{\theta}_i$ and focus on learning the linear part. That is, we run a linear bandit algorithm with the value of $\theta_i = \tilde{\theta}_i$ in the indicator functions, but keep learning the value of θ_i in the linear part. In more detail, if (4) holds, then we have

$$\begin{aligned}\mathbb{E}[y] = f_{\Theta^*}(x) &= \sum_{i\in[k]} (\mathbb{1}(\tilde{\theta}_i^\top x \ge 0)x)^\top \theta_i^* \\ &= \underbrace{\begin{bmatrix}\mathbb{1}(\tilde{\theta}_1^\top x \ge 0)x \\ \mathbb{1}(\tilde{\theta}_2^\top x \ge 0)x \\ \vdots \\ \mathbb{1}(\tilde{\theta}_k^\top x \ge 0)x\end{bmatrix}^\top}_{x^\dagger} \underbrace{\begin{bmatrix}\theta_1^* \\ \theta_2^* \\ \vdots \\ \theta_k^*\end{bmatrix}}_{\theta^\dagger}.\end{aligned} \tag{5}$$

Equivalently, we can run a linear bandit to learn $f_{\Theta^*}(x)$, however, with action being the term $x^\dagger \in \mathbb{R}^{dk}$ in (5), and parameter being the term $\theta^\dagger \in \mathbb{R}^{dk}$. The action $x^\dagger$ is a function of the original action x and the estimated parameter $\tilde{\Theta}$, and the parameter to update $\theta^\dagger$ is a vectorization of the parameter Θ of interest.

Challenges. Though the above two-stage design looks intuitive, there are still two challenges associated with (4) and (5). On one hand, even if we have a relatively accurate estimate $\tilde{\theta}_i$ of θ_i^*, (4) might not hold for some action x close to both of the estimate $\tilde{\theta}_i$ and the neuron θ_i^*. On the other hand, our theoretical result in §3 only provides a guarantee on $\tilde{\theta}_i$ up to a sign flip, and hence still (4) might not hold; thus, we also need to design our algorithm capturing this bias.

As suggested above, note that for any x such that $\theta_i^{*\top} x \approx 0$, even if $\tilde{\theta}_i \approx \theta_i^*$, it may still be the case that (4) fails to hold. As a consequence, (5) does not hold, and $f_{\Theta^*}(x)$ is not linear in $x^\dagger$ for x close to any of θ_i^*'s. In other words, the linear bandit is misspecified in some action region, so the algorithm may not converge if the optimal action lies in such regions. Thus, our regret bounds depend on the gap between the optimal action $x^* = \arg\max_{x\in\mathcal{X}} f_{\Theta^*}(x)$ and the nearest hyperplane that is perpendicular to one of the neurons in the ground-truth neural network.

Definition 4.1. A ReLU neural network with parameters Θ^* has a ν_*-gap for $\nu_* \in \mathbb{R}_{\ge 0}$ if $|\theta_i^{*\top} x^*| \ge \nu_*$ for all $i \in [k]$.

The following result ensures that a nontrivial gap $\nu_* > 0$ always exists for our ReLU structure.

Proposition 4.2. $\min_{i\in[k]} |\theta_i^{*\top} x^*| > 0$ *holds for the optimal action x^*.*

We give a proof in Appendix C. Note that our proof strategy relies on the ReLU structure where the weights of the second layer equal 1. In other words, any ReLU neural network structure as in (3) that satisfies our assumptions has a positive gap $\nu_* = \min_{i\in[k]} |\theta_i^{*\top} x^*| > 0$.

Given this gap ν_*, as long as our estimate $\tilde{\theta}_i$ has an estimation error of θ_i^* smaller than $\nu_*/2$, i.e., $\|\tilde{\theta}_i - \theta_i^*\| \leq \nu_*/2$, our bandit algorithm will be able to find the optimal action x^* in the action space $\mathcal{X}(\tilde{\Theta}, \nu_*/2)$, where

$$\mathcal{X}(\Theta, \nu) := \{x \in \mathcal{X} \mid |\theta_i^\top x| \geq \nu, \forall i \in [k]\}. \tag{6}$$

Intuitively, the above claim holds for two reasons: (i) $x^* \in \mathcal{X}(\tilde{\Theta}, \nu_*/2)$, and (ii) the bandit model $f_{\Theta^*}(x)$ is linear in $x^\dagger$ for any $x \in \mathcal{X}(\tilde{\Theta}, \nu_*/2)$ (recall $x^\dagger$ is a function of x). For (i), it suffices to show that $\mathcal{X}(\Theta^*, \nu_*) \subseteq \mathcal{X}(\tilde{\Theta}, \nu_*/2)$, as our ReLU neural network in (1) has a positive ν_*-gap and hence $x^* \in \mathcal{X}(\Theta^*, \nu_*)$. To this end, for any $x \in \mathcal{X}(\Theta^*, \nu_*)$, if $\theta_i^{*\top} x > 0$, we have

$$\begin{aligned}\tilde{\theta}_i^\top x &\geq \theta_i^{*\top} x - |\tilde{\theta}_i^\top x - \theta_i^{*\top} x| \\ &\geq \theta_i^{*\top} x - \|\tilde{\theta}_i - \theta_i^*\|_2 \\ &\geq \nu_* - \nu_*/2 = \nu_*/2 > 0. \end{aligned} \tag{7}$$

Similarly, if $\theta_i^{*\top} x < 0$, we have $\tilde{\theta}_i^\top x \leq \nu_*/2$. Next, to show (ii), it suffices to show that for any $x \in \mathcal{X}(\tilde{\Theta}, \nu_*/2)$, we have $\mathbb{1}(\theta_i^{*\top} x \geq 0) = \mathbb{1}(\tilde{\theta}_i^\top x \geq 0)$. Following a similar argument as (7), we can show that $\theta_i^{*\top} x^* \geq 0$ if $\tilde{\theta}_i^\top x \geq 0$ (and similarly for $\leq$). Thus, our plug-in indicator function is consistent with the true indicator function.

The other challenge is the sign misidentification from our Theorem 3.2. Specifically, $\tilde{\Theta}$ is close to the true parameters Θ^* only up to signs. In the worst case, the values of the corresponding indicators $\mathbb{1}(\tilde{\theta}_i^\top x \geq 0)$ may differ from the true values $\mathbb{1}(\theta_i^{*\top} x \geq 0)$ when $\tilde{\theta}_i$ is close to $-\theta_i^*$ instead of θ_i^*. In other words, the function $f_{\Theta^*}(x)$ will still be nonlinear of $x^\dagger$ and (5) may no longer hold, leading to misspecification, when $\|\tilde{\theta}_i + \theta_i^*\|_2 \leq \nu_*/2$ instead of $\|\tilde{\theta}_i - \theta_i^*\|_2 \leq \nu_*/2$, even if we reduce our search region to $\mathcal{X}(\tilde{\Theta}, \nu_*/2)$

In order to correct this misspecification bias, we propose to add additional k delicately designed linear components to (5). We show that the misspecification when $\|\tilde{\theta}_i + \theta_i^*\|_2 \leq \nu_*/2$ can be captured by a linear structure of k additional transformed features of x, enabling the linear bandit algorithm to function again. In more detail, we can write

$$\begin{aligned} f_{\Theta^*}(x) = &\sum_{i \in [k]} (\mathbb{1}(\tilde{\theta}_i^\top x \geq 0)x)^\top \theta_i^* \\ &+ \sum_{i \in [k]} ((\frac{1}{2} - \mathbb{1}(\tilde{\theta}_i^\top x \geq 0)x)^\top \\ &\cdot \left(\frac{\mathbb{1}(\theta_i^{*\top} x \geq 0) - \mathbb{1}(\tilde{\theta}_i^\top x \geq 0)}{\frac{1}{2} - \mathbb{1}(\tilde{\theta}_i^\top x \geq 0)} \theta_i^* \right). \end{aligned}$$

Note that compared to (5), we have an additional second term that captures the misspecification; this term equals 0 for any $x \in \mathcal{X}(\tilde{\Theta}, \nu_*/2)$ when $\|\tilde{\theta}_i - \theta_i^*\| \leq \nu_*/2$, as we have shown in (7). Similarly, when $\|\tilde{\theta}_i + \theta_i^*\| \leq \nu_*/2$, we can show that $\mathbb{1}(\tilde{\theta}_i^\top x \geq 0) = 1 - \mathbb{1}(\theta_i^{*\top} x \geq 0)$—i.e., if $\theta_i^{*\top} x \gtreqless 0$, we have $\tilde{\theta}_i^\top x \lesseqgtr 0$ for any $x \in \mathcal{X}(\tilde{\Theta}, \nu_*/2)$. In other words, we have

$$\frac{\mathbb{1}(\theta_i^{*\top} x \geq 0) - \mathbb{1}(\tilde{\theta}_i^\top x \geq 0)}{\frac{1}{2} - \mathbb{1}(\tilde{\theta}_i^\top x \geq 0)} = \begin{cases} 0, & \text{if } \|\tilde{\theta}_i - \theta_i^*\|_2 \leq \frac{\nu_*}{2} \\ 2, & \text{if } \|\tilde{\theta}_i + \theta_i^*\|_2 \leq \frac{\nu_*}{2} \end{cases}.$$

Therefore, the true reward function $f_{\Theta^*}(x)$ in (3) is equivalent to

$$f_{\theta^\ddagger}(x^\ddagger) := x^\ddagger(x, \tilde{\Theta})^\top \theta^\ddagger(\Theta^*, \tilde{\Theta}), \tag{8}$$

where $x^\ddagger : \mathcal{X} \times \mathbb{R}^{k \times d} \to \mathbb{R}^{2kd}$ and $\theta^\ddagger : \mathbb{R}^{k \times d} \times \mathbb{R}^{k \times d} \to \mathbb{R}^{2kd}$ are two mappings with

$$x^\ddagger(x, \tilde{\Theta}) = \begin{bmatrix} \mathbb{1}(\tilde{\theta}_1^\top x \geq 0)x \\ \mathbb{1}(\tilde{\theta}_2^\top x \geq 0)x \\ \vdots \\ \mathbb{1}(\tilde{\theta}_k^\top x \geq 0)x \\ (\frac{1}{2} - \mathbb{1}(\tilde{\theta}_1^\top x \geq 0))x \\ (\frac{1}{2} - \mathbb{1}(\tilde{\theta}_2^\top x \geq 0))x \\ \vdots \\ (\frac{1}{2} - \mathbb{1}(\tilde{\theta}_k^\top x \geq 0))x \end{bmatrix}, \tag{9}$$

and

$$\theta^\ddagger(\Theta^*, \tilde{\Theta}) = \begin{bmatrix} \theta_1^* \\ \theta_2^* \\ \vdots \\ \theta_k^* \\ 2\theta_1^* \mathbb{1}(\|\tilde{\theta}_1 + \theta_1^*\|_2 \leq \frac{\nu_*}{2}) \\ 2\theta_2^* \mathbb{1}(\|\tilde{\theta}_2 + \theta_2^*\|_2 \leq \frac{\nu_*}{2}) \\ \vdots \\ 2\theta_k^* \mathbb{1}(\|\tilde{\theta}_k + \theta_k^*\|_2 \leq \frac{\nu_*}{2}) \end{bmatrix}. \tag{10}$$

In the following, we will use $x^\ddagger$ and $\theta^\ddagger$ to denote the two vectors in (9) and (10) for simplicity whenever no ambiguity is raised. The reward function in (8) has additional k linear components compared to (5); $x^\ddagger$ builds upon $x^\dagger$ and contains additional k features that captures the misspecification due to sign flip. Now, we will be able to run a linear bandit algorithm with $2kd$ features to learn $\theta^\ddagger$ in an online manner.

