Combinatorial Neural Bandits

Taehyun Hwang [*1] **Kyuwook Chai** [*1] **Min-hwan Oh** [1]

Abstract

We consider a contextual combinatorial bandit problem where in each round a learning agent selects a subset of arms and receives feedback on the selected arms according to their scores. The score of an arm is an unknown function of the arm's feature. Approximating this unknown score function with deep neural networks, we propose algorithms: Combinatorial Neural UCB (`CN-UCB`) and Combinatorial Neural Thompson Sampling (`CN-TS`). We prove that `CN-UCB` achieves $\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{T})$ or $\widetilde{\mathcal{O}}(\sqrt{\widetilde{d}TK})$ regret, where $\widetilde{d}$ is the effective dimension of a neural tangent kernel matrix, K is the size of a subset of arms, and T is the time horizon. For `CN-TS`, we adapt an optimistic sampling technique to ensure the optimism of the sampled combinatorial action, achieving a worst-case (frequentist) regret of $\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{TK})$. To the best of our knowledge, these are the first combinatorial neural bandit algorithms with regret performance guarantees. In particular, `CN-TS` is the first Thompson sampling algorithm with the worst-case regret guarantees for the general contextual combinatorial bandit problem. The numerical experiments demonstrate the superior performances of our proposed algorithms.

1. Introduction

We consider a general class of contextual semi-bandits with combinatorial actions, where in each round the learning agent is given a set of arms, chooses a subset of arms, and receives feedback on each of the chosen arms along with the reward based on the combinatorial actions. The goal of the agent is to maximize cumulative rewards through these repeated interactions. The feedback is given as a function of the feature vectors (contexts) of the chosen arms. However, the functional form of the feedback model is unknown to the agent. Therefore, the agent needs to carefully balance exploration and exploitation in order to simultaneously learn the feedback model and optimize cumulative rewards.

Many real-world applications are naturally combinatorial action selection problems. For example, in most online recommender systems, such as streaming services and online retail, recommended items are typically presented as a set or a list. Real-time vehicle routing can be formulated as the shortest-path problem under uncertainty which is a classic combinatorial problem. Network routing is also another example of a combinatorial optimization problem. Often, in these applications, the response model is not fully known a priori (e.g., user preferences in recommender systems, arrival time in vehicle routing) but can only be queried by sequential interactions. Therefore, these applications can be formulated as a combinatorial bandit problem.

Despite the generality and wide applicability of the combinatorial bandit problem in practice, the combinatorial action space poses a greater challenge in balancing exploration and exploitation. To overcome such a challenge, parametric models such as the (generalized) linear model are often assumed for the feedback model (Qin et al., 2014; Wen et al., 2015; Kveton et al., 2015; Zong et al., 2016; Li et al., 2016; 2019; Oh & Iyengar, 2019). These works typically extend the techniques in the (generalized) linear contextual bandits (Abe & Long, 1999; Auer, 2002; Filippi et al., 2010; Rusmevichientong & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011; Chu et al., 2011; Li et al., 2017) to utilize contextual information and the structure of the feedback/reward model to avoid the naive exploration in combinatorial action space. However, the representation power of the (generalized) linear model can be limited in many real-world applications. When the model assumptions are violated, often the performances of the algorithms that exploit the structure of a model can severely deteriorate.

Beyond the parametric assumption for the feedback model, discretization-based techniques (Chen et al., 2018; Nika et al., 2020) have been proposed to capture the non-linearity of the base arm under the Lipschitz condition on the feedback model. These techniques split the context space and compute an upper confidence bound of rewards for each context partition. The performances of the algorithms strongly

[*]Equal contribution [1]Graduate School of Data Science, Seoul National University, Seoul, Republic of Korea. Correspondence to: Min-hwan Oh <minoh@snu.ac.kr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Table 1. Comparison with the related work. For the neural bandit algorithms with single arm selection (Zhou et al., 2020; Zhang et al., 2021), the reward function is not defined for a super arm (or the reward function can be viewed the same as the feedback for a single arm). All of the feedback models assume the boundedness of feedback. $\widetilde{\mathcal{O}}$ is a big-$\mathcal{O}$ notation up to logarithmic factors.

	Combinatorial	Feedback	Reward	Regret
`C`^{2}`UCB` (Qin et al., 2014)	Yes	Linear	Lipschitz	$\widetilde{\mathcal{O}}(d\sqrt{T})$
`CombLinUCB` (Wen et al., 2015)	Yes	Linear	Sum of feedback	$\widetilde{\mathcal{O}}(K\sqrt{dT\min\{\log N, d\}})$
`CombLinTS` (Wen et al., 2015)	Yes	Linear	Sum of feedback	$\widetilde{\mathcal{O}}(dK\sqrt{T})^{\dagger}$
`CC-MAB` (Chen et al., 2018)	Yes	Lipschitz	Sub-modular	$\widetilde{\mathcal{O}}(2^d T^{\frac{d+4}{d+6}})$
`ACC-UCB` (Nika et al., 2020)	Yes	Lipschitz	Sub-modular	$\widetilde{\mathcal{O}}(T^{\frac{\bar{d}+1}{\bar{d}+2}})^{\ddagger}$
`Neural-UCB` (Zhou et al., 2020)	No	General	-	$\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{T})$
`Neural-TS` (Zhang et al., 2021)	No	General	-	$\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{T})$
`CN-UCB` (**this work**)	Yes	General	Lipschitz	$\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{T})$ or $\widetilde{\mathcal{O}}(\sqrt{\widetilde{d}TK})$
`CN-TS` (**this work**)	Yes	General	Lipschitz	$\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{TK})$

†Bayesian regret, which is a weaker notion of regret than the worst-case regret.

$^{\ddagger}\bar{d}$ represents the approximate optimality dimension related to context space.

depend on the policy of how to partition the context space. However, splitting the context space is computationally expensive. As the reward function becomes more complex, so does the splitting procedure. Thus, it is challenging to apply these methods to high-dimensional contextual bandits. In addition, the Lipschitz assumption on the feedback model (not on the reward function) does not hold when contexts close in the context space yield significantly different outcomes, i.e., when context space cannot be partitioned with respect to the outcome.

Deep neural networks have shown remarkable empirical performances in various learning tasks (LeCun et al., 2015; Goodfellow et al., 2016; Silver et al., 2016). Incorporating the superior representation power and recent advances in generalization theory of deep neural networks (Jacot et al., 2018; Cao & Gu, 2019) into contextual bandits, an *upper confidence bound* (UCB) algorithm as an extension of the linear contextual bandit has been proposed (Zhou et al., 2020). Extending the UCB approach, Zhang et al. (2021) proposed a neural network-based *Thompson Sampling* (TS) algorithm (Thompson, 1933). However, these algorithms are proposed only for single-action selection. How these algorithms generalize to the combinatorial action selection has remained open.

In this paper, we study provably efficient contextual combinatorial bandit algorithms without any modeling assumptions on the feedback model (with mild assumptions on the reward function which takes the feedback as an input). The extension to the combinatorial actions and providing provable performance guarantees requires more involved analysis and novel algorithmic modifications, particularly for the TS algorithm. To briefly illustrate this challenge, even under the simple linear feedback model, a worst-case regret bound has not been known for a TS algorithm with various classes of combinatorial actions. This is due to the difficulty of ensuring the optimism of randomly sampled combinatorial actions (see Section 4.1). Addressing such challenges, we adapt an optimistic sampling technique to our proposed TS algorithm, which allows us to achieve a sublinear regret.

Our main contributions are as follows:

- We propose algorithms for a general class of contextual combinatorial bandits: *Combinatorial Neural UCB* (`CN-UCB`) and *Combinatorial Neural Thompson Sampling* (`CN-TS`). To the best of our knowledge, these are the first neural-network based combinatorial bandit algorithms with regret guarantees.
- We establish that `CN-UCB` is statistically efficient achieving $\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{T})$ or $\widetilde{\mathcal{O}}(\sqrt{\widetilde{d}TK})$ regret, where $\widetilde{d}$ is the effective dimension of a neural tangent kernel matrix, K is the size of a subset of arms, and T is the time horizon. This result matches the corresponding regret bounds of linear contextual bandits.
- The highlight of our contributions is that `CN-TS` is the first TS algorithm with the worst-case regret guarantees of $\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{TK})$ for a general class of contextual combinatorial bandits. To our best knowledge, even under a simpler, linear feedback model, the existing TS algorithms with various combinatorial actions (including semi-bandit) do not have the worst-case regret guarantees. This is due to the difficulty of ensuring the optimism of sampled combinatorial actions. We overcome this challenge by adapting optimistic sampling of the estimated reward while directly sampling in the reward space.

- The numerical evaluations demonstrate the superior performances of our proposed algorithms. We observe that the performances of the benchmark methods deteriorate significantly when the modeling assumptions are violated. In contrast, our proposed methods exhibit consistent competitive performances.

2. Problem setting

2.1. Notations

For a vector $\mathbf{x} \in \mathbb{R}^d$, we denote its ℓ_2-norm by $\|\mathbf{x}\|_2$ and its transpose by $\mathbf{x}^\top$. The weighted ℓ_2-norm associated with a positive definite matrix $\mathbf{A}$ is defined by $\|\mathbf{x}\|_{\mathbf{A}} := \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$. The trace of a matrix $\mathbf{A}$ is $tr(\mathbf{A})$. We define $[N]$ for a positive integer N to be a set containing positive integers up to N, i.e., $\{1, 2, \ldots, N\}$.

2.2. Contextual Combinatorial Bandit

In this work, we consider a contextual combinatorial bandit, where T is the total number of rounds, and N is the number of arms. At round $t \in [T]$, a learning agent observes the set of context vectors for all arms $\{\mathbf{x}_{t,i} \in \mathbb{R}^d \mid i \in [N]\}$ and chooses a set of arms $S_t \subset [N]$ with size constraint $|S_t| = K$. S_t is called a *super arm*. We introduce the notion of candidate super arm set $\mathcal{S} \subset 2^{[N]}$ defined as the set of all possible subsets of arms with size K, i.e., $\mathcal{S} := \{S \subset [N] \mid |S| = K\}$.

2.2.1. Score Function for Feedback

Once a super arm $S_t \in \mathcal{S}$ is chosen, the agent then observes the scores of the chosen arms $\{v_{t,i}\}_{i\in S_t}$ and receives a reward $R(S_t, \mathbf{v}_t)$ as a function of the scores $\mathbf{v}_t := [v_{t,i}]_{i=1}^N$ (which we discuss in the next section). This type of feedback is also known as *semi-bandit* feedback (Audibert et al., 2014). Note that in combinatorial bandits, feedback and reward are not necessarily the same as is the case in non-combinatorial bandits. For each $t \in [T]$ and $i \in [N]$, score $v_{t,i}$ is assumed to be generated as follows:

$$v_{t,i} = h(\mathbf{x}_{t,i}) + \xi_{t,i} \tag{1}$$

where h is an *unknown* function satisfying $0 \le h(\mathbf{x}) \le 1$ for any $\mathbf{x}$, and $\xi_{t,i}$ is a ρ-sub-Gaussian noise satisfying $\mathbb{E}[\xi_{t,i}|\mathcal{F}_t] = 0$ where $\mathcal{F}_t$ is the history up to round t.

To learn the score function h in Eq.(1), we use a fully connected neural network (Zhou et al., 2020; Zhang et al., 2021) with depth $L \ge 2$, defined recursively:

$$\begin{aligned} f_1 &= \mathbf{W}_1\mathbf{x} \\ f_\ell &= \mathbf{W}_\ell\phi(f_{\ell-1}), \quad 2 \le \ell \le L, \\ f(\mathbf{x};\boldsymbol{\theta}) &= \sqrt{m} f_L \end{aligned} \tag{2}$$

where $\boldsymbol{\theta} := [\mathrm{vec}(\mathbf{W}_1)^\top, ..., \mathrm{vec}(\mathbf{W}_L)^\top]^\top \in \mathbb{R}^p$ is the parameter of the neural network with $p = dm + m^2(L-2) + m$, $\phi(x) := \max\{x, 0\}$ is the ReLU activation function, and m is the width of each hidden layer. We denote the gradient of the neural network by $\mathbf{g}(\mathbf{x};\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta}) \in \mathbb{R}^p$.

2.2.2. Reward Function & Regret

$R(S, \mathbf{v})$ is a deterministic reward function that measures the quality of the super arm S based on the scores $\mathbf{v}$. For example, the reward of a super arm S_t can be the sum of the scores of arms in S_t, i.e., $R(S_t, \mathbf{v}_t) = \sum_{i\in S_t} v_{t,i}$. For our analysis, the reward function can be any function (linear or non-linear) which satisfies the following mild assumptions standard in the combinatorial bandit literature (Qin et al., 2014; Li et al., 2016).

Assumption 1 (Monotonicity). *$R(S, \mathbf{v})$ is monotone non-decreasing with respect to the score vector $\mathbf{v} = [v_i]_{i=1}^N$, which means, for any S, if $v_i \le v'_i$ for all $i \in [N]$, we have $R(S, \mathbf{v}) \le R(S, \mathbf{v}')$.*

Assumption 2 (Lipschitz continuity). *$R(S, \mathbf{v})$ is Lipschitz continuous with respect to the score vector $\mathbf{v}$ restricted on the arms in S, which means, there exists a constant $C_0 > 0$ such that for any $\mathbf{v}$ and $\mathbf{v}'$, we have $|R(S, \mathbf{v}) - R(S, \mathbf{v}')| \le C_0\sqrt{\sum_{i\in S}(v_i - v'_i)^2}$.*

Remark 1. *Reward function satisfying Assumptions 1 and 2 encompasses a wide range of combinatorial feedback models including semi-bandit, document-based or position based ranking models, and cascading models with little change to the learning algorithm. See Appendix G for more detailed discussions.*

Note that we do not require the agent to have direct knowledge on the explicit form of the reward function $R(S, \mathbf{v})$. For the sake of clear exposition, we assume that the agent has access to an exact optimization oracle $\mathbb{O}_{\mathcal{S}}(\mathbf{v})$ which takes a score vector $\mathbf{v}$ as an input and returns the solution of the maximization problem $\mathrm{argmax}_{S\in\mathcal{S}} R(S, \mathbf{v})$.