4.2. OFU-ReLU Algorithm

For now, we assume ν_* is known throughout our design; we describe how to remove this assumption in §4.3.

Our algorithm, which we name OFU-ReLU[2], has two phases:

[2] "OFU" standards for optimism in the face of uncertainty (Abbasi-Yadkori et al., 2011)

Algorithm 1 OFU-ReLU

```
**Input:** exploration length $t_0$, regularization parameter $\lambda$
Initialize $Z_0 \leftarrow \varnothing$
**for** $t \in [t_0]$ **do**
    Sample action $x_t \sim_{\text{i.i.d.}} p$
    Take action $x_t$ and obtain reward $y_t$
    $Z_t \leftarrow Z_{t-1} \cup \{(x_t, y_t)\}$
**end for**
Compute $\tilde{\Theta}_{t_0} \leftarrow \arg\min_{\Theta} \hat{L}_S(\Theta; Z_{t_0})$
**for** $t \in (t_0 + 1, T]$ **do**
    Compute confidence ellipsoid $C_t(\lambda, Z_t)$ for $\theta^\ddagger$
    $x_t^\ddagger \leftarrow \arg\max_{(x,\theta)\in\mathcal{X}^\ddagger(\tilde{\Theta}_{t_0},\nu^*/2)\times C_t(\lambda,Z_t)} x^\top\theta$
    Play $x_t$ with $x^\ddagger(x_t, \tilde{\Theta}_{t_0}) = x_t^\ddagger$ and obtain reward $y_t$
    $Z_t \leftarrow Z_{t-1} \cup \{(x_t, y_t)\}$
**end for**
```

- **Exploration.** Randomly sample exploratory actions $x_t \sim p$ for t_0 time steps until our estimate $\tilde{\Theta}_{t_0}$ (i.e., $\tilde{\Theta}$ estimated using the first t_0 samples) satisfies

  $$\min\{\|\tilde{\theta}_{t_0,i} - \theta_i^*\|_2, \|\tilde{\theta}_{t_0,i} + \theta_i^*\|_2\} \le \nu_*/2$$

  with high probability.

- **OFUL.** Run the OFUL algorithm (Abbasi-Yadkori et al., 2011) to learn the true reward function $f_{\theta^\ddagger}(x^\ddagger)$ in (8), which is linear in the parameter $\theta^\ddagger$ and features $x^\ddagger(x, \tilde{\Theta}_{t_0})$, over the region $x^\ddagger \in \mathcal{X}^\ddagger(\tilde{\Theta}_{t_0}, \nu^*/2)$, where

  $$\mathcal{X}^\ddagger(\Theta, \nu) := \{x^\ddagger(x, \Theta) \mid x \in \mathcal{X}(\Theta, \nu)\}. \tag{11}$$

  At each time period $t > t_0$, we follow OFUL, choosing arm $x_t^\ddagger = \arg\max_{(x,\theta)\in\mathcal{X}^\ddagger(\tilde{\Theta}_{t_0},\nu^*/2)\times C_t(\lambda,Z_t)} x^\top\theta$ and observing reward y_t; the confidence ellipsoid $C_t(\lambda, Z_t)$ for the true parameter $\theta^\ddagger$ depends on a regularization hyperparameter λ from OFUL, and can be computed using Theorem 2 in (Abbasi-Yadkori et al., 2011) with $S = \sqrt{5k}$ (note that $\|\theta^\ddagger\| \le \sqrt{5k}$) and all the data previously observed $Z_t = \{(x_\tau, y_\tau)\}_{\tau\in[t-1]}$.

We detail our algorithm in Algorithm 1. Given our design, we can obtain a regret bound scaling as $\tilde{O}(kd\sqrt{T})$; particularly, we control the parameter estimation error of $\tilde{\Theta}_{t_0}$ using Theorem 3.2 in §3, and analyze the regret of OFUL stage using Theorem 3 in (Abbasi-Yadkori et al., 2011). We provide a proof sketch below.

Theorem 4.3. *The cumulative regret of Algorithm 1 satisfies*

$$R_T = \tilde{O}\left(k^{14}d^3(1/\nu_*^8 \vee d^4) + kd\sqrt{T}\right).$$

Proof Sketch. Suppose the exploration stage ends at time t_0. We have

$$\min\left\{\|\tilde{\theta}_i - \theta_i^*\|_2, \|\tilde{\theta}_i + \theta_i^*\|_2\right\} \le \nu_*/2, \quad \forall i \in [k]$$

with probability at least $1-\delta/2$ by choosing t_0 large enough according to Theorem 3.2. In particular, it suffices to take $\delta = 1/\sqrt{T}$ and

$$t_0 = \tilde{\Omega}(k^{13}d^3(1/\nu_*^8 \vee d^4)). \tag{12}$$

Now we analyze the regret of our OFU-ReLU algorithm in three cases. First, at each time t, the per-period regret can be bounded trivially by

$$r_t \le (\sum_{i\in[k]} \|\theta_i^*\|_2)(\|x^*\|_2 + \|x_t\|_2) \le 2k.$$

Therefore, the regret during the exploration phase is upper bounded by $\tilde{O}(k^{14}d^3(1/\nu_*^8 \vee d^4))$.

In the second stage, we run OFUL to find the optimal policy for the linear function $f_{\theta^\ddagger}(x^\ddagger)$ given our forced-sample estimate $\tilde{\Theta}_{t_0}$. Applying the regret bound in Theorem 3 of (Abbasi-Yadkori et al., 2011) gives the regret bound in our second stage to be $\tilde{O}(kd\sqrt{T})$ with at least a probability of $1 - \delta/2$.

Finally, with a small probability $\delta = 1/\sqrt{T}$, we would have linear regret scaling as $2kT$; thus, the expected regret in this case is bounded by $2kT\delta = O(k\sqrt{T})$. Our claim then follows. □

We provide a detailed proof in Appendix D. Theorem 4.3 shows that our algorithm obtains a $\tilde{O}(kd\sqrt{T})$ regret guarantee as long as the time horizon T is large enough.

4.3. OFU-ReLU+ Algorithm

Algorithm 1 requires the knowledge of the gap ν_*, which is typically unknown. We can remove this assumption based on a batching strategy; we provide a schematic representation of our algorithm in Figure 1. Algorithm 2 summarizes our algorithm OFU-ReLU+ based on this insight.

At a high level, we split the entire time horizon T into M increasing batches with a grid $0 = T_0 \le T_1 \le \cdots \le T_M = T$. Each batch $i \in [M-1]$, i.e., $t \in (T_{i-1}, T_i]$, satisfies $a(T_i - T_{i-1}) = T_{i+1} - T_i$—i.e., $T_i = (a^i - 1)T_1$ for some constant $a > 1$ and $T_1 > 1$. Note that

$$M = \left\lceil \frac{\log(T/T_1 + 1)}{\log(a)} \right\rceil. \tag{13}$$

For each batch, we take a fixed guess of ν_*; we reduce this guess geometrically from one batch to the next. Specifically, let ν_0 be our initial guess of ν_* at $t = 0$; then, the guess ν_i for batch i is $\nu_i = \nu_0/b^i$ for some constant $b > 1$. Our ν_i will become sufficiently small that $\nu_i \le \nu_*$ for batch $i > \log(\nu_0/\nu_*)/\log(b)$. Our regret analysis in §4.2 can then be applied from that batch onwards.

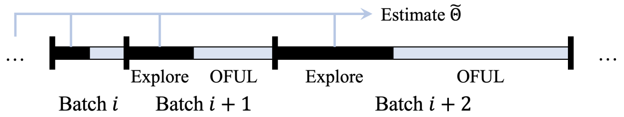


Figure 1: Schematic representation of OFU-ReLU+.

In detail, within each batch i, we have an exploration phase and a following OFUL phase as in §4.2. We initialize the batch with the exploration period as $t \in (T_{i-1}, T_{i-1} + t_{0,i}]$. Different from the conventional batching strategies (see, e.g., Golrezaei et al., 2019; Luo et al., 2022), we estimate $\tilde{\Theta}$ and run OFUL using samples from all previous batches without discarding data from previous batches. Particularly, we estimate $\tilde{\Theta}$ based on the accumulative samples from all the previous exploration phases $t \in \cup_{j\in[i]}(T_{j-1}, T_{j-1} + t_{0,j}]$ so that the estimation error satisfies

$$\min\{\|\tilde{\theta}_i - \theta_i^*\|_2, \|\tilde{\theta}_i + \theta_i^*\|_2\} \leq \nu_i/2, \quad \forall i \in [k]. \quad (14)$$

Define a function t_0[3] of an arbitrary ν as

$$t_0(\nu) = \tilde{\Theta}\left(k^{13}d^3(1/\nu^8 \vee d^4)\right) \quad (15)$$

according to (12) (detailed expression in Appendix D). Then, it suffices to have $t_{0,i} = t_0(\nu_i) - t_0(\nu_{i-1})$ according to (12). Let our estimate be $\tilde{\Theta}_{t_0(\nu_i)}$ (i.e., $\tilde{\Theta}$ estimated using all $t_0(\nu_i)$ random samples from previous batches).

In the rest of the batch i, we run OFUL; at each time step $t \in (T_{i-1} + t_{0,i}, T_i]$, we choose arm $x_t^\ddagger = \arg\max_{(x,\theta)\in\mathcal{X}^\ddagger(\tilde{\Theta}_{t_0(\nu_i)}, \nu_i/2)\times C_t(\lambda, Z_t)} x^\top\theta$ and observing reward y_t. Again, we compute the $C_t(\lambda, Z_t)$ based on (Abbasi-Yadkori et al., 2011) with $S = \sqrt{5k}$ and all the data observed $Z_t = \{(x_\tau, y_\tau)\}_{\tau\in[t-1]}$. The features are $x^\ddagger(x_\tau, \tilde{\Theta}_{t_0(\nu_i)})$ and the parameter to estimate is $\theta^\ddagger(\Theta^*, \tilde{\Theta}_{t_0(\nu_i)})$. Note that the features and the parameter are invariant of $\tilde{\Theta}_{t_0(\nu_i)}$ once $\nu_i \leq \nu_*$; thus, our estimate becomes consistent and OFUL is valid from then on.

To bound the regret, we decompose the whole time horizon into three parts and analyze them respectively:

(i). All batch i satisfying $\nu_i > \nu_*$,

(ii). $t \in (T_{i-1}, T_{i-1} + t_{0,i}]$ for all batch i with $\nu_i \leq \nu_*$,

(iii). $t \in (T_{i-1} + t_{0,i}, T_i]$ for all batch i with $\nu_i \leq \nu_*$.

The regret in (i) is independent of T and is based on our choices of a, b, T_1 and ν_0. Our analysis in §4.2 can be applied similarly to analyze the exploration regret in (ii) and the regret (iii) from OFUL. However, for (ii), since ν_i decreases over time, the regret per batch grows over time.

Algorithm 2 OFU-ReLU+

```
Input: regularization parameter $\lambda$, parameters $\nu_0, T_1, a, b$
Initialize $Z_0 \leftarrow \varnothing$, $E_0 \leftarrow \varnothing$
for $i \in [M]$ do
  $\nu_i \leftarrow \nu_{i-1}/b$, $t_{0,i} \leftarrow t_0(\nu_i) - t_0(\nu_{i-1})$
  for $t \in (T_{i-1}, T_{i-1} + t_{0,i}]$ do
    Sample action $x_t \sim_{\text{i.i.d.}} p$
    Take action $x_t$ and obtain reward $y_t$
    $E_t \leftarrow E_{t-1} \cup \{(x_t, y_t)\}$, $Z_t \leftarrow Z_{t-1} \cup \{(x_t, y_t)\}$
  end for
  Compute $\tilde{\Theta}_{t_0(\nu_i)} \leftarrow \arg\min_\Theta \hat{L}_S(\Theta; E_{T_i + t_{0,i}})$
  for $t \in (T_{i-1} + t_{0,i}, T_i]$ do
    Compute confidence ellipsoid $C_t(\lambda, Z_t)$ for $\theta^\ddagger$
    $x_t^\ddagger \leftarrow \arg\max_{(x,\theta)\in\mathcal{X}^\ddagger(\tilde{\Theta}_{t_0(\nu_i)}, \nu_i/2)\times C_t(\lambda, Z_t)} x^\top\theta$
    Play $x_t$ with $x^\ddagger(x_t, \tilde{\Theta}_{t_0(\nu_i)}) = x_t^\ddagger$ and obtain $y_t$
    $Z_t \leftarrow Z_{t-1} \cup \{(x_t, y_t)\}$
  end for
  $T_{i+1} \leftarrow (a^{i+1} - 1)T_1$
end for
```

We show in the following theorem that this regret scales as a polynomial term of T of which the order depends on the relative size between a and b, and can be chosen to be asymptotically less than $O(\sqrt{T})$.

Theorem 4.4. *The cumulative regret of Algorithm 2 has*

$$R_T = \tilde{O}\left(k^{14}d^7 + k^{14}d^3T^{8\frac{\log(b)}{\log(a)}} + kd\sqrt{T}\right).$$

Proof Sketch. We bound the regret for the three cases above respectively. First, in case (i), we have $i \leq \log(\nu_0/\nu_*)/\log(b)$. Recall that the per-period regret can be trivially bounded by $2k$. Thus, the regret in this case is upper bounded by

$$2k(a^{\log(\nu_0/\nu_*)/\log(b)} - 1)T_1 \leq 2k(\nu_0/\nu_*)^{\frac{\log(a)}{\log(b)}}T_1.$$

Similarly, regret in case (ii) can also be bounded by

$$2k\sum_{i=\lceil\frac{\log(\nu_0/\nu_*)}{\log(b)}\rceil}^{M-1} t_{0,i} = \tilde{O}\left(k^{14}d^7 + k^{14}d^3T^{8\frac{\log(b)}{\log(a)}}\right),$$

where we use the definition of $t_{0,i}$, that of $t_0(\nu)$ in (15), and the value of M in (13).

Next, we calculate the regret of running OFUL in case (iii), which is upper bounded again by $\tilde{O}(kd\sqrt{T})$ following the proof strategy of Theorem 3 in (Abbasi-Yadkori et al., 2011), similar to our proof for OFU-ReLU.

Finally, with a union bound, there's a small probability $M/\sqrt{T}$ that we will have a linear regret since the above analysis holds only with high probability. We can show the

[3] Note that t_0 in §4.2 is a function of ν_*. Here we abuse the notation t_0 slightly.

regret of this part scales as $\tilde{O}(k\sqrt{T})$. Combining all the above gives our final result. □

We provide a detailed proof in Appendix E. Compared with Theorem 4.3, here we gain an additional $T^{8\frac{\log(b)}{\log(a)}}$ dependence due to the increasing difficulty of learning the unknown gap ν_*. Theorem 4.4 implies that, as long as our choices of the multipliers a and b satisfy $8\log(b)/\log(a) \leq 1/2$, we recover a $\tilde{O}(\sqrt{T})$ regret dependence—i.e., when the length of exploration period grows slower than the batch time horizon. for instance, taking $a = 2$ and $b = 2^{1/32}$, we obtain

$$R_T = \tilde{O}\left(k^{14}d^7 + k^{14}d^3T^{1/4} + kd\sqrt{T}\right).$$

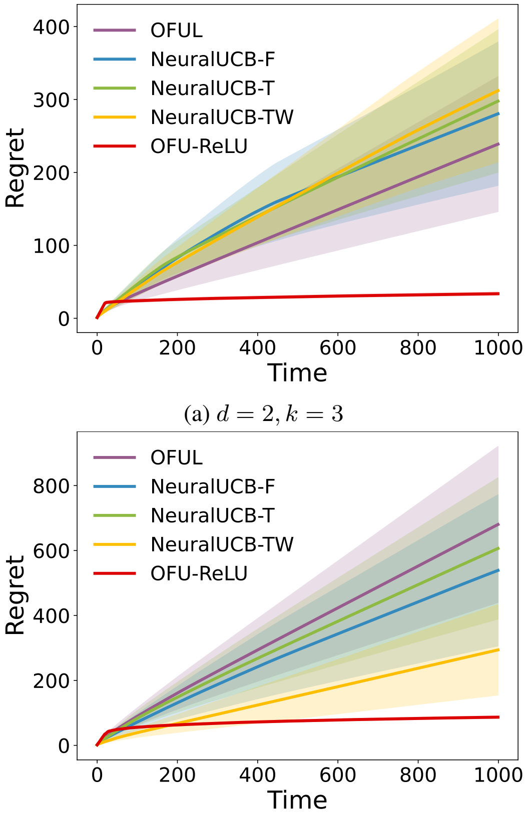
Finally, if the time horizon T is at least a polynomial term of k and d, then as before, we recover an $\tilde{O}(kd\sqrt{T})$ regret guarantee.

5. Experiments

We compare our algorithm OFU-ReLU with several benchmarks, including OFUL (Abbasi-Yadkori et al., 2011), which assumes the true model is linear and introduces misspecification errors, and three different versions of NeuralUCB (Zhou et al., 2020), i.e., NeuralUCB-F, NeuralUCB-T and NeuralUCB-TW. Particularly, NeuralUCB-F follows the setup in §7.1 of (Zhou et al., 2020) with $m = 20$ neurons and two layers; NeuralUCB-T assumes the knowledge of the neural network structure of the true reward, i.e., $m = k$ neurons and one layer; and NeuralUCB-TW inherits the structure from NeuralUCB-T but expands the layer size into $m = 2k$.

We consider the true model of a ReLU structure as in (3), with multiple settings presented in Figure 2. The parameter Θ^* is randomly sampled from the sphere $\|\theta_i\| = 1$ for $i \in [k]$. The noise follows a normal distribution $N(0, 0.01)$. We randomly draw $1,000$ arms from the unit sphere in each round t and choose an optimal arm from this arm set. Note that with a discretized arm set, our claim of a nontrivial gap ν_* always holds. For OFUL and OFU-ReLU, we use the theoretically suggested confidence ellipsoid for UCB. Since we do not know the gap ν_*, we set the length of exploration phase for OFU-ReLU to be 20 for our method. We tune the hyperparameters λ for all the methods.

Figure 2 shows the performance of our bandit algorithm versus the other benchmarks with a 95% confidence interval. We find our algorithm significantly outperforms all the other benchmarks. OFUL assumes a linear reward model structure and thus incurs large regret due to the misspecification error. All NeuralUCB benchmarks use gradient descent to learn the model structure over time and thus take long time to converge in general. Note that even as a fair comparison with NeuralUCB-T, where the true network structure is given, our method is still significantly better in terms of regret. It is worth noting that our method takes only 20 time steps to converges in a time horizon of $T = 1,000$, while NeuralUCB algorithms generally take a long time to converge (e.g., (Zhou et al., 2020) consider a longer horizon $T \approx 10,000$ in all their experiments). Our empirical results complement our theoretical analysis and suggest the efficiency of our algorithm in practice, especially in a short time horizon, despite a theoretically long exploration phase due to our parameter estimation error bound.



(b) $d = 2, k = 10$

Figure 2: Cumulative regret of a time horizon $T = 1,000$ over 50 trials with 95% confidence interval.

6. Conclusion

We analyze a bandit problem with the reward given by one-layer ReLU neural network structure, and propose algorithms that can provide a regret bound of $\tilde{O}(\sqrt{T})$. To the best of our knowledge, our work is the first to obtain such regret guarantee for bandit learning with neural networks (without an effective dimension assumption). Furthermore, we demonstrate the efficiency of our algorithm in a synthetic experiment, which suggests its practical potentials. We believe both our theoretical and empirical results provide the first insight into an efficient design of bandit algorithms based on ReLU neural network structures.

We conclude by providing directions for future research. Due to the complexity of the problem, we tailor our focus to the one-layer ReLU activations. A natural extension is to generalize our result to piecewise linear activation functions. It is more challenging to explore whether our insight can be generalized to bandit problems with more complex activation functions or multiple-layer architectures.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pp. 2312–2320, 2011.

Bastani, H. and Bayati, M. Online decision making with high-dimensional covariates. *Operations Research*, 68 (1):276–294, 2020.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. 2008.

Dong, K., Yang, J., and Ma, T. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. *Advances in Neural Information Processing Systems*, 34:26168–26182, 2021.

Du, S. and Lee, J. On the power of over-parametrization in neural networks with quadratic activation. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2018.

Golrezaei, N., Javanmard, A., and Mirrokni, V. Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Advances in Neural Information Processing Systems*, 32, 2019.

Gu, Q., Karbasi, A., Khosravi, K., Mirrokni, V., and Zhou, D. Batched neural bandits. *ACM/JMS Journal of Data Science*, 1(1):1–18, 2024.

Huang, B., Huang, K., Kakade, S., Lee, J. D., Lei, Q., Wang, R., and Yang, J. Optimal gradient-based algorithms for non-concave bandit optimization. *Advances in Neural Information Processing Systems*, 34:29101–29115, 2021.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Kassraie, P. and Krause, A. Neural contextual bandits without regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 240–278. PMLR, 2022.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Luo, Y., Sun, W. W., and Liu, Y. Contextual dynamic pricing with unknown noise: Explore-then-ucb strategy and improved regrets. *Advances in Neural Information Processing Systems*, 35:37445–37457, 2022.