Remark 2. *One can trivially extend the exact optimization oracle to an α-approximation oracle without altering the learning algorithm or regret analysis. For problems such as semi-bandit algorithms choosing top-K arms, exact optimization can be done by simply sorting base scores. Even for more challenging assortment optimization, there are many polynomial-time (approximate) optimization methods available (Rusmevichientong et al., 2010; Davis et al., 2014). For this reason, we present the regret analysis without α-approximation assumption. Extension of our regret analysis to an α-approximation oracle is given in Appendix E.*

The goal of the agent is to minimize the following (worst-case) cumulative expected regret:

$$\mathcal{R}(T) = \sum_{t=1}^{T} \mathbb{E}\left[R(S_t^*, \mathbf{v}_t^*) - R(S_t, \mathbf{v}_t^*)\right] \tag{3}$$

where $\mathbf{v}_t^* := [h(\mathbf{x}_{t,i})]_{i=1}^N$ is the expected score which is unknown, and $S_t^* := \text{argmax}_{S\in\mathcal{S}} R(S, \mathbf{v}_t^*)$ is the offline *optimal* super arm at round t under the expected score.

3. Combinatorial Neural UCB (CN-UCB)

3.1. CN-UCB Algorithm

In this section, we present our first algorithm, Combinatorial Neural UCB (CN-UCB). CN-UCB is a neural network-based UCB algorithm that operates using the *optimism in the face of uncertainty* (OFU) principle (Lai & Robbins, 1985) for combinatorial actions.

In our proposed method, the neural network used for feedback model approximation is initialized by randomly generating each entry of $\boldsymbol{\theta}_0 = [\text{vec}(\mathbf{W}_1)^\top, ..., \text{vec}(\mathbf{W}_L)^\top]^\top$, where for each $\ell \in [L-1]$, $\mathbf{W}_\ell = (\mathbf{W}, \mathbf{0}; \mathbf{0}, \mathbf{W})$ with each entry of $\mathbf{W}$ generated independently from $\mathcal{N}(0, 4/m)$ and $\mathbf{W}_L = (\mathbf{w}^\top, -\mathbf{w}^\top)$ with each entry of $\mathbf{w}$ generated independently from $\mathcal{N}(0, 2/m)$. At each round $t \in [T]$, the algorithm observes the contexts for all arms, $\{\mathbf{x}_{t,i}\}_{i\in[N]}$ and computes an upper confidence bound $u_{t,i}$ of the expected score for each arm i, based on $\mathbf{x}_{t,i}, \boldsymbol{\theta}_{t-1}$, and the exploration parameter γ_{t-1}. Then, the sum of upper confidence bound score vector $\mathbf{u}_t := [u_{t,i}]_{i=1}^N$ and the offset term vector $\mathbf{e}_t := [e_t, \cdots, e_t]$, (specified in Lemma 1), is passed to the optimization oracle $\mathbb{O}_\mathcal{S}$ as input. Then, the agent plays $S_t = \mathbb{O}_\mathcal{S}(\mathbf{u}_t + \mathbf{e}_t)$ and receives the corresponding scores $\{v_{t,i}\}_{i\in S_t}$ as feedback along with the reward associated with super arm S_t. Then the algorithm updates $\boldsymbol{\theta}_t$ by minimizing the following loss function in Eq.(4) using gradient descent with step size η for J times.

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{k=1}^{n}\left(f(\mathbf{x}^k; \boldsymbol{\theta}) - v^k\right)^2 + \frac{m\lambda}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2 \quad (4)$$

Here, the loss is minimized using ℓ_2-regularization. Hyperparameter λ controls the level of regularization, where the regularization centers at the randomly initialized neural network parameter $\boldsymbol{\theta}_0$. The CN-UCB algorithm is summarized in Algorithm 1.

3.2. Regret of CN-UCB

For brevity, we denote $\{\mathbf{x}^k\}_{k=1}^{TN}$ be the collection of all contexts $\{\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{T,N}\}$.

Definition 1. *(Jacot et al., 2018; Cao & Gu, 2019) Define*

$$\widetilde{\mathbf{H}}_{i,j}^{(1)} = \boldsymbol{\Sigma}_{i,j}^{(1)} = \langle \mathbf{x}^i, \mathbf{x}^j \rangle, \mathbf{A}_{i,j}^{(\ell)} = \begin{pmatrix} \boldsymbol{\Sigma}_{i,i}^{(\ell)} & \boldsymbol{\Sigma}_{i,j}^{(\ell)} \\ \boldsymbol{\Sigma}_{j,i}^{(\ell)} & \boldsymbol{\Sigma}_{j,j}^{(\ell)} \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{i,j}^{(\ell+1)} = 2\mathbb{E}_{(y,z)\sim\mathcal{N}(\mathbf{0},\mathbf{A}_{i,j}^{(\ell)})}\left[\phi(y)\phi(z)\right],$$

$$\widetilde{\mathbf{H}}_{i,j}^{(\ell+1)} = 2\widetilde{\mathbf{H}}_{i,j}^{(\ell)}\mathbb{E}_{(y,z)\sim\mathcal{N}(\mathbf{0},\mathbf{A}_{i,j}^{(\ell)})}\left[\phi'(y)\phi'(z)\right] + \boldsymbol{\Sigma}_{i,j}^{(\ell+1)}.$$

Then, $\mathbf{H} = (\widetilde{\mathbf{H}}^{(L)} + \boldsymbol{\Sigma}^{(L)})/2$ is called the neural tangent kernel (NTK) matrix on the context set $\{\mathbf{x}^k\}_{k=1}^{TN}$.

The NTK matrix $\mathbf{H}$ on the contexts $\{\mathbf{x}^k\}_{k=1}^{TN}$ is defined recursively from the input layer to the output layer of the network (Zhou et al., 2020; Zhang et al., 2021). Then, we define the effective dimension of the NTK matrix $\mathbf{H}$.

Definition 2. *The effective dimension $\widetilde{d}$ of the NTK matrix $\mathbf{H}$ with regularization parameter λ is defined as*

$$\widetilde{d} = \frac{\log\det(\mathbf{I} + \mathbf{H}/\lambda)}{\log(1 + TN/\lambda)}. \quad (5)$$

The effective dimension can be thought of as the actual dimension of contexts in the Reproducing Kernel Hilbert Space spanned by the NTK. For further detailed information, we refer the reader to Jacot et al. (2018). We proceed under the following assumption regarding contexts:

Assumption 3. *For any $k \in [TN]$, $\|\mathbf{x}^k\|_2 = 1$ and $[\mathbf{x}^k]_j = [\mathbf{x}^k]_{j+\frac{d}{2}}$ for $1 \le j \le \frac{d}{2}$. Furthermore, for some $\lambda_0 > 0$, $\mathbf{H} \succeq \lambda_0 \mathbf{I}$.*

This is a mild assumption commonly used in the neural contextual bandits (Zhou et al., 2020; Zhang et al., 2021). $\|\mathbf{x}\|_2 = 1$ is only imposed for simplicity of exposition. For the condition on the entries of $\mathbf{x}$, we can always reconstruct a new context $\mathbf{x}' = [\mathbf{x}^\top, \mathbf{x}^\top]^\top/\sqrt{2}$. A positive definite NTK matrix is a standard assumption in the NTK literature (Du et al., 2019; Arora et al., 2019), also used in the aforementioned neural contextual bandit literature. The following theorem provides the regret bound of Algorithm 1.

Theorem 1. *Suppose Assumptions 1-3 hold. Let $\mathbf{h} = \left[h(\mathbf{x}^k)\right]_{k=1}^{TN} \in \mathbb{R}^{TN}$. If we run CN-UCB with*

$$m \ge \textit{poly}(T, L, N, \lambda^{-1}, \lambda_0^{-1}, \log T),$$
$$\eta = \bar{C}_1(TKmL + m\lambda)^{-1},\ \lambda \ge \bar{C}_2 LK,$$
$$J = 2\log\left(\sqrt{\lambda/TK}/(\lambda + \bar{C}_3 TKL)\right) TKL/(\bar{C}_1\lambda)$$

for some positive constants $\bar{C}_1, \bar{C}_2, \bar{C}_3$ with $\bar{C}_2 \ge \sqrt{\max_{t,i}\|\mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m}\|_2^2/L}$ and $B \ge \sqrt{2\mathbf{h}^\top\mathbf{H}^{-1}\mathbf{h}}$, then the cumulative expected regret of CN-UCB over horizon T is upper-bounded by

$$\mathcal{R}(T) = \widetilde{\mathcal{O}}\left(\sqrt{\widetilde{d}T\max\{\widetilde{d}, K\}}\right).$$

Discussion of Theorem 1. Theorem 1 establishes that the cumulative regret of CN-UCB is $\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{T})$ or $\widetilde{\mathcal{O}}(\sqrt{\widetilde{d}TK})$, whichever is higher. This result matches the state-of-the-art regret bounds for the contextual combinatorial bandits with the linear feedback model (Li et al., 2016; Zong et al., 2016; Li & Zhang, 2018). Note that the existence of $\bar{C}_2$ in Theorem 1 follows from Lemma B.6 in Zhou et al. (2020)

Algorithm 1 Combinatorial Neural UCB (`CN-UCB`)

Input: Number of rounds T, regularization parameter λ, norm parameter B, step size η, network width m, number of gradient descent steps J, network depth L.
Initialization: Randomly initialize $\boldsymbol{\theta}_0$ as described in Section 3.1 and $\mathbf{Z}_0 = \lambda \mathbf{I}$
for $t = 1, ..., T$ **do**
  Observe $\{\mathbf{x}_{t,i}\}_{i\in[N]}$
  Compute $\hat{v}_{t,i} = f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})$ and $u_{t,i} = \hat{v}_{t,i} + \gamma_{t-1} \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|_{\mathbf{Z}_{t-1}^{-1}}$ for $i \in [N]$
  Let $S_t = \mathbb{O}_{\mathcal{S}}(\mathbf{u}_t + \mathbf{e}_t)$
  Play super arm S_t and observe $\{v_{t,i}\}_{i\in S_t}$
  Update $\mathbf{Z}_t = \mathbf{Z}_{t-1} + \sum_{i\in S_t} \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})\mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})^\top/m$
  Update $\boldsymbol{\theta}_t$ to minimize the loss in Eq.(4) using gradient descent with η for J times
  Compute γ_t and e_{t+1} described in lemma 1
end for

and Lemma B.3 in Cao & Gu (2019). While the regret analysis for Theorem 1 has its own merit, the technical lemmas for Theorem 1 also provide the building block for the more challenging analysis of the TS algorithm which is presented in Section 4.

3.3. Proof Sketch of Theorem 1

In this section, we provide a proof sketch of the regret upper bound in Theorem 1 and the key lemmas whose proofs are deferred to Appendix A.

Recall that we do not make any parametric assumption on the score function, but a neural network is used to approximate the unknown score function. Hence, we need to carefully control the approximation error. To achieve this, we use an over-parametrized neural network, for which the following condition on the neural network width is required.

Condition 1. *The network width m satisfies*

$$
\begin{aligned}
m &\geq C \max\{L^{-\frac{3}{2}} K^{-\frac{1}{2}} \lambda^{\frac{1}{2}} \left(\log(TNL^2/\delta)\right)^{\frac{3}{2}}, \\
&\quad T^6 N^6 L^6 \log(T^2 N^2 L/\delta) \max\{\lambda_0^{-4}, 1\}\}, \\
m\,(\log m)^{-3} &\geq C T^4 K^4 L^{21} \lambda^{-4} (1 + \sqrt{T/\lambda})^6 \\
&\quad + C T K L^{12} \lambda^{-1} + C T^4 K^4 L^{18} \lambda^{-10} (\lambda + TL)^6 ,
\end{aligned}
$$

where C is a positive absolute constant.

Unlike the analysis of the (generalized) linear UCB algorithms (Abbasi-Yadkori et al., 2011; Li et al., 2017), we do not have guarantees on the upper confidence bound $u_{t,i}$ being higher than the expected score $v^*_{t,i} = h(\mathbf{x}_{t,i})$ due to the approximation error. Therefore, we consider adding the offset term to the the upper confidence bound to ensure optimism. The following lemma shows that the upper confidence bounds $u_{t,i}$ do not deviate far from the expected score $h(\mathbf{x}_{t,i})$ and specifies the value of the offset term.

Lemma 1. *For any $\delta \in (0, 1)$, suppose the width of the neural network m satisfies Condition 1. Let γ_t be a positive scaling factor defined as*

$$
\begin{aligned}
\gamma_t = \Gamma_{1,t} &\left(\rho \sqrt{\log \frac{\det \mathbf{Z}_t}{\det \lambda \mathbf{I}} + \Gamma_{2,t} - 2\log\delta} + \sqrt{\lambda} B \right) \\
&+ (\lambda + C_1 tKL) \left((1-\eta m\lambda)^{\frac{J}{2}} \sqrt{tK/\lambda} + \Gamma_{3,t} \right) ,
\end{aligned}
$$

where

$$
\begin{aligned}
\Gamma_{1,t} &= \sqrt{1 + C_{\Gamma,1} t^{\frac{7}{6}} K^{\frac{7}{6}} L^4 \lambda^{-\frac{7}{6}} m^{-\frac{1}{6}} \sqrt{\log m}} , \\
\Gamma_{2,t} &= C_{\Gamma,2} t^{\frac{5}{3}} K^{\frac{5}{3}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} , \\
\Gamma_{3,t} &= C_{\Gamma,3} t^{\frac{7}{6}} K^{\frac{7}{6}} L^{\frac{7}{2}} \lambda^{-\frac{7}{6}} m^{-\frac{1}{6}} \sqrt{\log m} (1 + \sqrt{tK/\lambda}) ,
\end{aligned}
$$

for some constants $C_1, C_{\Gamma,1}, C_{\Gamma,2}, C_{\Gamma,3} > 0$. If $\eta \leq C_2(TKmL + m\lambda)^{-1}$ for some $C_2 > 0$, then for any $t \in [T]$ and $i \in [N]$, with probability at least $1-\delta$ we have

$$
|u_{t,i} - h(\mathbf{x}_{t,i})| \leq 2\gamma_{t-1} \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|_{\mathbf{Z}_{t-1}^{-1}} + e_t ,
$$

where e_t is defined for some absolute constants $C_3, C_4 > 0$ as follows.

$$
\begin{aligned}
e_t := C_3 \gamma_{t-1} t^{\frac{1}{6}} K^{\frac{1}{6}} L^{\frac{7}{2}} \lambda^{-\frac{2}{3}} m^{-\frac{1}{6}} \sqrt{\log m} \\
+ C_4 t^{\frac{2}{3}} K^{\frac{2}{3}} \lambda^{-\frac{2}{3}} m^{-\frac{1}{6}} \sqrt{\log m} .
\end{aligned}
$$

The next corollary shows that the surrogate upper confidence bound $u_{t,i} + e_t$ is higher than true mean score $h(\mathbf{x}_{t,i})$ with high probability.

Corollary 1. *With probability at least $1-\delta$*

$$
u_{t,i} + e_t \geq h(\mathbf{x}_{t,i}) .
$$

The point of Corollary 1 is that in Zhou et al. (2020), to bound the instantaneous regret, it is enough for the agent to choose only one optimistic action (see Lemma 5.3 in Zhou et al. (2020)), while in our case, the agent has to choose the optimistic super arm in order to bound the instantaneous regret (See Eq. (8) in Proof of Theorem 1). However, in order to ensure the optimism of the chosen super arm, it is necessary to guarantee the optimism of all individual arms in the chosen super arm, which is represented in Corollary 1.

The following technical lemma bounds the sum of weighted norms which is similar to Lemma 4.2 in Qin et al. (2014) and Lemma 5.4 in Zhou et al. (2020).

Lemma 2. *For any $\delta \in (0, 1)$ suppose the width of the neural network m satisfies Condition 1. If $\eta \leq C_1(TKmL + m\lambda)^{-1}$, and $\lambda \geq C_2 LK$, for some positive constant C_1, C_2 with $C_2 \geq \sqrt{\max_{t,i} \|\mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m}\|_2^2 / L}$, then with probability at least $1 - \delta$, for some $C_3 > 0$,*

$$\sum_{t=1}^{T} \sum_{i \in S_t} \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|^2_{\mathbf{Z}_{t-1}^{-1}} \leq 2\widetilde{d} \log(1+TN/\lambda) + 2 + C_3 T^{\frac{5}{3}} K^{\frac{3}{2}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m}\,.$$

Combining these results, we can derive the regret bound in Theorem 1. First, using the Lipschitz continuity of the reward function, we bound the instantaneous regret with the sum of scores for each individual arm within the super arm. By Lemma 1, the upper confidence bound of the over-parametrized neural network concentrates well around the true score function. By adding an arm-independent offset term, we can ensure the optimism of the surrogate upper confidence bound. Then, we apply Lemma 2 to derive the desired cumulative regret bound.

4. Combinatorial Neural TS (CN-TS)

4.1. Challenges in Worst-Case Regret Analysis for Combinatorial Actions

The challenges in the worst-case (non-Bayesian) regret analysis for TS algorithms with combinatorial actions lie in the difficulty of ensuring optimism of a sampled combinatorial action. The key analytical element to drive a sublinear regret for any TS algorithm, either combinatorial or non-combinatorial, is to show that a sampled action is optimistic with sufficient frequency (Agrawal & Goyal, 2013; Abeille & Lazaric, 2017). With combinatorial actions, however, ensuring optimism becomes more challenging than single-action selection. In particular, if the structure of the reward and feedback model is not known, one can only resort to hoping that all of the sampled base arms in the chosen super arm S_t are optimistic, i.e., the scores of all sampled base arms are higher than their expected scores. The probability of such an event can be exponentially small in the size of the super arm K.

For example, let the probability that the sampled score of the i-th arm is higher than the corresponding expected score be at least $\widetilde{p}$, i.e., $\mathbb{P}(\widetilde{v}_i > h(\mathbf{x}_i)) \geq \widetilde{p}$. If the sampled score of every arm is optimistic, by the monotonicity property of the reward function, the reward induced by the sampled scores would be larger than the reward induced by the expected score, i.e., $R(S, \widetilde{\mathbf{v}}) \geq R(S, \mathbf{v}^*)$. However, the probability of the event that all the K sampled scores are higher than their corresponding expected scores would be in the order of $\widetilde{p}^K$. Hence, the probability of such an event can be exponentially small in the size of the super arm K.

Note that in the UCB exploration, one can ensure high-probability optimism even with combinatorial actions in a straightforward manner since action selection is deterministic. However, in TS with combinatorial actions, suitable random exploration with provable efficiency is much more challenging to guarantee. This challenge is further exacerbated by the complex analysis based on neural networks that we consider in this work.

4.2. CN-TS Algorithm

To address the challenge of TS exploration with combinatorial actions described above, we present CN-TS, a neural network-based TS algorithm. We make two modifications from conventional TS for parametric bandits. First, instead of maintaining an actual Bayesian posterior as in the canonical TS algorithms, CN-TS is a generic randomized algorithm that samples rewards from a Gaussian distribution. The algorithm directly samples estimated rewards from a Gaussian distribution, rather than sampling network parameters – this modification is adapted from Zhang et al. (2021).

Second, in order to ensure sufficient *optimistic* sampling in combinatorial action space, we draw multiple M independent score samples for each arm instead of drawing a single sample. Leveraging these multiple samples, we compute the most optimistic (the highest estimated) score for each arm. We demonstrate that implementing this modification effectively ensures the required optimism of samples, formalized in Lemma 3. The algorithm is summarized in Algorithm 2.

4.3. Regret of CN-TS

Under the same assumptions introduced in the analysis of CN-UCB, we present the worst-case regret bound for CN-TS in Theorem 2.

Theorem 2. *Suppose Assumptions 1-3 hold and m satisfies Condition 1. If we run CN-TS with*

$$\begin{aligned}
&\eta = \bar{C}_1(TKmL + m\lambda)\,, \lambda = \max\{1 + 1/T, \bar{C}_2 LK\}, \\
&J = 2\log\left(\sqrt{\lambda/TKL}/(4\bar{C}_3 T)\right) TKL/(\bar{C}_1\lambda) \\
&\nu = B + \rho\sqrt{\widetilde{d}\log(1 + TN/\lambda) + 2 + 2\log T}\,, \\
&B = \max\{1/(22e\sqrt{\pi}), \sqrt{2\mathbf{h}^\top \mathbf{H}\mathbf{h}}\}\,, \\
&M = \lceil 1 - \log K/\log(1 - \widetilde{p}) \rceil
\end{aligned}$$

for some positive constants $\bar{C}_1 > 0, \bar{C}_3 > 0$, and $\bar{C}_2 \geq$

Algorithm 2 Combinatorial Neural Thompson Sampling (CN-TS)

Input: Number of rounds T, regularization parameter λ, exploration variance ν, step size η, network width m, number of gradient descent steps J, network depth L, sample size M.
Initialization: Randomly initialize $\boldsymbol{\theta}_0$ as described in Section 3.1 and $\widetilde{\mathbf{Z}}_0 = \lambda \mathbf{I}$
for $t = 1, ..., T$ **do**
  Observe $\{\mathbf{x}_{t,i}\}_{i\in[N]}$
  Compute $\sigma_{t,i}^2 = \lambda \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})^\top \widetilde{\mathbf{Z}}_{t-1}^{-1} \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/m$ for each $i \in [N]$
  Sample $\{\widetilde{v}_{t,i}^{(j)}\}_{j=1}^M$ independently from $\mathcal{N}(f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}), \nu^2\sigma_{t,i}^2)$ for each $i \in [N]$
  Compute $\widetilde{v}_{t,i} = \max_j \widetilde{v}_{t,i}^{(j)}$ for each $i \in [N]$
  Let $S_t = \mathbb{O}_{\mathcal{S}}(\widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})$
  Play super arm S_t and observe $\{v_{t,i}\}_{i\in S_t}$
  Update $\widetilde{\mathbf{Z}}_t = \widetilde{\mathbf{Z}}_{t-1} + \sum_{i\in S_t} \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}) \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})^\top/m$
  Update $\boldsymbol{\theta}_t$ to minimize the loss (4) using gradient descent with η for J times
end for

$\sqrt{\max_{t,i} \|\mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m}\|_2^2/L}$*, then the cumulative expected regret of* CN-TS *over horizon* T *is upper-bounded by*

$$\mathcal{R}(T) = \widetilde{\mathcal{O}}(\widetilde{d}\sqrt{TK}) .$$

Discussion of Theorem 2. Theorem 2 establishes that the cumulative regret of CN-TS is $\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{TK})$. To the best of our knowledge, this is the first TS algorithm with the worst-case regret guarantees for general combinatorial action settings. This is crucial since various combinatorial bandit problems were prohibitive for TS methods due to the difficulty of ensuring the optimism of randomly selected super-action as discussed in Section 4.1. Our result also encompasses the linear feedback model setting, for which, to our best knowledge, a worst-case regret bound has not been proven for TS with combinatorial actions in general.

Remark 3. *Both* CN-UCB *and* CN-TS *depend on the condition of network size* m*. However, our experiments show superior performances of the proposed algorithms even when they are implemented with much smaller* m *(see Section 5). The large value of* m *is sufficient for regret analysis, due to the current state of the NTK theory. The same phenomenon is also present in the single action selection version of the neural bandits (Zhang et al., 2021; Zhou et al., 2020).*

Remark 4. *For a clear exposition of main ideas, the knowledge of* T *is assumed for both* CN-UCB *and* CN-TS*. This knowledge was also assumed in the previous neural bandit literature (Zhang et al., 2021; Zhou et al., 2020). We can replace this requirement of knowledge on* T *by using a doubling technique. We provide modified algorithms that do not depend on such knowledge of* T *in Appendix F.*

Remark 5. *The proposed optimistic sampling technique can be applied to the regret analysis for TS algorithms with combinatorial actions other than neural bandit settings. Regarding the cost of the optimistic sampling, this salient feature of the algorithm is controlled by the number of multiple samples* M*. A notable feature is that while this technique provides provably sufficient optimism, the proposed optimistic sampling technique comes at a minimal cost of* $\log M$*. That is, even if we over-sample by the factor of 2, the additional cost in the regret bound only increases by the additive logarithmic factor, i.e.,* $\log 2M = \log M + \log 2$*. Also, given that a theoretically suggested value of* M *as shown in Theorem 2 is only* $\Omega(\log K)$*, the regret caused by the optimistic sampling is of* $\mathcal{O}(\log \log K)$*.*

4.4. Proof Sketch of Theorem 2

For any $t \in [T]$, we define events $\mathcal{E}_t^\sigma$ and $\mathcal{E}_t^\mu$ similar to the prior literature on TS (Agrawal & Goyal, 2013; Zhang et al., 2021) defined as follows.

$$\mathcal{E}_t^\sigma := \{\omega \in \mathcal{F}_{t+1} \mid \forall i, |\widetilde{v}_{t,i} - f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})| \le \beta_t \nu \sigma_{t,i}\}$$
$$\mathcal{E}_t^\mu := \{\omega \in \mathcal{F}_{t+1} \mid \forall i, |f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}) - h(\mathbf{x}_{t,i})| \le \nu \sigma_{t,i} + \epsilon\}$$

where for some constants $\{C_{\epsilon,k}\}_{k=1}^4$, ϵ is defined as

$$\begin{aligned}
\epsilon := {} & C_{\epsilon,1} T^{\frac{2}{3}} K^{\frac{2}{3}} L^3 \lambda^{-\frac{2}{3}} m^{-\frac{1}{6}} \sqrt{\log m} \\
& + C_{\epsilon,2} (1 - \eta m \lambda)^{J/2} \sqrt{TKL/\lambda} \\
& + C_{\epsilon,3} T^{\frac{7}{6}} K^{\frac{7}{6}} L^4 \lambda^{-\frac{7}{6}} m^{-\frac{1}{6}} \sqrt{\log m} (1 + \sqrt{TK/\lambda}) \\
& + C_{\epsilon,4} T^{\frac{7}{6}} K^{\frac{7}{6}} \lambda^{-\frac{2}{3}} L^{\frac{9}{2}} m^{-\frac{1}{6}} \sqrt{\log m} \\
& \cdot \left(B + \rho \sqrt{\widetilde{d} \log(1 + TN/\lambda) + 2 - 2\log\delta} \right) .
\end{aligned}$$

Under event $\mathcal{E}_t^\sigma$, the difference between the optimistic sampled score and the estimated score can be controlled by the score's approximate posterior variance. Under the event $\mathcal{E}_t^\mu$, the estimated score based on the neural network does not deviate far from the expected score up to the approximate error term. Note that both events $\mathcal{E}_t^\mu, \mathcal{E}_t^\sigma$ happen with high probability. The remaining part is a guarantee on the probability of optimism for randomly sampled actions. Lemma 3 shows that the proposed optimistic sampling ensures a constant probability of optimism.

Lemma 3. *Suppose we take optimistic samples of size* $M = \lceil 1 - \frac{\log K}{\log(1-\widetilde{p})} \rceil$ *where* $\widetilde{p} := 1/(4e\sqrt{\pi})$. *Then we have*

$$\mathbb{P}\bigg(R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) > R(S_t^*, \mathbf{v}_t^*) | \mathcal{F}_t, \mathcal{E}_t^\mu \bigg) \geq \widetilde{p}$$

where $\boldsymbol{\epsilon} = [\epsilon, \dots, \epsilon] \in \mathbb{R}^N$.

Lemma 3 implies that even in the worst case, our randomized action selection still provides optimistic rewards at least with constant frequency. Hence, the regret pertaining to random sampling can be upper-bounded based on this frequent-enough optimism. The complete proof is deferred to Appendix B.

5. Numerical Experiments

In this section, we perform numerical evaluations on `CN-UCB` and `CN-TS`. For each round in `CN-TS`, we draw $M = 10$ samples for each arm. We also present the performances of `CN-TS(M=1)`, which is a special case of `CN-TS` drawing only one sample per arm. We perform synthetic experiments and measure the cumulative regret of each algorithm. In Experiment 1, we compare our algorithms with contextual combinatorial bandits based on a linear assumption: `CombLinUCB` and `CombLinTS` (Wen et al., 2015).

In Experiment 2, we demonstrate the empirical performances of our algorithms as the context dimension d increases. The contexts given to the agent in each round are randomly generated from a unit ball. The dimension of each context is $d = 80$ for Experiment 1, and $d = 40, 80, 120$ for Experiment 2. For each round, the agent of each algorithm chooses $K = 4$ arms among $N = 20$.

Similar to the experiments in Zhou et al. (2020), we assume three unknown score functions

$$h_1(\mathbf{x}_{t,i}) = \mathbf{x}_{t,i}^\top \mathbf{a}\,,$$
$$h_2(\mathbf{x}_{t,i}) = (\mathbf{x}_{t,i}^\top \mathbf{a})^2\,,$$
$$h_3(\mathbf{x}_{t,i}) = \cos(\pi \mathbf{x}_{t,i}^\top \mathbf{a})\,,$$

where $\mathbf{a}$ has the same dimension of the context and is randomly generated from a unit ball and remains fixed during the horizon. We suppose a top-K problem and use the sum of scores $R(S_t, \mathbf{v}_t) = \sum_{i \in S_t} v_{t,i}$ as the reward function. However, as mentioned in Remark 1, the reward function can be any function that satisfies Assumptions 1 and 2. For example, $R(S_t, \mathbf{v}_t)$ can be the quality of positions of a position-based click model (Lattimore & Szepesvari, 2020) or the expected revenue given by a multinomial logit (MNL) choice model (Oh & Iyengar, 2019) although the regret bound under the MNL choice model is not provided under the current theoretical result.