Rajaraman, N., Han, Y., Jiao, J., and Ramchandran, K. Beyond ucb: Statistical complexity and optimal algorithms for non-linear ridge bandits. *arXiv preprint arXiv:2302.06025*, 2023.

Rigollet, P. and Hütter, J.-C. High dimensional statistics. *Lecture notes for course 18S997*, 813(814):46, 2015.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35 (2):395–411, 2010.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Salgia, S. Provably and practically efficient neural contextual bandits. In *International Conference on Machine Learning*, pp. 29800–29844. PMLR, 2023.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. Finite-time analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, 2013.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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

Xu, K., Bastani, H., and Bastani, O. Robust generalization of quadratic neural networks via function identification. *arXiv preprint arXiv:2109.10935*, 2021a.

Xu, K., Zhao, X., Bastani, H., and Bastani, O. Group-sparse matrix factorization for transfer learning of word embeddings. *arXiv preprint arXiv:2104.08928*, 2021b.

Xu, P., Wen, Z., Zhao, H., and Gu, Q. Neural contextual bandits with deep representation and shallow exploration. *arXiv preprint arXiv:2012.01780*, 2020.

Zhang, W., Zhou, D., Li, L., and Gu, Q. Neural thompson sampling. *arXiv preprint arXiv:2010.00827*, 2020.

Zhang, X., Yu, Y., Wang, L., and Gu, Q. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1524–1534. PMLR, 2019.

Zhou, D., Li, L., and Gu, Q. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pp. 11492–11502. PMLR, 2020.

A. Proof of Proposition 3.1

We give the proof of Proposition 3.1, followed by proofs of the lemmas used in this proof.

A.1. Intuition

We illustrate our proof strategy in Figure 3. We first define the following set

$$\mathcal{J}_i^\alpha = \{l \in [k] \mid \|\theta_l - \theta_i^*\|_2 \leq \alpha, \text{ or } \|\theta_l + \theta_i^*\|_2 \leq \alpha\}$$

for all $i \in [k]$. It suffices to prove that $\mathcal{J}_i^\alpha$ is a singleton set for every $i \in [k]$ in order to prove Proposition 3.1. Note that when $\alpha \leq \alpha_0/2$, $\mathcal{J}_i^\alpha$'s are disjoint (i.e., $\mathcal{J}_i^\alpha \cap \mathcal{J}_{i'}^\alpha = \varnothing$ for any $i, i' \in [k], i \neq i'$); in particular, if $i'' \in \mathcal{J}_i^\alpha$, then for any other $i' \in [k]$, we have

$$\|\theta_{i''} \pm \theta_{i'}^*\|_2 \geq \|\theta_i^* - \theta_{i'}^*\|_2 - \|\theta_{i''} - \theta_i^*\|_2 \geq \alpha_0 - \alpha \geq \alpha,$$

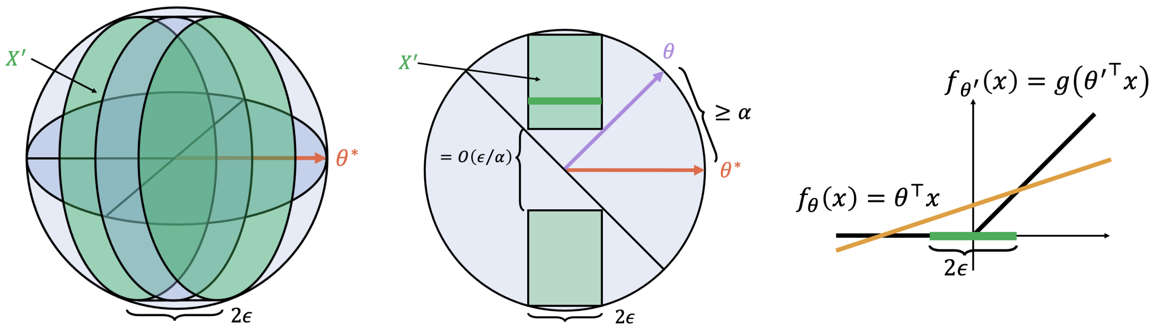
that is, $i'' \notin J_{i'}^\alpha$, as claimed, where we use Assumption 2.2. As a consequence, it suffices to show that $\mathcal{J}_i^\alpha \neq \varnothing$ for every $i \in [k]$. To this end, we prove its contrapositive—i.e., if there exists $j \in [k]$ such that $\mathcal{J}_j^\alpha = \varnothing$, then

$$L_{X_j}(\Theta) \geq \eta = \frac{\epsilon^2(1 - d\epsilon^2/2)|S^{d-2}|}{8} - \frac{k\epsilon^3|S^{d-3}|}{2\alpha} - (6k + 3\sigma)d\epsilon^3|S^{d-2}|.$$

Intuitively, if θ_j^* does not have a matching neuron θ_j (up to a sign flip) in $\mathcal{J}_j^\alpha$, then we can show that $g(\theta^\top x)$ is linear for any

$$\theta \in \bar{\Theta}_{\neg j}^* := \{\theta_i\}_{i \in [k]} \cup \{\theta_i^*\}_{i \in [k], i \neq j}$$

except θ_j^* on majority of the strip X_j' (Figure 3 (b), formally defined in (17)), which is a close approximation of X_j. Therefore, $f_\Theta(x) - f_{\Theta^*}(x)$ can be additively decomposed into a linear term plus $g(\theta_j^{*\top} x)$. Besides, we prove that any linear function cannot approximate $g(\theta_j^{*\top} x)$ well on X_j' (Figure 3 (c)). As a result, given the definition of $L_{X_j}(\Theta)$, we can show that $L_{X_j}(\Theta)$ is lower-bounded.



(a) (b) (c)

Figure 3: Illustrations for proof sketch of the estimation error for ReLU Neural Networks. (a) The region X' is the cylinder with caps consisting of the two green circles and radius 2ϵ. (b) Projected version of subfigure (a). (c) The green region is X' with a section of length $O(\epsilon/\alpha)$ cut out.

A.2. Proof of Proposition 3.1

We list the details of our proof strategy in this section. We introduce additional notation that we will use in the following proof. With a slight abuse of notation, for any vector $x \in \mathbb{R}^d$ Let x_i denote the i^{th} element of x and $x_{i:j} \in \mathbb{R}^{j-i+1}$ the subvector of x consisting of the i^{th} to j^{th} elements.

Step 1. Note that X_j is a slice of the sphere $\mathcal{X} = S^{d-1}$. To simplify our following analysis, we approximate X_j using a cylinder $X'_j \subseteq \mathbb{R}^d$—i.e., without loss of generality assuming that for one specific $j \in [k]$

$$\theta^*_j = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^\top , \tag{16}$$

then

$$X'_j = [-\epsilon, \epsilon] \times Z = \{x \in \mathbb{R}^d \mid |x_1| \le \epsilon, x_{2:d} \in Z\}, \quad \text{where} \quad Z = S^{d-2}\left(\sqrt{1-\epsilon^2}\right),$$

or equivalently

$$X'_j = \{\phi(x) \mid x \in X_j\}, \quad \text{where} \quad \phi(x) = \begin{bmatrix} x_1 & \sqrt{\frac{1-\epsilon^2}{1-x_1^2}} x_2 & \cdots & \sqrt{\frac{1-\epsilon^2}{1-x_1^2}} x_d \end{bmatrix}^\top . \tag{17}$$

This region is visualized in Figure 3 (a); its projection to two dimensions is shown in Figure 3 (b). We show that the loss restricted to X'_j is approximately equal to the loss restricted to X_j:

Lemma A.1. *Given X'_j, it holds that*

$$\left| \int_{X_j} |f_\Theta(x) - f_{\Theta^*}(x)| dx - \int_{X'_j} |f_\Theta(x) - f_{\Theta^*}(x)| dx \right| \le 6kd\epsilon^3 |S^{d-2}|.$$

The proof is provided in Appendix A.3.

Step 2. Next, we decompose X'_j into strips—i.e.,

$$X'_j = \bigcup_{z \in Z} X^z, \quad \text{where} \quad X^z = [-\epsilon, \epsilon] \times \{z\}.$$

Note that X^z is a one-dimensional manifold; one such strip is shown as the horizontal green line in Figure 3 (b). Our strategy is to lower bound the loss restricted to each of these strips, after which we can integrate over them to obtain a lower bound on the overall loss. With a slight abuse of notation, we further define

$$f_{\Theta_{\neg j}} = \sum_{i \in [k], i \ne j} g(\theta_i^\top x).$$

Precisely, we provide a lower bound of the loss for those $z \in Z$ where f_Θ and $f_{\Theta^*_{\neg j}}$ is linear on X^z; in particular, for such z, we have

$$X^z \cap \mathcal{V}^{\bar{\Theta}^*_{\neg j}} = \varnothing, \quad \text{where} \quad \mathcal{V}^{\bar{\Theta}^*_{\neg j}} = \bigcup_{\theta \in \bar{\Theta}^*_{\neg j}} \{x \in X'_j \mid \theta^\top x = 0\}.$$

Note that $\mathcal{V}^{\bar{\Theta}^*_{\neg j}}$ is the boundary at which one of the ReLUs, i.e., $g(\theta^\top x)$ with $\theta \in \bar{\Theta}^*_{\neg j}$, transitions from inactive to active. If X^z does not intersect $\mathcal{V}^{\bar{\Theta}^*_{\neg j}}$, then $\theta^\top x \ne 0$ on X^z for all $\theta \in \bar{\Theta}^*_{\neg j}$ and, hence, $f_\Theta - f_{\Theta^*_{\neg j}}$ must be linear on such X^z. In the following, we show that such z's make up a large proportion of Z; equivalently, we show that the following subset is small:

$$Z^{\bar{\Theta}^*_{\neg j}} = \bigcup_{\theta \in \bar{\Theta}^*_{\neg j}} Z^\theta, \quad \text{where} \quad Z^\theta = \{z \in Z \mid \exists x_1 \in [-\epsilon, \epsilon], \theta^\top (x_1 \circ z) = 0\}, \tag{18}$$

where $x_1 \circ z := \begin{bmatrix} x_1 & z_1 & \cdots & z_{d-1} \end{bmatrix}^\top$.

Lemma A.2. *For any $\theta \in \bar{\Theta}^*_{\neg j}$, we have*

$$|Z^\theta| \le \frac{2\epsilon}{\alpha} |S^{d-3}|.$$

The proof is given in Appendix A.4. This result is illustrated in Figure 3 (b); the set of X^z for which $z \in Z^{\bar{\Theta}^*_{\neg j}}$ (which has size $O(\epsilon/\alpha)$) has been removed from X'_j. Note that as α becomes larger, the size of $Z^{\bar{\Theta}^*_{\neg j}}$ becomes smaller.