We use regularization parameter $\lambda = 1$ for all methods, confidence bound coefficient $\alpha = 1$ for `CombLinUCB` and $\gamma = 1$ for `CN-UCB`, and exploration variance $\nu = 1$ for `CN-TS`, `CN-TS(M=1)` and `CombLinTS`. To estimate the score of each arm, we design a neural network with depth $L = 2$ and hidden layer width $m = 100$. The number of parameters is $p = md + m = 8100$ for Experiment 1, and $p = 4100, 8100, 12100$ for Experiment 2. The activation function is the rectified linear unit (ReLU). We use the loss function in Eq.(4) and use stochastic gradient descent with a batch of 100 super arms. We train the neural network every 10 rounds. The training epoch is 100, and the learning rate is 0.01.

Experiment 1. We evaluate the cumulative regret of the algorithms for each score function h. For score functions $h_1(\mathbf{x})$ and $h_2(\mathbf{x})$, we set the number of rounds, T, to 2000, while T is set to 4000 for $h_3(\mathbf{x})$. We then present the average results, derived from 20 independent runs for each score function instance. The results are depicted in Figure 1. Our proposed algorithms show significant improvements over those based on linear models. In contrast to linear baselines, the cumulative regrets for `CN-UCB` and `CN-TS` demonstrate a sub-linear trend, even when the score function is quadratic or non-linear. These findings suggest that our algorithms can be readily applied to a diverse range of complex reward functions.

Experiment 2. We present the results of our proposed algorithms for context dimensions $d = 40, 80, 120$. To highlight the advantage of optimistic sampling, we show a comparison between `CN-TS` and `CN-TS(M=1)`. For these experiments, we utilize the quadratic score function $h_2(\mathbf{x})$. The number of rounds, T, is set to 2000 for $d = 40, 80$ and 4000 for $d = 120$. Similar to Experiment 1, the results represent averages derived from 20 independent runs. Figure 2 demonstrates the proficient performance of our algorithms, even as the feature dimension increases. The empirical results suggest a scalability of our algorithms in d that is no greater than linear. Furthermore, when d is large, `CN-TS` exhibits a marginally lower cumulative regret compared to `CN-TS(M=1)`. This observation substantiates our assertion that `CN-TS` ensures a constant probability of optimism by drawing multiple M samples.

6. Conclusion

In this paper, we study a general class of a contextual combinatorial bandit problem, where the model of the score function is unknown. Approximating the score function with deep neural networks, we propose two algorithms: `CN-UCB` and `CN-TS`. We prove that `CN-UCB` achieves $\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{T})$ or $\widetilde{\mathcal{O}}(\sqrt{\widetilde{d}TK})$ regret. For `CN-TS`, we adapt an optimistic sampling technique to ensure the optimism of the sampled combinatorial action, establish a worst-case (frequentist) regret

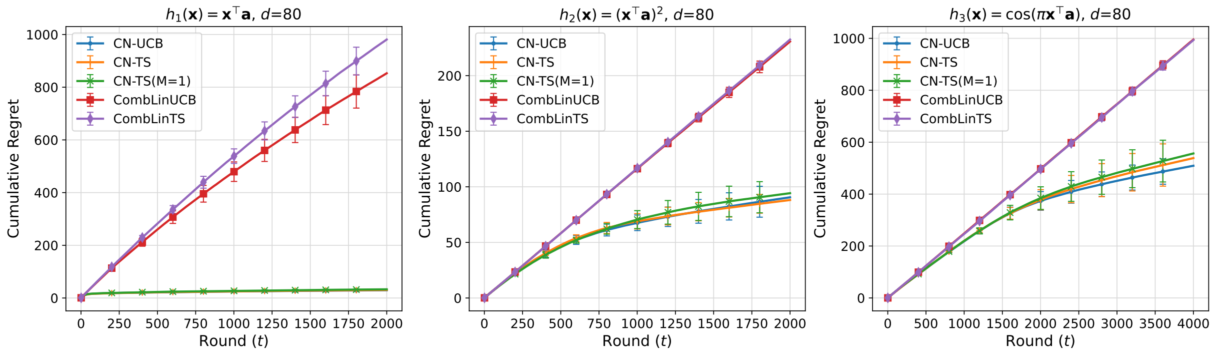

Figure 1. Cumulative regret of `CN-UCB` and `CN-TS` compared with algorithms based on linear models.

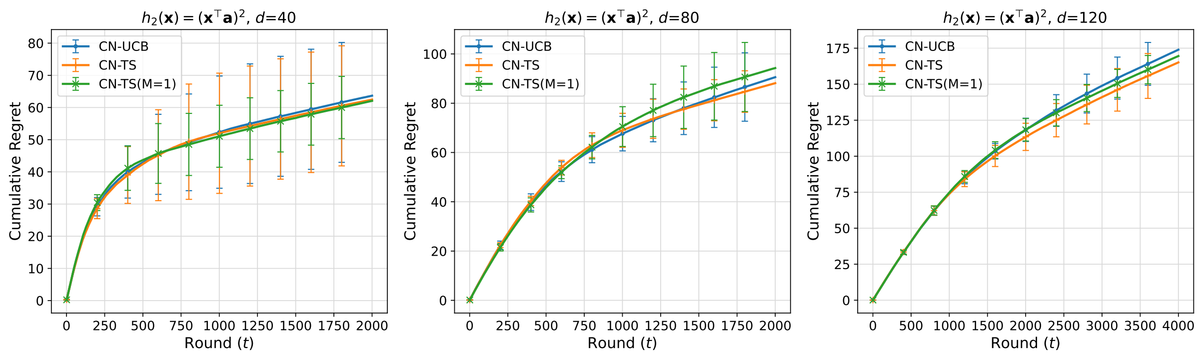

Figure 2. Experiment results of `CN-UCB`, `CN-TS`, and `CN-TS(M=1)` as context dimension d increases.

of $\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{TK})$. To our knowledge, these are the first combinatorial neural bandit algorithms with sub-linear regret guarantees. In particular, `CN-TS` is the first general contextual combinatorial Thompson sampling algorithm with the worst-case regret guarantees. Compared to the benchmark methods, our proposed methods exhibit consistent competitive performances, hence achieving both provable efficiency and practicality.

Acknowledgements

This work was supported by the New Faculty Startup Fund from Seoul National University and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1C1C1006859, No. 2022R1A4A103057912, No. 2021M3E5D2A01024795).

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *International Conference on Machine Learning*, pp. 3–11, 1999.

Abeille, M. and Lazaric, A. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pp. 176–184. PMLR, 2017.

Abramowitz, M. and Stegun, I. A. (eds.). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1964.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135. PMLR, 2013.

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization, 2019b.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International*

Conference on Machine Learning, pp. 322–332. PMLR, 2019.

Audibert, J.-Y., Bubeck, S., and Lugosi, G. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Besson, L. and Kaufmann, E. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10836–10846, 2019.

Chen, L., Xu, J., and Lu, Z. Contextual combinatorial multi-armed bandits with volatile arms and submodular reward. *Advances in Neural Information Processing Systems*, 31, 2018.

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, 2011.

Davis, J. M., Gallego, G., and Topaloglu, H. Assortment optimization under variants of the nested logit model. *Mathematics of Operations Research*, 62(2):250–273, 2014.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2010.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8571–8580, 2018.

Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pp. 767–776, 2015.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.

Lattimore, T. and Szepesvari, C. *Bandit Algorithms*. Cambridge: Cambridge University Press, 2020.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080, 2017.

Li, S. and Zhang, S. Online clustering of contextual cascading bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Li, S., Wang, B., Zhang, S., and Chen, W. Contextual combinatorial cascading bandits. In *International conference on machine learning*, pp. 1245–1253. PMLR, 2016.

Li, S., Lattimore, T., and Szepesvári, C. Online learning to rank with features. In *International Conference on Machine Learning*, pp. 3856–3865. PMLR, 2019.

Nika, A., Elahi, S., and Tekin, C. Contextual combinatorial volatile multi-armed bandit with adaptive discretization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1486–1496. PMLR, 2020.

Oh, M.-h. and Iyengar, G. Thompson sampling for multinomial logit contextual bandits. In *Advances in Neural Information Processing Systems*, pp. 3151–3161, 2019.

Qin, L., Chen, S., and Zhu, X. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 461–469. SIAM, 2014.

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

Rusmevichientong, P., Shen, Z.-J. M., and Shmoys, D. B. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research*, 58(6):1666–1680, 2010.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Wen, Z., Kveton, B., and Ashkan, A. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 1113–1122, 2015.

Zhang, W., Zhou, D., Li, L., and Gu, Q. Neural thompson sampling. In *International Conference on Learning Representation (ICLR)*, 2021.

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.

Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., and Kveton, B. Cascading bandits for large-scale recommendation problems. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pp. 835–844, 2016.

Nomenclature

$v^*_{t,i} = h(\mathbf{x}_{t,i})$	Expected score of arm i at time t
K	The size of super arm
ρ	Sub-Gaussian parameter
N	The number of arms
T	The number of rounds
$\xi_{t,i}$	Sub-Gaussian noise
$R(S, \mathbf{v})$	Reward for the super arm S based on scores $\mathbf{v}$
S^*_t	The offline optimal super arm at time t
S_t	The super arm played at time t
$\widetilde{d}$	The effective dimension
J	The number of gradient descent steps
$\mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})$	Gradient of the neural network for arm i at time t
$\hat{v}_{t,i} = f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})$	Estimated score of arm i at time t
L	The number of hidden layer of the neural network
m	The hidden layer width of the neural network
$\boldsymbol{\theta}_t$	The parameter of the neural network at time t
B	$\geq \max\{\sqrt{2\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}, 1/(22e\sqrt{\pi})\}$
$\mathbf{H}$	The Neural Tangent Kernel matrix
λ	Regularziation parameter
λ_0	NTK matrix parameter
η	Step size for gradient descent
p	The number of parameters of the neural network
$\mathcal{E}^\mu_t$	Event $\forall i \in [N] : \mid f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}) - h(\mathbf{x}_{t,i}) \mid \leq \nu\sigma_{t,i} + \epsilon$
$\mathcal{E}^\sigma_t$	Event $\forall i \in [N] : \mid \widetilde{v}_{t,i} - f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}) \mid \leq \beta_t \nu \sigma_{t,i}$
ν	$= B + \rho\sqrt{\widetilde{d}\log(1 + TN/\lambda) + 2 + 2\log T}$
$\widetilde{p}$	$= 1/(4e\sqrt{\pi})$
$\sigma^2_{t,i}$	$= \lambda \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})^\top \widetilde{\mathbf{Z}}^{-1}_{t-1} \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/m$
β_t	$= \sqrt{4\log t + 2\log K + 2\log M}$
$\widetilde{v}_{t,i}$	$= \max_j \widetilde{v}^{(j)}_{t,i}$
$\widetilde{v}^{(j)}_{t,i}$	j-th sampled score generated by from distribution $\mathcal{N}\left(f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}), \nu\sigma^2_{t,i}\right)$
M	$= \lceil 1 - \log K / \log(1 - \widetilde{p}) \rceil$
ϵ	Approximation error of the neural network
$\widetilde{\mathbf{Z}}_t$	$= \lambda \mathbf{I} + \sum_{k=1}^{t} \sum_{i \in S_k} \mathbf{g}(\mathbf{x}_{k,i}; \boldsymbol{\theta}_{k-1}) \mathbf{g}(\mathbf{x}_{k,i}; \boldsymbol{\theta}_{k-1})^\top / m$

γ_t	Exploration scaling parameter for upper confidence bound
$\Gamma_{1,t}$	$= \sqrt{1 + C_{\Gamma,1} t^{\frac{7}{6}} K L^4 \lambda^{-\frac{7}{6}} m^{-\frac{1}{6}} \sqrt{\log m}}$
$\Gamma_{2,t}$	$= C_{\Gamma,2} t^{\frac{5}{3}} K^{\frac{3}{2}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m}$
$\Gamma_{3,t}$	$= C_{\Gamma,3} t^{\frac{7}{6}} K^{\frac{7}{6}} L^{\frac{7}{2}} \lambda^{-\frac{7}{6}} m^{-\frac{1}{6}} \sqrt{\log m}(1 + \sqrt{tK/\lambda})$
e_t	Offset term added to the upper confidence bound at time t
$\mathbf{Z}_t$	$= \lambda \mathbf{I} + \sum_{k=1}^{t} \sum_{i \in S_k} \mathbf{g}(\mathbf{x}_{k,i}; \boldsymbol{\theta}_{k-1}) \mathbf{g}(\mathbf{x}_{k,i}; \boldsymbol{\theta}_{k-1})^\top / m$
$u_{t,i}$	Upper confidence bound of the expected score for arm i at time t

Appendix

A. Regret Bound for CN-UCB

In this section, we present all the necessary technical lemmas and their proof, followed by the proof of Theorem 1.

A.1. Proof of Lemma 1

We introduce the following technical lemmas which are necessary for proof of Lemma 1.

Lemma 4 (Lemma 5.1 in Zhou et al. (2020))**.** *For any $\delta \in (0,1)$, suppose that there exists a positive constant $\bar{C}$ such that*

$$m \geq \bar{C} T^4 N^4 L^6 \lambda_0^{-1} \log(T^2 N^2 L/\delta)\,.$$

Then, with probability at least $1-\delta$, there exists a $\boldsymbol{\theta}^* \in \mathbb{R}^p$ such that for all $i \in [TN]$,

$$h(\mathbf{x}^k) = \mathbf{g}(\mathbf{x}^k; \boldsymbol{\theta}_0)^\top (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\,,$$
$$\sqrt{m}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2 \leq \sqrt{\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}\,,$$

where $\mathbf{H}$ is the NTK matrix defined in Definition 1 and $\mathbf{h} = [h(\mathbf{x}^k)]_{k=1}^{TN}$.

Lemma 5 (Lemma 4.1 in Cao & Gu (2019))**.** *Suppose that there exist $\bar{C}_1, \bar{C}_2 > 0$ such that for any $\delta \in (0,1)$, τ satisfies*

$$\bar{C}_1 m^{-\frac{3}{2}} L^{-\frac{3}{2}} \left(\log(TNL^2/\delta)\right)^{\frac{3}{2}} \leq \tau \leq \bar{C}_2 L^{-6} \left(\log m\right)^{-\frac{3}{2}}\,.$$

Then, with probability at least $1-\delta$, for all $\widetilde{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}$ satisfying $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \leq \tau$, $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \leq \tau$ and $k \in [TN]$, we have

$$\left| f(\mathbf{x}^k; \widetilde{\boldsymbol{\theta}}) - f(\mathbf{x}^k; \hat{\boldsymbol{\theta}}) - \mathbf{g}(\mathbf{x}^k; \hat{\boldsymbol{\theta}})^\top (\widetilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \right| \leq \bar{C}_3 \tau^{\frac{4}{3}} L^3 \sqrt{m \log m},$$

where $\bar{C}_3 \geq 0$ is an absolute constant.

Lemma 6 (Lemma 5 in Allen-Zhu et al. (2019b))**.** *For any $\delta \in (0,1)$, suppose that there exist $\bar{C}_1, \bar{C}_2 > 0$ such that if τ satisfies*

$$\bar{C}_1 m^{-\frac{3}{2}} L^{-\frac{3}{2}} \max\left\{(\log m)^{-\frac{3}{2}}, (\log(TN/\delta))^{\frac{3}{2}}\right\} \leq \tau \leq \bar{C}_2 L^{-\frac{9}{2}} (\log m)^{-3}\,.$$

Then, with probability at least $1-\delta$, for all $\boldsymbol{\theta}$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq \tau$ and $k \in [TN]$ we have

$$\|\mathbf{g}(\mathbf{x}^k; \boldsymbol{\theta}) - \mathbf{g}(\mathbf{x}^k; \boldsymbol{\theta}_0)\|_2 \ \leq \bar{C}_3 \sqrt{\log m}\, \tau^{\frac{1}{3}} L^3 \|\mathbf{g}(\mathbf{x}^k; \boldsymbol{\theta}_0)\|_2\,,$$

where $\bar{C}_3 > 0$ is an absolute constant.

Lemma 7 (Lemma B.3 in Cao & Gu (2019))**.** *Suppose that there exist $\bar{C}_1, \bar{C}_2 > 0$ such that for any $\delta \in (0,1)$, τ satisfies*

$$\bar{C}_1 m^{-\frac{3}{2}} L^{-\frac{3}{2}} \left(\log(TNL^2/\delta)\right)^{\frac{3}{2}} \leq \tau \leq \bar{C}_2 L^{-6} \left(\log m\right)^{-\frac{3}{2}}\,.$$

Then with probability at least $1-\delta$, for any $\boldsymbol{\theta}$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq \tau$ and $k \in [TN]$ we have

$$\|\mathbf{g}(\mathbf{x}^k; \boldsymbol{\theta})\|_F \leq \bar{C}_3 \sqrt{mL}$$

where $\bar{C}_3 > 0$ is an absolute constant.

Proof of Lemma 1. First of all, note that because m satisfies Condition 1, the required conditions in Lemma 4-7 are satisfied. For any $t \in [T], i \in [N]$, by definition of $u_{t,i}$ and $v^*_{t,i}$, we have

$$\begin{aligned}
\left|u_{t,i} - v^*_{t,i}\right| &= \left| f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}) + \gamma_{t-1} \left\|\mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|_{\mathbf{Z}_t^{-1}} - h(\mathbf{x}_{t,i}) \right| \\
&= \left| f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}) + \gamma_{t-1} \left\|\mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|_{\mathbf{Z}_{t-1}^{-1}} - \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_0)^\top (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) \right| \\
&\leq \underbrace{\left| f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}) - \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_0)^\top (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) \right|}_{I_0} + \underbrace{\gamma_{t-1} \left\|g(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|_{\mathbf{Z}_{t-1}^{-1}}}_{I_1}
\end{aligned}$$

where the second equality holds due to Lemma 4, and the inequality follows from the triangle inequality.

For I_0, we have

$$\begin{aligned}
I_0 &= \left|f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) - \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 + \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_{t-1})\right| \\
&= \left|f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_0) - \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_0) - \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1})\right| \\
&\le \underbrace{\left|f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_0) - \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_0)\right|}_{I_2} + \underbrace{\left|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1})\right|}_{I_3}
\end{aligned} \tag{6}$$

where the second equality holds due to the initial condition of f, i.e., $f(\mathbf{x};\boldsymbol{\theta}_0) = 0$ for all $\mathbf{x}$, and the inequality comes from the triangle inequality.

To bound I_2, we have

$$\begin{aligned}
I_2 &= \left|f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_0) - \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_0)\right| \\
&\le C_3' \tau^{\frac{4}{3}} L^3 \sqrt{m \log m} \\
&= C_3 t^{\frac{2}{3}} K^{\frac{2}{3}} \lambda^{-\frac{2}{3}} m^{-\frac{1}{6}} \sqrt{\log m}
\end{aligned}$$

where the first inequality follows from Lemma 5 for some constant $C_3' > 0$, and the second equality is due to setting τ of Lemma 5 as $2\sqrt{tK/(m\lambda)}$ of Lemma 11, i.e., $\tau = 2\sqrt{tK/(m\lambda)}$.

To bound I_3, we have

$$I_3 = \left|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1})\right| \le \|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)\|_{\mathbf{Z}_{t-1}^{-1}} \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1}\|_{\mathbf{Z}_{t-1}} \le \frac{\gamma_{t-1}}{\sqrt{m}} \|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)\|_{\mathbf{Z}_{t-1}^{-1}}$$

where the first inequality holds due to the Cauchy-Schwarz inequality, and the second inequality follows from Lemma 11.

Combining the results, we have

$$\begin{aligned}
\left|u_{t,i} - v_{t,i}^*\right| &\le I_2 + I_3 + I_1 \\
&\le C_3 t^{\frac{2}{3}} K^{\frac{2}{3}} \lambda^{-\frac{2}{3}} m^{-\frac{1}{6}} \sqrt{\log m} + \frac{\gamma_{t-1}}{\sqrt{m}} \|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)\|_{\mathbf{Z}_{t-1}^{-1}} + \frac{\gamma_{t-1}}{\sqrt{m}} \|g(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\|_{\mathbf{Z}_{t-1}^{-1}} \\
&= C_3 t^{\frac{2}{3}} K^{\frac{2}{3}} \lambda^{-\frac{2}{3}} m^{-\frac{1}{6}} \sqrt{\log m} + \frac{\gamma_{t-1}}{\sqrt{m}} \underbrace{\left(\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)\|_{\mathbf{Z}_{t-1}^{-1}} + \|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\|_{\mathbf{Z}_{t-1}^{-1}}\right)}_{I_4}.
\end{aligned}$$

Now I_4 can be bounded as

$$\begin{aligned}
I_4 &= \|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0) + \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) - \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\|_{\mathbf{Z}_{t-1}^{-1}} + \|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\|_{\mathbf{Z}_{t-1}^{-1}} \\
&\le \|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0) - \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\|_{\mathbf{Z}_{t-1}^{-1}} + 2\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\|_{\mathbf{Z}_{t-1}^{-1}} \\
&\le \frac{1}{\sqrt{\lambda}} \|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0) - \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\|_2 + 2\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\|_{\mathbf{Z}_{t-1}^{-1}} \\
&\le \frac{1}{\sqrt{\lambda}} C_2' \sqrt{\log m} \left(2\sqrt{tK/(m\lambda)}\right)^{\frac{1}{3}} L^3 \|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)\|_2 + 2\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\|_{\mathbf{Z}_{t-1}^{-1}} \\
&\le C_2 t^{\frac{1}{6}} K^{\frac{1}{6}} \lambda^{-\frac{2}{3}} L^{\frac{7}{2}} m^{\frac{1}{3}} \sqrt{\log m} + 2\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\|_{\mathbf{Z}_{t-1}^{-1}}
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality holds due to the property $\|\mathbf{x}\|_{\mathbf{Z}_{t-1}^{-1}} \le \frac{1}{\sqrt{\lambda}}\|\mathbf{x}\|_2$, the third inequality follows from Lemma 6 with $\tau = 2\sqrt{tK/(m\lambda)}$ in Lemma 11, and the last inequality holds due to Lemma 7.

Finally, by taking a union bound about δ, with probability at least $1 - 5\delta$, we have

$$\begin{aligned}
\left|u_{t,i} - v_{t,i}^*\right| &\le C_3 t^{\frac{2}{3}} K^{\frac{2}{3}} \lambda^{-\frac{2}{3}} m^{-\frac{1}{6}} \sqrt{\log m} + \frac{\gamma_{t-1}}{\sqrt{m}} I_4 \\
&\le 2\gamma_{t-1} \left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|_{\mathbf{Z}_{t-1}^{-1}} + C_2 \gamma_{t-1} t^{\frac{1}{6}} K^{\frac{1}{6}} \lambda^{-\frac{2}{3}} L^{\frac{7}{2}} m^{-\frac{1}{6}} \sqrt{\log m} \\
&\quad + C_3 t^{\frac{2}{3}} K^{\frac{2}{3}} \lambda^{-\frac{2}{3}} m^{-\frac{1}{6}} \sqrt{\log m}.
\end{aligned}$$

In particular, if we define

$$e_t = C_2\gamma_{t-1}t^{\frac{1}{6}}K^{\frac{1}{6}}\lambda^{-\frac{2}{3}}L^{\frac{7}{2}}m^{-\frac{1}{6}}\sqrt{\log m} + C_3 t^{\frac{2}{3}}K^{\frac{2}{3}}\lambda^{-\frac{2}{3}}m^{-\frac{1}{6}}\sqrt{\log m}$$

and we replace δ with $\delta/5$, then we have the desired result. $\square$

A.2. Proof of Corollary 1

Proof of Corollary 1. Suppose that Lemma 1 holds. Let us denote

$$\bar{u}_{t,i} = u_{t,i} + \underbrace{C_2\gamma_{t-1}t^{\frac{1}{6}}K^{\frac{1}{6}}\lambda^{-\frac{2}{3}}L^{\frac{7}{2}}m^{-\frac{1}{6}}\sqrt{\log m} + C_3 t^{\frac{2}{3}}K^{\frac{2}{3}}\lambda^{-\frac{2}{3}}m^{-\frac{1}{6}}\sqrt{\log m}}_{e_t}.$$

Then, we have

$$\begin{aligned}
\bar{u}_{t,i} - v^*_{t,i} &= f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) + \gamma_{t-1}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|_{\mathbf{Z}_{t-1}^{-1}} + e_t - \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) \\
&\geq -\underbrace{\left|f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) - \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\right|}_{I_0} + \gamma_{t-1}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|_{\mathbf{Z}_{t-1}^{-1}} + e_t \\
&\geq -\underbrace{C_3 t^{\frac{2}{3}}K^{\frac{2}{3}}\lambda^{-\frac{2}{3}}m^{-\frac{1}{6}}\sqrt{\log m}}_{I_2} - \underbrace{\frac{\gamma_{t-1}}{\sqrt{m}}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)\right\|_{\mathbf{Z}_{t-1}^{-1}}}_{I_3} + \frac{\gamma_{t-1}}{\sqrt{m}}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\right\|_{\mathbf{Z}_{t-1}^{-1}} \\
&\quad + e_t + \frac{\gamma_{t-1}}{\sqrt{m}}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\right\|_{\mathbf{Z}_{t-1}^{-1}} - \frac{\gamma_{t-1}}{\sqrt{m}}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\right\|_{\mathbf{Z}_{t-1}^{-1}} \\
&= -C_3 t^{\frac{2}{3}}K^{\frac{2}{3}}\lambda^{-\frac{2}{3}}m^{-\frac{1}{6}}\sqrt{\log m} + 2\frac{\gamma_{t-1}}{\sqrt{m}}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\right\|_{\mathbf{Z}_{t-1}^{-1}} + e_t \\
&\quad - \frac{\gamma_{t-1}}{\sqrt{m}}\underbrace{\left(\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)\right\|_{\mathbf{Z}_{t-1}^{-1}} + \left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\right\|_{\mathbf{Z}_{t-1}^{-1}}\right)}_{I_4} \\
&\geq 0\,,
\end{aligned}$$

where the first equation comes from Lemma 4 and the second inequality follows from Eq.(6). $\square$

A.3. Proof of Lemma 2

The following lemma is necessary for our proof.

Lemma 8 (Lemma B.7 in Zhang et al. (2021))**.** *For any $t \in [T]$, suppose that there exists $\bar{C} > 0$ such that the network width m satisfies*

$$m \geq \bar{C}T^6N^6L^6\log(TLN/\delta).$$

Then with probability at least $1-\delta$,

$$\log\det(\mathbf{I} + \lambda^{-1}\mathbf{K}_t) \leq \log\det(\mathbf{I} + \lambda^{-1}\mathbf{H}) + 1\,,$$

where $\mathbf{K}_t = \bar{\mathbf{J}}_t^\top\bar{\mathbf{J}}_t/m$, $\bar{\mathbf{J}}_t = [\mathbf{g}(\mathbf{x}_{1,a_{11}};\boldsymbol{\theta}_0), \cdots, \mathbf{g}(\mathbf{x}_{t,a_{tK}};\boldsymbol{\theta}_0)] \in \mathbb{R}^{p\times tK}$, and a_{tk} means k-th action in the super arm S_t at time t, i.e., $S_t := \{a_{t1}, \ldots, a_{tK}\}$.