Step 3. Next, for $z \in Z \setminus Z^{\bar{\Theta}^*_{\neg j}}$, we lower bound the loss on X^z. Remember the loss is

$$|f_\Theta(x) - f_{\Theta^*}(x)| = |(f_\Theta(x) - f_{\Theta^*_{\neg j}}(x)) - g(\theta_j^{*\top} x)|,$$

where we argue in Step 2 that on such X_z the first term is linear. Therefore, we can lower bound the loss using the following lemma:

Lemma A.3. *For any $\beta_0, \beta_1 \in \mathbb{R}$, we have*

$$\int_{-\epsilon}^{\epsilon} |(\beta_0 + \beta_1 w) - g(w)| dw \geq \frac{\epsilon^2}{8}.$$

We provide the proof in Appendix A.5. Since our loss is the mean absolute error, this result follows from a geometric argument. Intuitively, as illustrated in Figure 3 (c), there is a triangular gap between $f_\Theta - f_{\Theta^*_{\neg j}}$ (which is linear—i.e., $f_\Theta(x) - f_{\Theta^*_{\neg j}}(x) = \beta^\top x$ for some $\beta \in \mathbb{R}^d$) and $g(\theta_j^{*\top} x)$ on X^z; this gap equals the loss, and it cannot be reduced regardless of the value of β.

Step 4. Finally, the proof of Proposition 3.1 consists of integrating the lower bound from Step 3 over $z \in Z \setminus Z^{\bar{\Theta}^*_{\neg j}}$ to obtain a lower bound on $L_{X_j}(\Theta)$.

Proof. First, note that

$$\begin{aligned}
\int_{X_j'} |f_\Theta(x) - f_{\Theta^*}(x)| dx &\geq \int_Z \int_{-\epsilon}^{\epsilon} |f_\Theta(x_1 \circ z) - f_{\Theta^*}(x_1 \circ z)| dx_1 dz \\
&\geq \int_{Z \setminus Z^{\bar{\Theta}^*_{\neg j}}} \int_{-\epsilon}^{\epsilon} |f_\Theta(x_1 \circ z) - f_{\Theta^*}(x_1 \circ z)| dx_1 dz \\
&= \int_{Z \setminus Z^{\bar{\Theta}^*_{\neg j}}} \int_{-\epsilon}^{\epsilon} |(f_\Theta(x_1 \circ z) - f_{\Theta^*_{\neg j}}(x_1 \circ z)) - g(\theta_j^{*\top}(x_1 \circ z))| dx_1 dz. \qquad (19)
\end{aligned}$$

Remember for any given $z \in Z \setminus Z^{\bar{\Theta}^*_{\neg j}}$, the first term $f_\Theta(x_1 \circ z) - f_{\Theta^*_{\neg j}}(x_1 \circ z)$ is linear in $x_1 \circ z$, i.e., there exists a parameter $\tilde{\theta}_j$ such that $\tilde{\theta}_j^\top (x_1 \circ z) = f_\Theta(x_1 \circ z) - f_{\Theta^*_{\neg j}}(x_1 \circ z)$. Without loss of generality, we can modify the coordinate system so that

$$\tilde{\theta}_j = \begin{bmatrix} t_1 & t_2 & 0 & \cdots & 0 \end{bmatrix}^\top$$

without affecting θ_j^* in (16). Then, we have

$$\begin{aligned}
\int_{-\epsilon}^{\epsilon} |(f_\Theta(x_1 \circ z) - f_{\Theta^*_{\neg j}}(x_1 \circ z)) - g(\theta_j^{*\top}(x_1 \circ z))| dx_1 &= \int_{-\epsilon}^{\epsilon} |\tilde{\theta}_j^\top (x_1 \circ z) - g(\theta_j^{*\top}(x_1 \circ z))| dx_1 \\
&= \int_{-\epsilon}^{\epsilon} |(t_1 x_1 + t_2 z_1) - g(x_1)| dx_1 \\
&\geq \frac{\epsilon^2}{8},
\end{aligned}$$

where the last inequality uses Lemma A.3. Given the above result, we can derive from (19) that

$$\int_{X_j'} |f_\Theta(x) - f_{\Theta^*}(x)| dx \geq |Z \setminus Z^{\Theta^*_{\neg j}}| \frac{\epsilon^2}{8} = \left((1 - \epsilon^2)^{\frac{d-2}{2}} |S^{d-2}| - \frac{4k\epsilon |S^{d-3}|}{\alpha} \right) \frac{\epsilon^2}{8}.$$

This combined with Lemma A.1 gives that

$$\int_{X_j} |f_\Theta(x) - f_{\Theta^*}(x)| dx \geq \left((1 - \frac{d}{2}\epsilon^2) |S^{d-2}| - \frac{4k\epsilon |S^{d-3}|}{\alpha} \right) \frac{\epsilon^2}{8} - 6kd\epsilon^3 |S^{d-2}|.$$

The result then follows. □

A.3. Proof of Lemma A.1

First, note that

$$\begin{aligned}\int_{X_j'} |f_\Theta(x) - f_{\Theta^*}(x)| dx &= \int_{X_j} |f_\Theta(\phi(x)) - f_{\Theta^*}(\phi(x))| \cdot |\det \nabla_x \phi(x)| dx \\ &= \int_{X_j} |f_\Theta(\phi(x)) - f_{\Theta^*}(\phi(x))| \left(\frac{1-\epsilon^2}{1-x_1^2}\right)^{\frac{d-1}{2}} dx,\end{aligned}$$

where the second equality follows the fact that $\nabla_x \phi(x)$ is a lower triangular matrix. Then, we have

$$\begin{aligned}&\int_{X_j'} |f_\Theta(x) - f_{\Theta^*}(x)| dx - \int_{X_j} |f_\Theta(x) - f_{\Theta^*}(x)| dx \\ &= \int_{X_j} \left(\frac{1-\epsilon^2}{1-x_1^2}\right)^{\frac{d-1}{2}} |f_\Theta(\phi(x)) - f_{\Theta^*}(\phi(x))| - |f_\Theta(x) - f_{\Theta^*}(x)| dx \\ &\leq \int_{X_j} |(f_\Theta(x) - f_\Theta(\phi(x))) - (f_{\Theta^*}(x) - f_{\Theta^*}(\phi(x)))| dx \\ &\leq \int_{X_j} |f_\Theta(x) - f_\Theta(\phi(x))| + |f_{\Theta^*}(x) - f_{\Theta^*}(\phi(x))| dx,\end{aligned}$$

and

$$\begin{aligned}&\int_{X_j} |f_\Theta(x) - f_{\Theta^*}(x)| dx - \int_{X_j'} |f_\Theta(x) - f_{\Theta^*}(x)| dx \\ &= \int_{X_j} |f_\Theta(x) - f_{\Theta^*}(x)| - \left(\frac{1-\epsilon^2}{1-x_1^2}\right)^{\frac{d-1}{2}} |f_\Theta(\phi(x)) - f_{\Theta^*}(\phi(x))| dx \\ &\leq \int_{X_j} |f_\Theta(x) - f_{\Theta^*}(x)| - (1-\epsilon^2)^{\frac{d-1}{2}} |f_\Theta(\phi(x)) - f_{\Theta^*}(\phi(x))| dx \\ &\leq \int_{X_j} |f_\Theta(x) - f_{\Theta^*}(x)| - (1 - \frac{d}{2}\epsilon^2) |f_\Theta(\phi(x)) - f_{\Theta^*}(\phi(x))| dx \\ &\leq \int_{X_j} |f_\Theta(x) - f_\Theta(\phi(x))| + |f_{\Theta^*}(x) - f_{\Theta^*}(\phi(x))| dx \\ &\qquad + \frac{d}{2}\epsilon^2 \int_{X_j} |f_\Theta(\phi(x)) - f_{\Theta^*}(\phi(x))| dx,\end{aligned}$$

where we use $|x_1| \leq \epsilon$ and Bernoulli's inequality. Therefore, we have

$$\begin{aligned}&\left| \int_{X_j} |f_\Theta(x) - f_{\Theta^*}(x)| dx - \int_{X_j'} |f_\Theta(x) - f_{\Theta^*}(x)| dx \right| \\ &\qquad \leq \int_{X_j} |f_\Theta(x) - f_\Theta(\phi(x))| + |f_{\Theta^*}(x) - f_{\Theta^*}(\phi(x))| dx + \frac{d}{2}\epsilon^2 \int_{X_j} |f_\Theta(\phi(x)) - f_{\Theta^*}(\phi(x))| dx.\end{aligned}$$

Note that both f_Θ and f_{Θ^*} are k-Lipschitz. In particular, for any $x, x' \in \mathbb{R}^d$, we have

$$|f_\Theta(x) - f_\Theta(x')| = \left| \sum_{i=1}^k (g(\theta_i^\top x) - g(\theta_i^\top x')) \right| \leq \sum_{i=1}^k |\theta_i^\top (x - x')| \leq k \|x - x'\|_2.$$

The same result holds for f_{Θ^*}. As a result, we can derive

$$\int_{X_j} |f_\Theta(x) - f_\Theta(\phi(x))| + |f_{\Theta^*}(x) - f_{\Theta^*}(\phi(x))| dx \leq 2k|X_j| \max_{x \in X_j} \|x - \phi(x)\|,$$

and

$$\int_{X_j} |f_\Theta(\phi(x)) - f_{\Theta^*}(\phi(x))| dx \leq 2k|X_j| \max_{x \in X_j} \|\phi(x)\|.$$

Note that

$$\max_{x \in X_j} \|x - \phi(x)\|_2 = \max_{x \in X_j} \sqrt{1 - x_1^2} - \sqrt{1-\epsilon^2} \leq \epsilon^2,$$

and

$$\max_{x \in X_j} \|\phi(x)\|_2 \leq 1.$$

Besides, we have

$$|X_j| = \int_{-\epsilon}^{\epsilon} A^{d-2}\left(\sqrt{1-x_1^2}\right) dx_1 \leq 2\epsilon |S^{d-2}|,$$

where $A^n(r)$ is the area of the n-sphere with radius r. Then, the claim follows.

A.4. Proof of Lemma A.2

Consider the set Z^{θ_i} for some $i \in [k]$ first. Without loss of generality, we can modify the coordinate system so that

$$\theta_i = \begin{bmatrix} t_1 & t_2 & 0 & \dots & 0 \end{bmatrix}^\top$$

without affecting θ_j^* in (16). By assumption, we have $\|\theta_i\|_2 = \sqrt{t_1^2 + t_2^2} = 1$. In the following, we first consider $t_1 \geq 0$. Remember

$$\alpha^2 < \|\theta_i - \theta_j^*\|_2^2 = (1-t_1)^2 + t_2^2 = 2(1-t_1) = 2\left(1 - \frac{t_1}{\sqrt{t_1^2+t_2^2}}\right) = 2\left(1 - \frac{1}{\sqrt{1 + t_2^2/t_1^2}}\right).$$

This implies

$$\frac{|t_2|}{|t_1|} > \sqrt{\left(\frac{1}{1-\alpha^2/2}\right)^2 - 1} \geq \sqrt{\left(1 + \frac{\alpha^2}{2}\right)^2 - 1} \geq \alpha$$

For any $z \in Z^{\theta_i}$, the condition $\theta_i^\top (x_1 \circ z) = 0$ is equivalent to

$$t_1 x_1 + t_2 z_1 = 0.$$

As a consequence, we have

$$|z_1| \leq \frac{|t_1| \cdot |x_1|}{|t_2|} \leq \frac{\epsilon}{\alpha}.$$

Therefore, we can obtain that

$$|Z^{\theta_j}| \leq \int_{-\epsilon/\alpha}^{\epsilon/\alpha} A^{d-3}\left(\sqrt{1-\epsilon^2 - z_1^2}\right) dz_1 \leq \frac{2\epsilon}{\alpha}|S^{d-3}|,$$

where $A^n(r)$ is the area of the n-sphere with radius r. The claim for $\theta \in \{\theta_i\}_{i\in[k]}$ then follows. The case $t_1 < 0$ can be analyzed similarly using $\alpha < \|\theta_i + \theta_j^*\|_2$. In addition, the claim for $\theta \in \{\theta_i^*\}_{i\in[k], i\neq j}$ also holds noticing $2\alpha < \alpha_0 < \|\theta_i^* \pm \theta_j^*\|_2$ for $i \neq j$ by Assumption 2.2.