Proof. Note that we have

$$\begin{aligned}
\det(\mathbf{Z}_T) &= \det\left(\mathbf{Z}_{T-1} + \sum_{i\in S_T} \mathbf{g}(\mathbf{x}_{T,i};\boldsymbol{\theta}_{T-1})\mathbf{g}(\mathbf{x}_{T,i};\boldsymbol{\theta}_{T-1})^\top/m\right) \\
&= \det\left(\mathbf{Z}_{T-1}^{\frac{1}{2}}\left(\mathbf{I} + \sum_{i\in S_T} \mathbf{Z}_{T-1}^{-\frac{1}{2}}(\mathbf{g}(\mathbf{x}_{T,i};\boldsymbol{\theta}_{T-1})/\sqrt{m})(\mathbf{g}(\mathbf{x}_{T,i};\boldsymbol{\theta}_{T-1})/\sqrt{m})^\top \mathbf{Z}_{T-1}^{-\frac{1}{2}}\right)\mathbf{Z}_{T-1}^{\frac{1}{2}}\right) \\
&= \det(\mathbf{Z}_{T-1})\cdot\det\left(\mathbf{I} + \sum_{i\in S_T}\left(\mathbf{Z}_{T-1}^{-\frac{1}{2}}\mathbf{g}(\mathbf{x}_{T,i};\boldsymbol{\theta}_{T-1})/\sqrt{m}\right)\left(\mathbf{Z}_{T-1}^{-\frac{1}{2}}\mathbf{g}(\mathbf{x}_{T,i};\boldsymbol{\theta}_{T-1})/\sqrt{m}\right)^\top\right) \\
&= \det(\mathbf{Z}_{T-1})\cdot\left(1 + \sum_{i\in S_T}\left\|\mathbf{g}(\mathbf{x}_{T,i};\boldsymbol{\theta}_{T-1})/\sqrt{m}\right\|^2_{\mathbf{Z}_{T-1}^{-1}}\right) \\
&= \det(\mathbf{Z}_0)\prod_{t=1}^{T}\left(1 + \sum_{i\in S_t}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|^2_{\mathbf{Z}_{t-1}^{-1}}\right).
\end{aligned}$$

Then, we have

$$\log\frac{\det(\mathbf{Z}_T)}{\det(\mathbf{Z}_0)} = \sum_{t=1}^{T}\log\left(1 + \sum_{i\in S_t}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|^2_{\mathbf{Z}_{t-1}^{-1}}\right).$$

On the other hand, for any $t\in[T]$, we have

$$\begin{aligned}
\sum_{i\in S_t}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|^2_{\mathbf{Z}_{t-1}^{-1}} &\le \sum_{i\in S_t}\frac{1}{\lambda}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})\right\|_2^2/m \\
&\le \sum_{i\in S_t}\frac{1}{\lambda m}\left(C_2\sqrt{mL}\right)^2 \\
&\le 1,
\end{aligned}$$

where the first inequality comes from the property $\|\mathbf{x}\|^2_{\mathbf{A}^{-1}} \le \|\mathbf{x}\|_2^2/\lambda_{\min}(\mathbf{A})$ for any positive definite matrix A, the constant C_2 of the second inequality can be derived by Lemma 7, and the last inequality holds due to the assumption of λ. Then using the inequality, $x \le 2\log(1+x)$ for any $x\in[0,1]$, we have

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i\in S_t}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|^2_{\mathbf{Z}_{t-1}^{-1}} &\le 2\sum_{t=1}^{T}\log\left(1 + \sum_{i\in S_t}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|^2_{\mathbf{Z}_{t-1}^{-1}}\right) \\
&\le 2\left|\log\frac{\det\mathbf{Z}_T}{\det\lambda\mathbf{I}}\right| \\
&\le 2\left|\log\frac{\det\bar{\mathbf{Z}}_T}{\det\lambda\mathbf{I}}\right| + C_3 T^{\frac{5}{3}}K^{\frac{5}{3}}L^4\lambda^{-\frac{1}{6}}m^{-\frac{1}{6}}\sqrt{\log m},
\end{aligned}$$

where the last inequality holds due to Lemma 13 for some $C_3 > 0$. Furthermore, since we have

$$\begin{aligned}
\log\frac{\det\bar{\mathbf{Z}}_T}{\det\lambda\mathbf{I}} &= \log\det\left(\bar{\mathbf{Z}}_T(\lambda\mathbf{I})^{-1}\right) \\
&= \log\det\left(\mathbf{I} + \sum_{t=1}^{T}\sum_{i\in S_t}\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)^\top/(m\lambda)\right) \\
&= \log\det\left(\mathbf{I} + \lambda^{-1}\bar{\mathbf{J}}_T\bar{\mathbf{J}}_T^\top/m\right) \\
&= \log\det\left(\mathbf{I} + \lambda^{-1}\bar{\mathbf{J}}_T^\top\bar{\mathbf{J}}_T/m\right) \\
&= \log\det\left(\mathbf{I} + \lambda^{-1}\mathbf{K}_T\right) \\
&\le \log\det\left(\mathbf{I} + \lambda^{-1}\mathbf{H}\right) + 1 \\
&= \widetilde{d}\log(1 + TN/\lambda) + 1,
\end{aligned} \tag{7}$$

where the first, second equation and the first inequality holds naively, the third equality uses the definition of $\bar{\mathbf{J}}_t$, the fourth equality holds since for any matrix $\mathbf{A} \in \mathbb{M}_n(\mathbb{R})$ the nonzero eigenvalues of $\mathbf{I} + \mathbf{A}\mathbf{A}^\top$ and $\mathbf{I} + \mathbf{A}^\top\mathbf{A}$ are same, which means $\det(\mathbf{I} + \mathbf{A}\mathbf{A}^\top) = \det(\mathbf{I} + \mathbf{A}^\top\mathbf{A})$, the first inequality follows from Lemma 8, and the last equality uses the definition of effective dimension in Definition 2. Finally, by taking a union bound about δ, with probability at least $1 - 2\delta$, we have

$$\sum_{t=1}^{T} \sum_{i \in S_t} \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|^2_{\mathbf{Z}_{t-1}^{-1}} \leq 2\tilde{d} \log(1 + TN/\lambda) + 2 + C_3 T^{\frac{5}{3}} K^{\frac{5}{3}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m}\,.$$

By replacing δ with $\delta/2$, we have the desired result. □

A.4. Proof of Theorem 1

Proof of Theorem 1. We define the following event:

$$\begin{aligned}
&\mathfrak{E}_1 := \left\{ |u_{t,i} - h(\mathbf{x}_{t,i})| \leq 2\gamma_{t-1} \|\mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}} + e_t, \forall i \in [N], 1 \leq t \leq T \right\}, \\
&\mathfrak{E}_2 := \left\{ \sum_{t=1}^{T} \sum_{i \in S_t} \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|^2_{\mathbf{Z}_{t-1}^{-1}} \leq 2\tilde{d} \log(1 + TN/\lambda) + 2 + C_3 T^{\frac{5}{3}} K^{\frac{5}{3}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} \right\}, \\
&\mathfrak{E} := \mathfrak{E}_1 \cap \mathfrak{E}_2\,.
\end{aligned}$$

Then, we decompose the cumulative expected regret into two components: when $\mathfrak{E}$ occurs and when $\mathfrak{E}$ does not happen.

$$\begin{aligned}
\mathcal{R}(T) &= \mathbb{E}\left[\sum_{t=1}^{T} \left(R(S_t^*, \mathbf{v}_t^*) - R(S_t, \mathbf{v}_t^*) \right) \mathbb{I}(\mathfrak{E}) \right] + \mathbb{E}\left[\sum_{t=1}^{T} \left(R(S_t^*, \mathbf{v}_t^*) - R(S_t, \mathbf{v}_t^*) \right) \mathbb{I}(\mathfrak{E}^c) \right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \underbrace{\left(R(S_t^*, \mathbf{v}_t^*) - R(S_t, \mathbf{v}_t^*) \right) \mathbb{I}(\mathfrak{E})}_{\mathcal{I}_t} \right] + \mathcal{O}(1)\,,
\end{aligned}$$

where the inequality holds since we have $\mathfrak{E}$ holds with probability at least $1 - T^{-1}$ by Lemma 1 and Lemma 2. To bound $\mathcal{I}_t$, we have

$$\begin{aligned}
\mathcal{I}_t &\leq R(S_t^*, \mathbf{v}_t^*) - R(S_t, \mathbf{v}_t^*) \\
&\leq R(S_t^*, \mathbf{u}_t + \mathbf{e}_t) - R(S_t, \mathbf{v}_t^*) \qquad (8) \\
&\leq R(S_t, \mathbf{u}_t + \mathbf{e}_t) - R(S_t, \mathbf{v}_t^*) \\
&\leq C_\ell \sqrt{\sum_{i \in S_t} \left(u_{t,i} + e_t - v_{t,i}^* \right)^2} \\
&\leq C_\ell \sqrt{\sum_{i \in S_t} \left(2\gamma_{t-1} \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|_{\mathbf{Z}_{t-1}^{-1}} + 2e_t \right)^2} \\
&\leq 4C_\ell \sqrt{\sum_{i \in S_t} \left(\max\left\{ \gamma_{t-1} \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|_{\mathbf{Z}_{t-1}^{-1}}, e_t \right\} \right)^2}\,, \qquad (9)
\end{aligned}$$

where C_ℓ is a Lipschitz constant, the first inequality holds due to the monotonicity of the reward function, the second inequality comes from the choice of the oracle, i.e., $S_t = \mathbb{O}_{\mathcal{S}}(\mathbf{u}_t + \mathbf{e}_t)$, the third inequality follows from the Lipschitz continuity of the reward function, the fourth inequality comes from Lemma 1 and the last inequality holds due to the property, $a + b \leq 2\max\{a, b\}$.

On the other hand, if we denote $\mathcal{A}_i := \gamma_{t-1} \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|_{\mathbf{Z}_{t-1}^{-1}}$, then we have

$$\begin{aligned}
\sqrt{\sum_{i \in S_t} \left(\max\left\{ \gamma_{t-1} \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|_{\mathbf{Z}_{t-1}^{-1}}, e_t \right\} \right)^2} &= \sqrt{\sum_{\mathcal{A}_i \geq e_t} \mathcal{A}_i^2 + \sum_{\mathcal{A}_i < e_t} e_t^2} \\
&\leq \sqrt{\sum_{i \in S_t} \mathcal{A}_i^2 + \sum_{i \in S_t} e_t^2} \\
&\leq \sqrt{\sum_{i \in S_t} \mathcal{A}_i^2} + \sqrt{\sum_{i \in S_t} e_t^2} \\
&= \sqrt{\sum_{i \in S_t} \mathcal{A}_i^2} + \sqrt{K} e_t \, . \qquad (10)
\end{aligned}$$

By substituting Eq.(10) into Eq.(9), we have

$$R(S_t^*, \mathbf{v}_t^*) - R(S_t, \mathbf{v}_t^*) \leq 4C_\ell \left(\sqrt{\sum_{i \in S_t} \gamma_{t-1}^2 \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|_{\mathbf{Z}_{t-1}^{-1}}^2} + \sqrt{K} e_t \right) \qquad (11)$$

Therefore, by summing Eq.(11) over all $t \in [T]$, we have

$$\begin{aligned}
\mathcal{R}(T) &\leq 4C_\ell \sum_{t=1}^{T} \left(\sqrt{\sum_{i \in S_t} \gamma_{t-1}^2 \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|_{\mathbf{Z}_{t-1}^{-1}}^2} + \sqrt{K} e_t \right) \\
&\leq 4C_\ell \gamma_T \sum_{t=1}^{T} \sqrt{\sum_{i \in S_t} \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|_{\mathbf{Z}_{t-1}^{-1}}^2} + 4C_\ell \sqrt{K} T e_T \\
&\leq 4C_\ell \gamma_T \sqrt{T \sum_{t=1}^{T} \sum_{i \in S_t} \left\| \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})/\sqrt{m} \right\|_{\mathbf{Z}_{t-1}^{-1}}^2} + 4C_\ell \sqrt{K} T e_T \\
&\leq 4C_\ell \gamma_T \sqrt{T \left(2\widetilde{d} \log(1 + TN/\lambda) + 2 + \bar{C}_1 T^{\frac{5}{3}} K^{\frac{5}{3}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} \right)} + 4C_\ell \sqrt{K} T e_T \, , \qquad (12)
\end{aligned}$$

where the second inequality holds since $\gamma_t \leq \gamma_T$ and $e_t \leq e_T$, the third inequality follows from the Cauchy-Schwarz inequality and the last inequality comes from Lemma 2 with an absolute constant $\bar{C}_1 > 0$.

Meanwhile, we bound γ_T as follows:

$$\begin{aligned}
\gamma_T &= \Gamma_{1,T} \left(\rho \sqrt{\log \frac{\det \mathbf{Z}_T}{\det \lambda \mathbf{I}} + C_{\Gamma,2} T^{\frac{5}{3}} K^{\frac{5}{3}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} - 2 \log \delta} + \sqrt{\lambda} B \right) \\
&\quad + (\lambda + \bar{C}_2 TKL) \left((1 - \eta m \lambda)^{\frac{J}{2}} \sqrt{TK/\lambda} + \Gamma_{3,T} \right) \\
&\leq \Gamma_{1,T} \left(\rho \sqrt{\log \frac{\det \bar{\mathbf{Z}}_T}{\det \lambda \mathbf{I}} + 2C_{\Gamma,2} T^{\frac{5}{3}} K^{\frac{5}{3}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} - 2 \log \delta} + \sqrt{\lambda} B \right) \\
&\quad + (\lambda + \bar{C}_2 TKL) \left((1 - \eta m \lambda)^{\frac{J}{2}} \sqrt{TK/\lambda} + \Gamma_{3,T} \right) \\
&\leq \Gamma_{1,T} \left(\rho \sqrt{\widetilde{d} \log(1 + TN/\lambda) + 1 + 2C_{\Gamma,2} T^{\frac{5}{3}} K^{\frac{5}{3}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} - 2 \log \delta} + \sqrt{\lambda} B \right) \\
&\quad + (\lambda + \bar{C}_2 TKL) \left((1 - \eta m \lambda)^{\frac{J}{2}} \sqrt{TK/\lambda} + \Gamma_{3,T} \right) \qquad (13)
\end{aligned}$$

where the fist inequality holds due to Lemma 13, the second inequality holds due to Eq.(7).

Note that by setting $\eta = C_1(TKmL + m\lambda)^{-1}$ and $J = 2\log\left(\frac{\sqrt{\lambda/TK}}{\lambda+\bar{C}_2TKL}\right)\frac{TKL}{C_1\lambda}$, we have

$$(\lambda + \bar{C}_2TKL)(1-\eta m\lambda)^{\frac{J}{2}}\sqrt{TK/\lambda} \le 1\,.$$

By choosing sufficiently large m such that

$$\begin{aligned}
&\Gamma_{1,T} = \sqrt{1 + C_{\Gamma,1}T^{\frac{7}{6}}K^{\frac{7}{6}}L^4\lambda^{-\frac{7}{6}}m^{-\frac{1}{6}}\sqrt{\log m}} \le 2\\
&\Gamma_{2,T} = C_{\Gamma,2}T^{\frac{5}{3}}K^{\frac{5}{3}}L^4\lambda^{-\frac{1}{6}}m^{-\frac{1}{6}}\sqrt{\log m} \le 1\,,\\
&C_1T^{\frac{5}{3}}K^{\frac{3}{2}}L^4\lambda^{-\frac{1}{6}}m^{-\frac{1}{6}}\sqrt{\log m} \le 1\,,\\
&(\lambda + \bar{C}_2TKL)\Gamma_{3,T} = (\lambda + \bar{C}_2TKL)C_{\Gamma,3}T^{\frac{7}{6}}K^{\frac{7}{6}}L^{\frac{7}{2}}\lambda^{-\frac{7}{6}}m^{-\frac{1}{6}}\sqrt{\log m}(1+\sqrt{TK/\lambda}) \le 1\,,\\
&Te_T \le \gamma_T + 1 \le 2\rho\sqrt{\widetilde{d}\log(1+TN/\lambda) + 3 - 2\log\delta} + 2\sqrt{\lambda}B + 3\,,
\end{aligned}$$

and combining all the results, $\mathcal{R}(T)$ can be bounded by

$$\begin{aligned}
\mathcal{R}(T) \le\, & 4C_\ell\sqrt{T\left(2\widetilde{d}\log(1+TN/\lambda)+3\right)}\left[2\rho\sqrt{\widetilde{d}\log(1+TN/\lambda)+3-2\log\delta} + 2\sqrt{\lambda}B + 2\right]\\
&+ 4C_\ell\sqrt{K}\left(2\rho\sqrt{\widetilde{d}\log(1+TN/\lambda)+3-2\log\delta} + 2\sqrt{\lambda}B + 3\right).
\end{aligned}$$