A.5. Proof of Lemma A.3

We first have

$$\begin{aligned}\int_{-\epsilon}^{\epsilon} |(\beta_0 + \beta_1 w) - g(w)| dw &= \int_{-\epsilon}^{0} |\beta_0 + \beta_1 w| dw + \int_{0}^{\epsilon} |\beta_0 + (\beta_1 - 1) w| dw \\ &= \int_{0}^{\epsilon} |\beta_0 - \beta_1 w| dw + \int_{0}^{\epsilon} |\beta_0 + (\beta_1 - 1) w| dw \\ &= F(\beta_0, \beta_1) + F(\beta_0, 1 - \beta_1),\end{aligned}$$

where we define $F(a,b) = \int_0^\epsilon |a - bw| dw$. Note that we must have either $\beta_1 \geq 1/2$ or $1 - \beta_1 \geq 1/2$. Additionally, it holds that $F(-|a|, b) \geq F(|a|, b)$ for $b > 0$. Therefore, it suffices to consider the case $a \geq 0, b > 1/2$ to provide a lower bound. In this case, $F(a,b)$ takes the minimum when $a \leq b\epsilon$. Therefore, we have

$$\begin{aligned}F(a,b) &= \int_0^{a/b} (a - bw) dw + \int_{a/b}^{\epsilon} (bw - a) dw \\ &= \left(\frac{a^2}{b} - \frac{a^2}{2b}\right) + \left(\frac{b\epsilon^2}{2} - a\epsilon\right) - \left(\frac{a^2}{2b} - \frac{a^2}{b}\right) \\ &= \frac{a^2}{b} + \frac{b\epsilon^2}{2} - a\epsilon.\end{aligned}$$

As a function of a, this expression is minimized when $a = b\epsilon/2$. In such case,

$$F(a,b) \geq F\left(\frac{b\epsilon}{2}, b\right) = \frac{b\epsilon^2}{4} \geq \frac{\epsilon^2}{8},$$

where we use $b \geq 1/2$. The result then follows.

B. Proof of Theorem 3.2

We give the proof of Theorem 3.2, followed by proofs of the lemmas used in this proof. First, we have the following technical lemma:

Lemma B.1. *We have (a) the mean squared loss $L_{S,p}$ is $4k$-Lipschitz in Θ with respect to $\ell_{2,1}$ norm, and (b) when $n \geq \log(\frac{2}{\delta})$, its empirical counterpart $\hat{L}_{S,p}$ is $(4k + 6\sigma)$-Lipschitz with at least a probability of $1 - \delta/2$.*

We give a proof in Appendix B.1. Next, we have the following useful lemma, which says that the empirical loss converges to the true loss.

Lemma B.2. *Given the same setup as in Lemma B.3, we have*

$$\mathbb{P}_p\left[\sup_{\Theta} |(\hat{L}_S(\Theta; Z) - \sigma(\xi)) - L_{S,p}(\Theta)| \leq \zeta\right] \geq 1 - \delta,$$

where $\sigma(\xi) \coloneqq \frac{1}{n}\sum_{i\in[n]} \xi_i^2$.

The proof is given in Appendix B.2. This lemma follows from a standard covering number argument. Using this result, we provide a sample complexity bound for learning the true parameter given the i.i.d. training examples $\{(x_i, y_i)\}_{i\in[n]}$ from p.

Now we define the mean absolute loss function and its corresponding empirical version:

$$L_{A,p}(\Theta) = \mathbb{E}_p\left[|f_\Theta(x) - f_{\Theta^*}(x)|\right], \quad \hat{L}_A(\Theta; Z) = \frac{1}{n}\sum_{i\in[n]} |f_\Theta(x_i) - y_i|.$$

Lemma B.3. *For any $\delta \in \mathbb{R}_{>0}$, we have*

$$\mathbb{P}_p\left[L_{A,p}(\hat{\Theta}) \leq \sqrt{2\zeta}\right] \geq 1 - \delta,$$

where ζ is defined as

$$\zeta = \sqrt{\frac{4096k^2(k \vee \sigma)^2}{n}\left(dk \max\left\{1, \log\left(1+\sqrt{\frac{n}{dk}}\right)\right\} + \log\frac{4}{\delta}\right)}.$$

We give a proof in Appendix B.3.

Lemma B.4. *Consider any small $\tilde{\eta} \in \mathbb{R}_{>0}$ satisfying $\tilde{\eta} \leq \min\{\frac{1}{1152^2\sqrt{2\pi}k^2d^{3/2}}, \frac{\alpha_0^2\pi^{1/2}}{1454^2k^2d^{1/2}}\}$. If the generalization error has $L_{A,p}(\Theta) \leq \tilde{\eta}$, where p is a distribution on the $(d-1)$-sphere such that*

$$\frac{1}{|S^{d-1}|}\int_{\mathcal{X}} |f_\Theta(x) - f_{\Theta^*}(x)|dx \leq \mathbb{E}_p\left[|f_\Theta(x) - f_{\Theta^*}(x)|\right],$$

then there exists a bijection $\sigma : [k] \to [k]$ such that

$$\min\{\|\theta_{\sigma(i)} - \theta_i^*\|_2, \|\theta_{\sigma(i)} + \theta_i^*\|_2\} \leq \tilde{\alpha} := 727\pi^{-\frac{1}{4}}kd^{\frac{1}{4}}\tilde{\eta}^{\frac{1}{2}}, \quad \forall i \in [k].$$

We give a proof in Appendix B.4. Note that the assumptions on the distribution p can be easily satisfied, e.g., when $p = \text{Uniform}(S^{d-1})$ is a uniform distribution on the sphere. The above result says given a small generalization error for certain parameter estimate Θ, its estimation error of the ground-truth parameter Θ^* up to a sign flip is also small correspondingly.

Finally, Theorem 3.2 follows directly from Lemmas B.3 & B.4 by taking $\tilde{\eta} = \sqrt{2\zeta}$.

B.1. Proof of Lemma B.1

The mean squared loss satisfies

$$\begin{aligned}
|L_{S,p}(\Theta) - L_{S,p}(\Theta')| &\leq \mathbb{E}_p[|(f_\Theta(x) - f_{\Theta^*}(x))^2 - (f_{\Theta'}(x) - f_{\Theta^*}(x))^2|] \\
&\leq \mathbb{E}_p[|(f_\Theta(x) - f_{\Theta^*}(x) + f_{\Theta'}(x) - f_{\Theta^*}(x))(f_\Theta(x) - f_{\Theta'}(x))|] \\
&\leq 4k\mathbb{E}_p[|f_\Theta(x) - f_{\Theta'}(x)|] \\
&\leq 4k\sum_{i\in[k]}\|\theta_i - \theta_i'\|_2,
\end{aligned}$$

where we use $\|x\|_2 = 1$ and $\|\theta_i\|_2 = \|\theta_i'\|_2 = 1$. Similarly, the empirical loss satisfies

$$\begin{aligned}
|\hat{L}_S(\Theta; Z) - \hat{L}_S(\Theta'; Z)| &= \left|\frac{1}{n}\sum_{i\in[n]}[(f_\Theta(x_i) - y_i)^2 - (f_{\Theta'}(x_i) - y_i)^2]\right| \\
&\leq \frac{1}{n}\sum_{i\in[n]}|(f_\Theta(x_i) - f_{\Theta^*}(x_i))^2 - (f_{\Theta'}(x_i) - f_{\Theta^*}(x_i))^2| + \frac{2}{n}\sum_{i\in[n]}|\xi_i(f_\Theta(x_i) - f_{\Theta'}(x_i))| \\
&\leq (4k + \frac{2}{n}\sum_{i\in[n]}|\xi_i|)\sum_{i\in[k]}\|\theta_i - \theta_i'\|_2.
\end{aligned}$$

Since ξ_i is σ-subgaussian, so is $|\xi_i|$. When $n \geq \log(\frac{2}{\delta})$, we have

$$\frac{2}{n}\sum_{i\in[n]}|\xi_i| \leq 2\left(E[|\xi_i|] + \sqrt{\frac{2\sigma^2\log(\frac{2}{\delta})}{n}}\right) \leq 6\sigma$$

with at least a probability of $1 - \delta/2$, where we use Lemma 1.4 from (Rigollet & Hütter, 2015) and Lemma F.2 by taking $t = \sqrt{2\sigma^2\log(\frac{2}{\delta})/n}$. The claim follows.