□

B. Regret Bound for CN-TS

B.1. Proof Lemma 3

proof of Lemma 3. For given $\mathcal{F}_t$, since $\widetilde{v}_{t,i}^{(j)} \sim \mathcal{N}(f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}), \nu^2\sigma_{t,i}^2)$, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_j \widetilde{v}_{t,i}^{(j)} + \epsilon > h(\mathbf{x}_{t,i}) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) = 1 - \mathbb{P}\left(\widetilde{v}_{t,i}^{(j)} + \epsilon \le h(\mathbf{x}_{t,i}), \forall j \in [M] \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right)\\
&= 1 - \mathbb{P}\left(\frac{\widetilde{v}_{t,i}^{(j)} - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) + \epsilon}{\nu\sigma_{t,i}} \le \frac{h(\mathbf{x}_{t,i}) - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})}{\nu\sigma_{t,i}}, \forall j \in [M] \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right)\\
&\ge 1 - \mathbb{P}\left(\frac{\widetilde{v}_{t,i}^{(j)} - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) + \epsilon}{\nu\sigma_{t,i}} \le \frac{|h(\mathbf{x}_{t,i}) - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})|}{\nu\sigma_{t,i}}, \forall j \in [M] \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right)\\
&= 1 - \mathbb{P}\left(\frac{\widetilde{v}_{t,i}^{(j)} - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})}{\nu\sigma_{t,i}} \le \frac{|h(\mathbf{x}_{t,i}) - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})| - \epsilon}{\nu\sigma_{t,i}}, \forall j \in [M] \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right)\\
&= 1 - \mathbb{P}\left(Z_j \le \frac{|h(\mathbf{x}_{t,i}) - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})| - \epsilon}{\nu\sigma_{t,i}}, \forall j \in [M] \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right)\,,
\end{aligned}$$

where the first inequality is due to $a \le |a|$, for the last equality we denote Z_j as a standard normal random variable. Note that under the event $\mathcal{E}_t^\mu$, we have $|f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) - h(\mathbf{x}_{t,i})| \le \nu\sigma_{t,i} + \epsilon$ for all $i \in [N]$. Hence, under the event $\mathcal{E}_t^\mu$,

$$\frac{|h(\mathbf{x}_{t,i}) - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})| - \epsilon}{\nu\sigma_{t,i}} \le \frac{\nu\sigma_{t,i} + \epsilon - \epsilon}{\nu\sigma_{t,i}} = 1\,.$$

Then, it follows that

$$\mathbb{P}\left(\max_j \widetilde{v}_{t,i}^{(j)} + \epsilon > h(\mathbf{x}_{t,i}) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \ge 1 - \left[\mathbb{P}\left(Z \le 1\right)\right]^M\,.$$

Using the anti-concentration inequality in Lemma 9, we have $\mathbb{P}(Z \leq 1) \leq 1 - \widetilde{p}$ where $\widetilde{p} := 1/(4e\sqrt{\pi})$. Then finally we have

$$\begin{aligned}
\mathbb{P}\Big(R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) \geq R(S_t^*, \mathbf{v}_t^*) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\Big) &\geq \mathbb{P}\Big(R(S_t^*, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) \geq R(S_t^*, \mathbf{v}_t^*) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\Big) \\
&\geq \mathbb{P}\Big(\widetilde{v}_{t,i} + \epsilon \geq h(\mathbf{x}_{t,i}), \forall i \in S_t^* \mid \mathcal{F}_t, \mathcal{E}_t^\mu\Big) \\
&= \prod_{i \in S_t} \mathbb{P}\Big(\widetilde{v}_{t,i} + \epsilon \geq h(\mathbf{x}_{t,i}) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\Big) \\
&\geq \Big(1 - [\mathbb{P}(Z \leq 1)]^M\Big)^K \\
&\geq \left[1 - (1 - \widetilde{p})^M\right]^K \\
&\geq 1 - K\left(1 - \widetilde{p}\right)^M \\
&\geq 1 - (1 - \widetilde{p}) \\
&= \widetilde{p}\,,
\end{aligned}$$

where the first inequality holds due to the choice of the oracle, the second inequality comes from the monotonicity of the reward function, the third inequality uses the Bernoulli's inequality, and the last inequality comes from the choice of $M = \lceil 1 - \frac{\log K}{\log(1-\widetilde{p})} \rceil$, which means $(1 - \widetilde{p})^M \leq \frac{1}{K}(1 - \widetilde{p})$. □

B.2. Proof of Theorem 2

Proof of Theorem 2. First of all, we decompose the expected cumulative regret as follows:

$$\mathcal{R}(T) = \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[R(S_t^*, \mathbf{v}_t^*) - R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})\right]}_{\mathcal{R}_1(T)} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) - R(S_t, \mathbf{v}_t^*)\right]}_{\mathcal{R}_2(T)}\,.$$

From now on, we derive the bounds for $\mathcal{R}_1(T)$ and $\mathcal{R}_2(T)$ respectively.

Bounding $\mathcal{R}_2(T)$

First we decompose $\mathcal{R}_2(T)$:

$$\mathcal{R}_2(T) = \sum_{t=1}^{T} \mathbb{E}\Bigg[\underbrace{R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) - R(S_t, \hat{\mathbf{v}}_t + \boldsymbol{\epsilon})}_{I_2}\Bigg] + \sum_{t=1}^{T} \mathbb{E}\Bigg[\underbrace{R(S_t, \hat{\mathbf{v}}_t + \boldsymbol{\epsilon}) - R(S_t, \mathbf{v}_t^*)}_{I_1}\Bigg]\,.$$

For I_1, we have

$$
\begin{aligned}
|R(S_t, \hat{\mathbf{v}}_t + \boldsymbol{\epsilon}) - R(S_t, \mathbf{v}_t^*)| &\leq C_0^{(1)} \sqrt{\sum_{i \in S_t} \left(f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}) + \epsilon - h(\mathbf{x}_{t,i}) \right)^2} \\
&\leq C_0^{(1)} \sqrt{\sum_{i \in S_t} \left(\nu \sigma_{t,i} + 2\epsilon \right)^2} \\
&\leq C_0^{(1)} \sqrt{\sum_{i \in S_t} \left(2 \max\{\nu \sigma_{t,i}, 2\epsilon\} \right)^2} \\
&\leq 2C_0^{(1)} \sqrt{\sum_{i \in S_t} (\nu \sigma_{t,i})^2 + \sum_{i \in S_t} 4\epsilon^2} \\
&\leq 2C_0^{(1)} \left[\sqrt{\sum_{i \in S_t} (\nu \sigma_{t,i})^2} + \sqrt{\sum_{i \in S_t} 4\epsilon^2} \right] \\
&= 2C_0^{(1)} \left[\nu \sqrt{\sum_{i \in S_t} \sigma_{t,i}^2} + 2\epsilon \sqrt{K} \right] ,
\end{aligned}
$$

where the first inequality holds due to the Lipschitz continuity for a constant $C_0^{(1)} > 0$, the second inequality holds due to the event $\mathcal{E}_t^{\mu}$ holds with high probability, the third inequality follows from the property that $a + b \leq 2\max\{a, b\}$, and the last inequality uses the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$.

On the other hand, for I_2 we have

$$
\begin{aligned}
|R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) - R(S_t, \hat{\mathbf{v}}_t + \boldsymbol{\epsilon})| &\leq C_0^{(2)} \sqrt{\sum_{i \in S_t} \left(\widetilde{v}_{t,i} - f(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}) \right)^2} \\
&\leq C_0^{(2)} \sqrt{\sum_{i \in S_t} \beta_t^2 \nu^2 \sigma_{t,i}^2} \\
&= C_0^{(2)} \beta_t \nu \sqrt{\sum_{i \in S_t} \sigma_{t,i}^2} \, ,
\end{aligned}
$$

where the first inequality holds for some Lipschitz continuity constant $C_0^{(2)} > 0$, the second inequality holds due to the event $\mathcal{E}_t^{\sigma}$ holds with high probability.

By combining the bounds of I_1 and I_2, we derive the bound for $\mathcal{R}_2(T)$ as follows:

$$
\begin{aligned}
\mathcal{R}_2(T) &\le 2C_0 \sum_{t=1}^{T} \mathbb{E}\left[\nu\sqrt{\sum_{i\in S_t}\sigma_{t,i}^2} + 2\epsilon\sqrt{K}\right] + C_0\nu\beta_T \sum_{t=1}^{T}\mathbb{E}\left[\sqrt{\sum_{i\in S_t}\sigma_{t,i}^2}\right] \\
&= C_0\nu(\beta_T+2)\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\sum_{i\in S_t}\sigma_{t,i}^2}\right] + 2C_0T\sqrt{K}\epsilon \\
&\le C_0\nu(\beta_T+2)\mathbb{E}\left[\sqrt{T\sum_{t=1}^{T}\sum_{i\in S_t}\sigma_{t,i}^2}\right] + 2C_0T\sqrt{K}\epsilon \\
&= C_0\nu(\beta_T+2)\mathbb{E}\left[\sqrt{T\lambda\sum_{t=1}^{T}\sum_{i\in S_t}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/\sqrt{m}\right\|_{\widetilde{\mathbf{Z}}_t^{-1}}^2}\right] + 2C_0T\sqrt{K}\epsilon \\
&\le C_0\nu(\beta_T+2)\sqrt{T\lambda\left(2\widetilde{d}\log(1+TN/\lambda) + 2 + C_1T^{\frac{5}{3}}K^{\frac{3}{2}}L^4\lambda^{-\frac{1}{6}}m^{-\frac{1}{6}}\sqrt{\log m}\right)} \\
&\quad + 2C_0T\sqrt{K}\epsilon\,,
\end{aligned} \tag{14}
$$

where $C_1 > 0$ is a constant, the first inequality uses $\beta_t \le \beta_T$ and $C_0 = \max\{C_0^{(1)}, C_0^{(2)}\}$, the second inequality follows from the Cauchy-Schwarz inequality, and the last inequality holds due to Lemma 2.

Bounding $\mathcal{R}_1(T)$

Note that a sufficient condition for ensuring the success of CN-TS is to show that the probability of sampling being optimistic is high enough. Lemma 3 gives a lower bound of the probability that the reward induced by sampled scores is larger than the reward induced by the expected scores up to the approximation error. For our analysis, first we define $\widetilde{\mathcal{V}}_t$ the set of concentrated samples for which the reward induced by sampled scores concentrate appropriately to the reward induced by the estimated scores. Also, we define the set of optimistic samples $\widetilde{\mathcal{V}}_t^{\text{opt}}$ which coinciding with $\widetilde{\mathcal{V}}_t$.

$$
\begin{aligned}
\widetilde{\mathcal{V}}_t &:= \left\{\{\widetilde{v}_{t,i}^{(j)} \mid i\in[N]\}_{j=1}^{M} =: \dot{\mathbf{v}}_t^{1:M} \mid R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) - R(S_t, \hat{\mathbf{v}}_t + \boldsymbol{\epsilon}) \le C_0\sqrt{\sum_{i\in S_t}(\beta_t\nu\sigma_{t,i})^2}\right\}\,, \\
\widetilde{\mathcal{V}}_t^{\text{opt}} &:= \left\{\{\widetilde{v}_{t,i}^{(j)} \mid i\in[N]\}_{j=1}^{M} =: \ddot{\mathbf{v}}_t^{1:M} \mid R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) > R(S_t^*, \mathbf{v}_t^*)\right\} \cap \widetilde{\mathcal{V}}_t\,.
\end{aligned}
$$

Also, note that the event $\mathcal{E}_t := \mathcal{E}_t^{\sigma} \cap \mathcal{E}_t^{\mu}$, which means

$$
\mathcal{E}_t = \{|f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) - h(\mathbf{x}_{t,i})| \le \nu\sigma_{t,i} + \epsilon, \forall i\in[N]\} \cap \{|\widetilde{v}_{t,i} - f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})| \le \beta_t\nu\sigma_{t,i}, \forall i\in[N]\}\,.
$$

For our notations, we denote $\dot{S}_t$ as the super arm induced by the sampled score $\dot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t$ and $\boldsymbol{\epsilon}$. Also we represent $R(S, \dot{\mathbf{v}}_t^{1:M} + \boldsymbol{\epsilon})$ the reward under the sampled score $\dot{\mathbf{v}}_t^{1:M}$ and $\boldsymbol{\epsilon}$. Also, we define $\ddot{S}_t$ as the super arm induced by $\ddot{\mathbf{v}}_t \in \widetilde{\mathcal{V}}_t^{\text{opt}}$ and $\boldsymbol{\epsilon}$. Similarly we can define $R(S, \ddot{\mathbf{v}}_t + \boldsymbol{\epsilon})$.