B.2. Proof of Lemma B.2

Then, construct an $\zeta/(16k+12\sigma)$-net $\mathcal{E}_\Theta$ with respect to $\ell_{2,1}$ norm of $\bar{\mathit{\Theta}} = \{\Theta \in \mathbb{R}^{k\times d} \mid \|\Theta\|_{2,1} = k\}$ (remember $\bar{\mathit{\Theta}}$ is a superset of Θ's domain $\mathit{\Theta}$). For any such $\Theta \in \bar{\mathit{\Theta}}$, there exists $\Theta' \in \mathcal{E}_\Theta$ such that

$$|(\hat{L}_S(\Theta;Z) - L_{S,p}(\Theta)) - (\hat{L}_S(\Theta';Z) - L_{S,p}(\Theta'))| \le (8k+6\sigma)\|\Theta - \Theta'\|_{2,1} \le \zeta/2,$$

with high probability $1-\delta/2$. Given the above inequality, we have

$$\begin{aligned}\mathbb{P}_p\left[\sup_{\Theta\in\mathit{\Theta}} |(\hat{L}_S(\Theta;Z) - \sigma(\xi)) - L_{S,p}(\Theta)| \ge \zeta\right] &\le \mathbb{P}_p\left[\sup_{\Theta\in\bar{\mathit{\Theta}}} |(\hat{L}_S(\Theta;Z) - \sigma(\xi)) - L_{S,p}(\Theta)| \ge \zeta\right] \\ &\le \mathbb{P}_p\left[\max_{\Theta\in\mathcal{E}_\Theta} |(\hat{L}_S(\Theta;Z) - \sigma(\xi)) - L_{S,p}(\Theta)| \ge \frac{\zeta}{2}\right] + \delta/2 \\ &\le \sum_{\Theta\in\mathcal{E}_\Theta} \mathbb{P}_p\left[|(\hat{L}_S(\Theta;Z) - \sigma(\xi)) - L_{S,p}(\Theta)| \ge \frac{\zeta}{2}\right] + \delta/2. \quad (20)\end{aligned}$$

Note that

$$(\hat{L}_S(\Theta;Z) - \sigma(\xi)) - L_{S,p}(\Theta) = \mathcal{L}_1 + 2\mathcal{L}_2,$$

where

$$\mathcal{L}_1 = \frac{1}{n}\sum_{i\in[n]} (f_\Theta(x_i) - f_{\Theta^*}(x_i))^2 - \mathbb{E}_p\left[(f_\Theta(x) - f_{\Theta^*}(x))^2\right], \quad \mathcal{L}_2 = \frac{1}{n}\sum_{i\in[n]} (f_\Theta(x_i) - f_{\Theta^*}(x_i))\xi_i.$$

The probability within the sum in (20) is then upper bounded by

$$\mathbb{P}_p\left[|(\hat{L}_S(\Theta;Z) - \sigma(\xi)) - L_{S,p}(\Theta)| \ge \frac{\zeta}{2}\right] \le \mathbb{P}_p\left[|\mathcal{L}_1| \ge \frac{\zeta}{4}\right] + \mathbb{P}_p\left[|\mathcal{L}_2| \ge \frac{\zeta}{8}\right].$$

In the following, we bound each of the above two terms respectively. For the first term, since $|f_\Theta(x_i) - f_{\Theta^*}(x_i)| \le 2k$, $(f_\Theta(x_i) - f_{\Theta^*}(x_i))^2$ is $4k^2$-subgaussian; thus, by Lemma F.2, we have

$$\mathbb{P}_p\left[|\mathcal{L}_1| \ge \frac{\zeta}{4}\right] \le 2\exp\left(-\frac{n\zeta^2}{2048k^4}\right). \quad (21)$$

Next, for the second term, since ξ_i is σ-subgaussian and $(f_\Theta(x_i) - f_{\Theta^*}(x_i))$ and ξ_i are independent, we can show that $(f_\Theta(x_i) - f_{\Theta^*}(x_i))\xi_i$ is $(2k\sigma)$-subgaussian; thus, by Lemma F.2, we have

$$\mathbb{P}_p\left[|\mathcal{L}_2| \ge \frac{\zeta}{8}\right] \le 2\exp\left(-\frac{n\zeta^2}{512k^2\sigma^2}\right). \quad (22)$$

Therefore, we obtain from (20) that

$$\begin{aligned}\mathbb{P}_p\left[\sup_{\Theta\in\mathit{\Theta}} |(\hat{L}_S(\Theta;Z) - \sigma(\xi)) - L_{S,p}(\Theta)| \ge \zeta\right] &\le 4|\mathcal{E}_\Theta|\exp\left(-\frac{n\zeta^2}{2048k^2(k^2+\sigma^2)}\right) \\ &\le 2\left(1 + \frac{2k(16k+12\sigma)}{\zeta}\right)^{dk}\exp\left(-\frac{n\zeta^2}{2048k^2(k^2+\sigma^2)}\right) \\ &\le 2\exp\left(-\frac{n\zeta^2}{2048k^2(k^2+\sigma^2)} + dk\log\left(1 + \frac{2k(16k+12\sigma)}{\zeta}\right)\right), \quad (23)\end{aligned}$$

where the second inequality follows Lemma F.1 by noticing that $\|\Theta\|_{2,1} = k$ for $\Theta \in \bar{\mathit{\Theta}}$. Finally, we choose ζ so that (23) is smaller than $\delta/2$—in particular, letting

$$\zeta = \sqrt{\frac{4096k^2(k\vee\sigma)^2}{n}\left(dk\max\left\{1, \log\left(1+\sqrt{\frac{n}{dk}}\right)\right\} + \log\frac{4}{\delta}\right)},$$

then we have (23) is upper bounded by $\delta/2$, where we use $k+\sigma \ge 1$. The result follows.

B.3. Proof of Lemma B.3

Since $\hat{\Theta}$ minimizes $\hat{L}_S(\Theta; Z)$, we have

$$\begin{aligned}
0 &\le L_{S,p}(\hat{\Theta}) - L_{S,p}(\Theta^*) \\
&\le L_{S,p}(\hat{\Theta}) - (\hat{L}_S(\hat{\Theta}; Z) - \sigma(\xi)) + (\hat{L}_S(\Theta^*; Z) - \sigma(\xi)) - L_{S,p}(\Theta^*) \\
&\le 2 \sup_{\Theta} |(\hat{L}_S(\Theta; Z) - \sigma(\xi)) - L_{S,p}(\Theta)| \le 2\zeta,
\end{aligned}$$

with probability at least $1-\delta$, where we use Lemma B.2. Thus, by Cauchy-Schwarz inequaltiy and the fact that $L_{S,p}(\Theta^*) = 0$, we have

$$L_{A,p}(\hat{\Theta}) \le \sqrt{L_{S,p}(\hat{\Theta})} \le \sqrt{2\zeta},$$

as claimed.

B.4. Proof of Lemma B.4

Using the condition on p, we have

$$\int_{X_i} |f_\Theta(x) - f_{\Theta^*}(x)| dx \le \int_{\mathcal{X}} |f_\Theta(x) - f_{\Theta^*}(x)| dx \le |S^{d-1}| L_{A,p}(\Theta) \le |S^{d-1}| \tilde{\eta}.$$

Then, letting $\eta = |S^{d-1}|\tilde{\eta}$, we have that $L_{X_i}(\Theta) \le \eta$ holds for all $i \in [k]$. Thus, by Proposition 3.1 (we will check the conditions on α later), we have

$$\min\{\|\theta_{\sigma(i)} - \theta_i^*\|_2, \|\theta_{\sigma(i)} + \theta_i^*\|_2\} \le \alpha$$

for all $i \in [k]$. Plugging the value of η above into α, it yields that

$$\alpha = \frac{\frac{k\epsilon^3}{2} \frac{|S^{d-3}|}{|S^{d-1}|}}{\frac{\epsilon^2(1-d\epsilon^2/2)}{8} \frac{|S^{d-2}|}{|S^{d-1}|} - \tilde{\eta} - 6kd\epsilon^3 \frac{|S^{d-2}|}{|S^{d-1}|}}. \tag{24}$$

Using $|S^{d-1}| = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ and Lemma 10 from (Xu et al., 2021b), it holds that for any $d > 1$

$$\frac{|S^{d-3}|}{|S^{d-1}|} \le \frac{d}{2\pi}, \qquad \frac{|S^{d-2}|}{|S^{d-1}|} \ge \frac{\sqrt{d}}{2\sqrt{2\pi}}.$$

Combining the above, we obtain from (24)

$$\alpha \le \frac{4\sqrt{2}kd\epsilon^3}{(\pi d)^{1/2}(1 - d\epsilon^2/2 - 48kd\epsilon)\epsilon^2 - 16\sqrt{2\pi}\tilde{\eta}}.$$

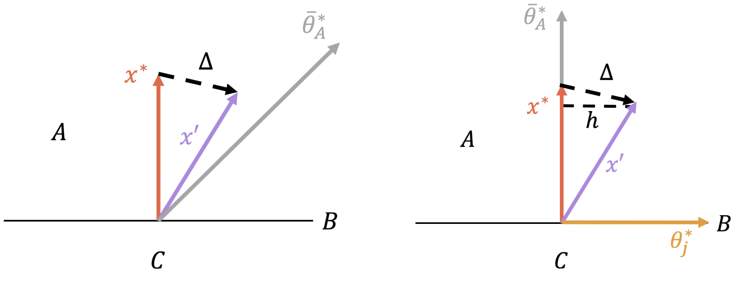
Take $\epsilon = (36\sqrt{2\pi}\frac{\tilde{\eta}}{\sqrt{d}})^{1/2}$. Given our choice of $\tilde{\eta}$, we can show that $d\epsilon^2/2 \le 1/4$ and $48kd\epsilon \le 1/4$. Thus, we have

$$\alpha \le 727\pi^{-\frac{1}{4}} k d^{\frac{1}{4}} \tilde{\eta}^{\frac{1}{2}}.$$

Note that such α satisfies $\alpha \le \alpha_0/2$ by our condition on $\tilde{\eta}$. The claim follows.

C. Proof of Proposition 4.2

We give a proof of Proposition 4.2.



(a) (b)

Figure 4: Illustrations for proof sketch of the optimal action gap ν_*.

C.1. Intuition

We illustrate our proof strategy in Figure 4. For all the neurons θ_i^*'s, we can decompose them into three subsets — i.e., the neurons that are "positively" activated, $A = \{i \in [k] \mid \theta_i^{*\top} x^* > 0\}$, those that are orthogonal to x^*, $B = \{i \in [k] \mid \theta_i^{*\top} x^* = 0\}$, and those that are inactive, $C = \{i \in [k] \mid \theta_i^{*\top} x^* < 0\}$. Define $\bar{\theta}_A^* = \sum_{i \in A} \theta_i^*$. We first show that $\bar{\theta}_A^*$ and x^* should be in the same direction (Figure 4 (a)). Then, given this fact, we prove by showing that there is no neuron in the set B; otherwise, we can always find an action x' to improve the value of $f_{\Theta^*}(x)$ (Figure 4 (b)).

To see that, the maximum value of $f_{\Theta^*}(x)$ is equal to $f_{\Theta^*}(x^*) = \bar{\theta}_A^{*\top} x^*$. Therefore, if $\bar{\theta}_A^*$ is not aligned with x^*, we can always move x^* a bit closer to $\bar{\theta}_A^*$ to a position x' such that $\bar{\theta}_A^{*\top} x' > \bar{\theta}_A^{*\top} x^*$ without negatively affecting the other neurons in $B \cup C$. Then, suppose the set $B \neq \varnothing$ and there's at least one $j \in B$. We can always find an increasing direction Δ by moving x^* a bit closer to θ_j^*. Intuitively, Δ is almost orthogonal to $\bar{\theta}_A^*$ and in parallel with θ_j^*; thus, the increase by moving x^* closer to θ_j^* exceeds the the decrease by moving it away from $\bar{\theta}_A^*$ when Δ is small enough.

C.2. Proof of Proposition 4.2

First, we show that $\bar{\theta}_A^*$ and x^* are in the same direction. We prove by contradiction. Suppose $\bar{\theta}_A^*$ and x^* are not aligned. We will show that we can find an action x' such that $f_{\Theta^*}(x') > f_{\Theta^*}(x^*)$. Take x' to be Δ away from x^*, i.e., $x' = x^* + \Delta$, such that x', x^* and $\bar{\theta}_A^*$ are in the same hyperplane (see Figure 4 (a)). We first let Δ be small enough such that the neurons in A remains strictly activated, i.e., $\theta_i^{*\top} x' > 0$ for $i \in A$ (it suffices to take $\|\Delta\|_2 < \min_{i \in A} |\theta_i^{*\top} x^*|$). Then, define the angle between two vectors z and y as $\angle(z, y) := \arccos(\frac{z^\top y}{\|z\|_2 \|y\|_2})$.

As long as Δ is such that $\angle(x^*, \bar{\theta}_A^*) > \angle(x', \bar{\theta}_A^*)$, it's easy to show that $\bar{\theta}_A^{*\top} x^* < \bar{\theta}_A^{*\top} x'$ by noting that $\|x^*\|_2 = \|x'\|_2 = 1$ and $\angle(x^*, \bar{\theta}_A^*), \angle(x', \bar{\theta}_A^*) < \pi/2$. Additionally, we also require Δ to be small enough such that the neurons in C remains inactive choosing action x', i.e., $\theta_i^{*\top} x' < 0$ for $i \in C$; note that it suffices to take $\|\Delta\|_2 < \min_{i \in C} |\theta_i^{*\top} x^*|$. Finally, we have $\sum_{i \in B} g(\theta_i^{*\top} x') \geq \sum_{i \in B} g(\theta_i^{*\top} x^*) = 0$ since $g(z)$ is non-negative. Therefore, we conclude that

$$f_{\Theta^*}(x') = \bar{\theta}_A^{*\top} x' + \sum_{i \in B} g(\theta_i^{*\top} x') > \bar{\theta}_A^{*\top} x^* + \sum_{i \in B} g(\theta_i^{*\top} x^*) = f_{\Theta^*}(x^*),$$

which is a contradiction.

Now, we further show that $B = \varnothing$. Similarly, we use a proof by contradiction. Suppose there is at least some $j \in B$. We take an action $x' = x^* + \Delta$, such that x', x^* and θ_j^* are in the same hyperplane (see Figure 4 (b)). Note that both $\bar{\theta}_A^*$ and x^* are orthogonal to θ_j^* by definition of the set B. Then, it holds that $\bar{\theta}_A^{*\top} x' + \theta_j^{*\top} x' > \bar{\theta}_A^{*\top} x^* + \theta_j^{*\top} x^*$, as long as Δ is such that

$$\theta_j^{*\top} \Delta > -\bar{\theta}_A^{*\top} \Delta. \tag{25}$$

Since $\|x^*\| = \|x'\| = 1$ and $x' = x^* + \Delta$, we have $-2x^{*\top} \Delta = \|\Delta\|^2$. Besides, let h denote the perpendicular distance from x' to x^* and it satisfies

$$\sqrt{1-h^2}+\sqrt{\|\Delta\|_2^2-h^2}=1 \Leftrightarrow h=\|\Delta\|_2\sqrt{1-\frac{\|\Delta\|_2^2}{4}}.$$

By noting that $\bar{\theta}_A^*/\|\bar{\theta}_A^*\|_2=x^*$, it suffices to have

$$\cos(\angle(\theta_j^*,\Delta))\|\Delta\|_2 \geq \frac{1}{2}\|\bar{\theta}_A^*\|\|\Delta\|_2^2 \Leftrightarrow \frac{h}{\|\Delta\|_2} \geq \frac{1}{2}\|\bar{\theta}_A^*\|\|\Delta\|_2 \Leftrightarrow \|\Delta\|_2 \leq \frac{4}{1+\|\bar{\theta}_A^*\|_2^2} \tag{26}$$

so that (25) holds. In addition, we take Δ to be small enough such that the neurons in A remains strictly activated and those in C remain inactive; to that end, it suffices to take $\|\Delta\|_2<\min_{i\in A\cup C}|\theta_i^{*\top}x^*|$. Thus, we reach a contradiction that $f_{\Theta^*}(x')>f_{\Theta^*}(x^*)$. Our claim follows.

D. Proof of Theorem 4.3

Suppose the exploration stage ends at time t_0. By Theorem 3.2, we can ensure

$$\min\{\|\tilde{\theta}_i-\theta_i^*\|_2, \|\tilde{\theta}_i+\theta_i^*\|_2\} \leq \nu_*/2, \quad \forall i\in[k]$$

with probability at least $1-\delta/2$ by choosing t_0 large enough so that $727\pi^{-\frac{1}{4}}kd^{\frac{1}{4}}(2\zeta)^{\frac{1}{4}}\leq \nu_*/2$ (note that ζ depends on the sample size $n=t_0$). In particular, it suffices to take $\delta=1/\sqrt{T}$ and

$$t_0 \geq t_1(\nu_*), \quad \text{where} \quad t_1(\nu):=\frac{C_1k^{10}d^2(k\vee\sigma)^2}{\nu^8}\left(dk(\log(d(k\vee\sigma))\vee\log\log T)+\log(64T)\right) \tag{27}$$

for some constant C_1. Recall that Theorem 3.2 requires $\sqrt{2\zeta}\leq\min\left\{\frac{1}{1152^2\sqrt{2\pi}k^2d^{3/2}}, \frac{\alpha_0^2\pi^{1/2}}{1454^2k^2d^{1/2}}\right\}$; thus, we also require

$$t_0 \geq t_2 := C_2k^{10}d^6(k\vee\sigma)^2\left(dk(\log(d(k\vee\sigma))\vee\log\log T)+\log(64T)\right) \tag{28}$$

for some constant C_2. Therefore, we have, with a slight abuse of notation,

$$t_0=t_0(\nu_*)=\tilde{\Theta}(k^{13}d^3(1/\nu_*^8\vee d^4)), \quad \text{where} \quad t_0(\nu)=\max\{t_1(\nu),t_2\}. \tag{29}$$

Now we analyze the regret of our algorithm. At each time t, the per-period regret r_t can be upper bounded by

$$r_t=f_{\Theta^*}(x^*)-f_{\Theta^*}(x_t)\leq(\sum_{i\in[k]}\|\theta_i^*\|_2)(\|x^*\|_2+\|x_t\|_2)\leq 2k.$$

Therefore, the regret during the exploration stage is upper bounded by

$$\sum_{t\in[t_0]}r_t\leq 2kt_0=\tilde{O}(k^{14}d^3(1/\nu_*^8\vee d^4)).$$

In the second stage, we run OFUL to find the optimal policy for the linear function $f_{\theta^\ddagger}(x^\ddagger)$ given our estimate $\tilde{\Theta}_{t_0}$, where $x^\ddagger$ and $\theta^\ddagger$ are defined in (9) and (10) respectively. Following the same proof strategy of Theorem 3 in (Abbasi-Yadkori et al., 2011), it gives the regret bound in the second stage to be

$$\sum_{t\in[T]\backslash[t_0]}r_t\leq C_3\sqrt{kdT\log(\lambda+T/(2kd))}\left(\lambda^{1/2}\sqrt{k}+\sigma\sqrt{\log(4T)+2kd\log(1+T/(2\lambda kd))}\right)$$

with a probability at least $1-\delta/2$, where the contextual dimension d here is $2kd$, and $S=\sqrt{5k}$ by noting that $\|\theta^\ddagger\|\leq\sqrt{5k}$.

Finally, the above analysis shows that with a probability at least $1-\delta$, we both have a small estimation error in the exploration stage and a guanranteed regret upper bound of OFUL in the second stage. Thus, with a small probability $\delta=1/\sqrt{T}$, we would have linear regret scaling as $2kT$; thus, the expected regret in this case is bounded by $2kT\delta=2k\sqrt{T}$. Our claim then follows.

E. Proof of Theorem 4.4

We bound the regret for the three cases respectively: (i) all batch i satisfying $\nu_i > \nu_*$, (ii) $t \in (T_{i-1}, T_{i-1} + t_{0,i}]$ for all batch i with $\nu_i \le \nu_*$, and (iii) $t \in (T_{i-1} + t_{0,i}, T_i]$ for all batch i with $\nu_i \le \nu_*$.

First, in case (i), we have $i \le \log(\nu_0/\nu_*)/\log(b)$. Recall that the per-period regret r_t can be trivially bounded by $2k$. Thus, the regret in this case is upper bounded by

$$2k(a^{\log(\nu_0/\nu_*)/\log(b)} - 1)T_1 \le 2k(\nu_0/\nu_*)^{\frac{\log(a)}{\log(b)}} T_1.$$

Once $\nu_i \le \nu_*$, the gap ν_i is sufficiently accurate that the optimal action x^* becomes feasible in the search region. Then, we can follow our proof strategy in Appendix D.

Consider case (ii). Due to our exploration strategy, we randomly collect $t_{0,i} = t_0(\nu_i) - t_0(\nu_{i-1})$ samples in each batch i ($t_0(\nu)$ defined in (29)); thus, the total number of random samples we collect over time horizon T is upper bounded by

$$\sum_{i=\lceil \frac{\log(\nu_0/\nu_*)}{\log(b)} \rceil}^{M-1} t_{0,i} \le t_0(\nu_{M-1}).$$

Thus, the regret in case (ii) is upper bounded by

$$2kt_0(\nu_{M-1}) = \tilde{O}\left(k^{14}d^7 + k^{14}d^3T^{8\frac{\log(b)}{\log(a)}}\right),$$

where we use the definition of t_0 in (28) and the value of M in (13).

Next, we calculate the regret of running OFUL in case (iii). Same as the proof in Appendix D, we have

$$\sum_{i=\lceil \frac{\log(\nu_0/\nu_*)}{\log(b)} \rceil}^{M-1} \sum_{t=T_{i-1}+t_{0,i}}^{T_i} r_t = \tilde{O}(kd\sqrt{T}).$$

Note that the proof strategy in (Abbasi-Yadkori et al., 2011) works as long as the data are fetched sequentially, and the length of the time periods without model misspecification (i.e., batch i with $\nu_i < \nu_*$) scales as T, which both satisfy in our algorithm.

Finally, recall that at each batch i, there's a probability of $\delta_i = 1/\sqrt{T}$ that our analysis above will fail. Thus, with a probability of at most $M/\sqrt{T}$ across all M batches, we will have a linear regret. The expected regret for this small-probability event is upper bounded by

$$2kT \cdot M/\sqrt{T} = \tilde{O}(k\sqrt{T}).$$

Combining all the above gives our final result.

F. Useful Lemmas

Lemma F.1. *For a ball in $\mathbb{R}^{d_1 \times d_2}$ with radius r with respect to any norm, there exists an ζ-net $\mathcal{E}$ such that*

$$|\mathcal{E}| \le \left(1 + \frac{2r}{\zeta}\right)^{d_1 d_2}.$$

Proof. This claim follows by a direct application of Proposition 4.2.12 in (Vershynin, 2018). □

Lemma F.2. *Letting $\{x_i\}_{i\in[n]}$ be a set of independent σ-subgaussian random variables with mean μ_i, then for all $t \ge 0$, we have*

$$\Pr\left[|\frac{1}{n}\sum_{i\in[n]}(x_i - \mu_i)| \ge t\right] \le 2\exp\left(-\frac{nt^2}{2\sigma^2}\right).$$

Proof. See Proposition 2.5 of (Wainwright, 2019). □