Recall that $S_t = \operatorname{argmax} R(S, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})$. Then, for any $\dot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t$, we have

$$
\left(R(S_t^*, \mathbf{v}_t^*) - R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})\right)\mathbb{I}\left(\mathcal{E}_t\right) \le \left(R(S_t^*, \mathbf{v}_t^*) - \inf_{\dot{\mathbf{v}}_t^{1:M}\in\widetilde{\mathcal{V}}_t}\max_{S} R(S, \dot{\mathbf{v}}_t^{1:M} + \boldsymbol{\epsilon})\right)\mathbb{I}\left(\mathcal{E}_t\right)\,.
$$

Note that we can decompose

$$
R(S_t^*, \mathbf{v}_t^*) - R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) = \left(R(S_t^*, \mathbf{v}_t^*) - R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})\right)\mathbb{I}(\mathcal{E}_t) + \left(R(S_t^*, \mathbf{v}_t^*) - R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})\right)\mathbb{I}(\mathcal{E}_t^{c})\,.
$$

Since the event $\mathcal{E}_t$ holds with high probability, we can bound the summation of the second term in the right hand side as follows:

$$\sum_{t=1}^{T} \mathbb{E}\left[R(S_t^*, \mathbf{v}_t^*) - R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})\right] = \sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\left(R(S_t^*, \mathbf{v}_t^*) - R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})\right) \mathbb{I}(\mathcal{E}_t)\right]}_{I_3} + \mathcal{O}(1)\,.$$

Therefore, we need to bound the summation of I_3. Note that we have

$$\begin{aligned}
&\mathbb{E}\left[\left(R(S_t^*, \mathbf{v}_t^*) - R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})\right) \mathbb{I}(\mathcal{E}_t) \mid \mathcal{F}_t\right] \\
&\le \mathbb{E}\left[\left(R(S_t^*, \mathbf{v}_t^*) - \inf_{\dot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t} \max_{S} R(S, \dot{\mathbf{v}}_t^{1:M} + \boldsymbol{\epsilon})\right) \mathbb{I}\left(\mathcal{E}_t\right) \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\left(R(S_t^*, \mathbf{v}_t^*) - \inf_{\dot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t} \max_{S} R(S, \dot{\mathbf{v}}_t^{1:M} + \boldsymbol{\epsilon})\right) \mathbb{I}\left(\mathcal{E}_t\right) \mid \mathcal{F}_t, \ddot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t^{\text{opt}}\right] \\
&\le \mathbb{E}\left[\left(R(\ddot{S}_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) - \inf_{\dot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t} \max_{S} R(S, \dot{\mathbf{v}}_t^{1:M} + \boldsymbol{\epsilon})\right) \mathbb{I}\left(\mathcal{E}_t\right) \mid \mathcal{F}_t, \ddot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t^{\text{opt}}\right] \\
&\le \mathbb{E}\left[\left(R(\ddot{S}_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) - \inf_{\dot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t} R(\ddot{S}_t, \dot{\mathbf{v}}_t^{1:M} + \boldsymbol{\epsilon})\right) \mathbb{I}\left(\mathcal{E}_t\right) \mid \mathcal{F}_t, \ddot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t^{\text{opt}}\right] \\
&= \mathbb{E}\left[\sup_{\dot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t} \left(R(\ddot{S}_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) - R(\ddot{S}_t, \dot{\mathbf{v}}_t^{1:M} + \boldsymbol{\epsilon})\right) \mathbb{I}\left(\mathcal{E}_t\right) \mid \mathcal{F}_t, \ddot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t^{\text{opt}}\right] \\
&= \mathbb{E}\left[\left(R(\ddot{S}_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) - R(\ddot{S}_t, \ddot{\mathbf{v}}_t^{1:M} + \boldsymbol{\epsilon})\right) \mathbb{I}\left(\mathcal{E}_t\right) \mid \mathcal{F}_t, \ddot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t^{\text{opt}}\right] \\
&= \mathbb{E}\left[\left(R(\ddot{S}_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) - R(\ddot{S}_t, \hat{\mathbf{v}}_t + \boldsymbol{\epsilon}) + R(\ddot{S}_t, \hat{\mathbf{v}}_t + \boldsymbol{\epsilon}) - R(\ddot{S}_t, \ddot{\mathbf{v}}_t^{1:M} + \boldsymbol{\epsilon})\right) \mathbb{I}\left(\mathcal{E}_t\right) \mid \mathcal{F}_t, \ddot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t^{\text{opt}}\right] \\
&\le \mathbb{E}\left[2C_0 \sqrt{\sum_{i \in \ddot{S}_t} (\beta_t \nu \sigma_{t,i})^2}\, \mathbb{I}\left(\mathcal{E}_t\right) \mid \mathcal{F}_t, \ddot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t^{\text{opt}}\right] \\
&= 2C_0 \beta_t \nu \mathbb{E}\left[\sqrt{\sum_{i \in \ddot{S}_t} \sigma_{t,i}^2}\, \mathbb{I}\left(\mathcal{E}_t\right) \mid \mathcal{F}_t, \ddot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t^{\text{opt}}\right] \\
&\le 2C_0 \beta_t \nu \mathbb{E}\left[\sqrt{\sum_{i \in \ddot{S}_t} \sigma_{t,i}^2} \mid \mathcal{F}_t, \ddot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t^{\text{opt}}\right] \cdot \mathbb{P}(\mathcal{E}_t)\,,
\end{aligned}$$

where $C_0 > 0$ is a Lipschitz constant. On the other hand, from Lemma 3, we have

$$\mathbb{P}\left(R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) > R(S_t^*, \mathbf{v}_t^*) \mid \mathcal{F}_t, \mathcal{E}_t\right) \ge 1/(4e\sqrt{\pi}) := \widetilde{p}\,,$$

which means that

$$\begin{aligned}
\mathbb{P}(\ddot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t^{\text{opt}} \mid \mathcal{F}_t, \mathcal{E}_t) &= \mathbb{P}\left(R(\ddot{S}_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) > R(S_t^*, \mathbf{v}_t^*) \text{ and } \ddot{\mathbf{v}}_t^{1:M} \in \widetilde{\mathcal{V}}_t \mid \mathcal{F}_t, \mathcal{E}_t\right) \\
&\ge \mathbb{P}\left(R(\ddot{S}_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) > R(S_t^*, \mathbf{v}_t^*) \mid \mathcal{F}_t, \mathcal{E}_t\right) - \mathbb{P}\left(\ddot{\mathbf{v}}_t^{1:M} \notin \widetilde{\mathcal{V}}_t \mid \mathcal{F}_t, \mathcal{E}_t\right) \\
&\ge \widetilde{p} - \mathcal{O}(t^{-1}) \\
&\ge \widetilde{p}/2.
\end{aligned}$$

Then, we can write

$$\mathbb{E}\left[\sqrt{\sum_{i\in S'_t}\sigma_{t,i}^2} \mid \mathcal{F}_t,\mathcal{E}_t\right] \geq \mathbb{E}\left[\sqrt{\sum_{i\in \ddot{S}_t}\sigma_{t,i}^2} \mid \mathcal{F}_t,\mathcal{E}_t,\ddot{\mathbf{v}}_t^{1:M}\in\widetilde{\mathcal{V}}_t^{\text{opt}}\right]\cdot \mathbb{P}\left(\ddot{\mathbf{v}}_t^{1:M}\in\widetilde{\mathcal{V}}_t^{\text{opt}} \mid \mathcal{F}_t,\mathcal{E}_t\right)$$
$$\geq \mathbb{E}\left[\sqrt{\sum_{i\in \ddot{S}_t}\sigma_{t,i}^2} \mid \mathcal{F}_t,\mathcal{E}_t,\ddot{\mathbf{v}}_t^{1:M}\in\widetilde{\mathcal{V}}_t^{\text{opt}}\right]\cdot \widetilde{p}/2\,,$$

where S'_t is a super arm induced by any sampled scores. By combining the results, we have

$$\mathbb{E}\left[\left(R(S_t^*,\mathbf{v}_t^*)-R(S_t,\widetilde{\mathbf{v}}_t+\boldsymbol{\epsilon})\right)\mathbb{I}(\mathcal{E}_t)\mid\mathcal{F}_t\right] \leq 2C_0\beta_t\nu\mathbb{E}\left[\sqrt{\sum_{i\in \ddot{S}_t}\sigma_{t,i}^2} \mid \mathcal{F}_t,\mathcal{E}_t,\ddot{\mathbf{v}}_t^{1:M}\in\widetilde{\mathcal{V}}_t^{\text{opt}}\right]\cdot\mathbb{P}(\mathcal{E}_t)$$
$$\leq \frac{4C_0\beta_t\nu}{\widetilde{p}}\mathbb{E}\left[\sqrt{\sum_{i\in S'_t}\sigma_{t,i}^2} \mid \mathcal{F}_t,\mathcal{E}_t\right]\cdot\mathbb{P}(\mathcal{E}_t)$$
$$\leq \frac{4C_0\beta_t\nu}{\widetilde{p}}\mathbb{E}\left[\sqrt{\sum_{i\in S'_t}\sigma_{t,i}^2} \mid \mathcal{F}_t\right]\,.$$

Summing over all $t\in[T]$ and the failure event into consideration, we have

$$\sum_{t=1}^{T}\mathbb{E}\left[\left(R(S_t^*,\mathbf{v}_t^*)-R(S_t,\widetilde{\mathbf{v}}_t+\boldsymbol{\epsilon})\right)\mathbb{I}(\mathcal{E}_t)\mid\mathcal{F}_t\right] \leq \frac{4C_0\beta_T\nu}{\widetilde{p}}\sum_{t=1}^{T}\mathbb{E}\left[\sqrt{\sum_{i\in S'_t}\sigma_{t,i}^2} \mid \mathcal{F}_t\right]\,. \tag{15}$$

Note that the summation on the RHS contains an expectation, so we cannot directly apply Lemma 2. Instead, since we can write

$$\sum_{t=1}^{T}\mathbb{E}\left[\sqrt{\sum_{i\in S'_t}\sigma_{t,i}^2} \mid \mathcal{F}_t\right] = \sum_{t=1}^{T}\sqrt{\sum_{i\in S''_t}\sigma_{t,i}^2} + \sum_{t=1}^{T}\left(\mathbb{E}\left[\sqrt{\sum_{i\in S'_t}\sigma_{t,i}^2} \mid \mathcal{F}_t\right] - \sqrt{\sum_{i\in S''_t}\sigma_{t,i}^2}\right)\,,$$

where S''_t is any super arm induced by arbitrary sampled scores. By using Lemma 2 we have

$$\sum_{t=1}^{T}\sqrt{\sum_{i\in S''_t}\sigma_{t,i}^2} \leq \sqrt{T\sum_{t=1}^{T}\sum_{i\in S''_t}\sigma_{t,i}^2}$$
$$\leq \sqrt{T\lambda\left(2\widetilde{d}\log(1+TN/\lambda)+2+C_1T^{\frac{5}{3}}K^{\frac{3}{2}}L^4\lambda^{-\frac{1}{6}}m^{-\frac{1}{6}}\sqrt{\log m}\right)}\,, \tag{16}$$

where $C_1>0$ is a constant.

On the other hand, let $Y_t=\sum_{k=1}^{t}\left(\mathbb{E}\left[\sqrt{\sum_{i\in S'_k}\sigma_{k,i}^2}\mid\mathcal{F}_k\right]-\sqrt{\sum_{i\in S''_k}\sigma_{k,i}^2}\right)$. Since we have

$$Y_t-Y_{t-1}=\mathbb{E}\left[\sqrt{\sum_{i\in S'_t}\sigma_{t,i}^2} \mid \mathcal{F}_t\right]-\sqrt{\sum_{i\in S''_t}\sigma_{t,i}^2}\,,$$

which implies,

$$\mathbb{E}\left[Y_t-Y_{t-1}\mid\mathcal{F}_t\right]=\mathbb{E}\left[\mathbb{E}\left[\sqrt{\sum_{i\in S'_t}\sigma_{t,i}^2} \mid \mathcal{F}_t\right]\mid\mathcal{F}_t\right]-\mathbb{E}\left[\sqrt{\sum_{i\in S''_t}\sigma_{t,i}^2} \mid \mathcal{F}_t\right]=0\,,$$

then Y_t is a martingale for all $1 \le t \le T$.

Note that we can bound $|Y_t - Y_{t-1}|$ as follows:

$$\begin{aligned}
|Y_t - Y_{t-1}| &= \left| \mathbb{E}\left[\sqrt{\sum_{i \in S'_t} \sigma^2_{t,i}} \mid \mathcal{F}_t \right] - \sqrt{\sum_{i \in S''_t} \sigma^2_{t,i}} \right| \\
&\le \mathbb{E}\left[\sqrt{\sum_{i \in S'_t} (C_2 \sqrt{L})^2} \mid \mathcal{F}_t \right] + \sqrt{\sum_{i \in S''_t} (C_2 \sqrt{L})^2} \\
&= 2C_2 \sqrt{LK}\,,
\end{aligned}$$

where the inequality holds due to Lemma 7 for some positive constant C_2. Then, applying the Azuma-Hoeffding inequality (Lemma 10), which means,

$$\sum_{t=1}^{T} \left(\mathbb{E}\left[\sqrt{\sum_{i \in S'_t} \sigma^2_{t,i}} \mid \mathcal{F}_t \right] - \sqrt{\sum_{i \in S''_t} \sigma^2_{t,i}} \right) \le C_2 \sqrt{8TLK \log T}\,, \tag{17}$$

with probability $1 - T^{-1}$. Combining Eq.(16) and Eq.(17), we have

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\sum_{i \in S'_t} \sigma^2_{t,i}} \mid \mathcal{F}_t \right] &\le \sqrt{T\lambda \left(2\widetilde{d} \log(1 + TN/\lambda) + 2 + C_1 T^{\frac{5}{3}} K^{\frac{3}{2}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} \right)} \\
&\quad + C_2 \sqrt{8TLK \log T}
\end{aligned} \tag{18}$$

By substituting Eq.(18) for Eq.(15), we have the bound for $\mathcal{R}_1(T)$ as follows:

$$\begin{aligned}
\mathcal{R}_1(T) \le \frac{4C_0 \beta_t \nu}{\widetilde{p}} &\left[\sqrt{T\lambda \left(2\widetilde{d} \log(1 + TN/\lambda) + 2 + C_1 T^{\frac{5}{3}} K^{\frac{3}{2}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} \right)} \right. \\
&\left. + C_2 \sqrt{8TLK \log T} \right] + \mathcal{O}(1)
\end{aligned} \tag{19}$$

Finally, combining Eq.(19) and Eq.(14) we have

$$\begin{aligned}
\mathcal{R}(T) \le \frac{4C_0 \beta_T \nu}{\widetilde{p}} &\left[\sqrt{T\lambda \left(2\widetilde{d} \log(1 + TN/\lambda) + 2 + C_1 T^{\frac{5}{3}} K^{\frac{3}{2}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} \right)} \right. \\
&\left. + C_2 \sqrt{8TLK \log T} \right] + 2C_0 T \sqrt{K} \epsilon + \mathcal{O}(1) \\
&+ C_0 \nu (\beta_T + 2) \sqrt{T\lambda \left(2\widetilde{d} \log(1 + TN/\lambda) + 2 + C_1 T^{\frac{5}{3}} K^{\frac{3}{2}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} \right)}
\end{aligned} \tag{20}$$

Then choosing m such that

$$\begin{aligned}
&C_1 T^{\frac{5}{3}} K^{\frac{3}{2}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} \le 1\,, \\
&C_{\epsilon,1} T^{\frac{5}{3}} K^{\frac{2}{3}} L^3 \lambda^{-\frac{2}{3}} m^{-\frac{1}{6}} \sqrt{\log m} \le \frac{1}{4}\,, \\
&C_{\epsilon,3} T^{\frac{13}{6}} K^{\frac{7}{6}} L^4 \lambda^{-\frac{7}{6}} m^{-\frac{1}{6}} \sqrt{\log m} (1 + \sqrt{TK/\lambda}) \le \frac{1}{4}\,, \\
&C_{\epsilon,4} T^{\frac{13}{6}} K^{\frac{7}{6}} \lambda^{-\frac{2}{3}} L^{\frac{9}{2}} m^{-\frac{1}{6}} \sqrt{\log m} \left(B + \rho \sqrt{\log \det(\mathbf{I} + \mathbf{H}/\lambda) + 2 - 2\log \delta} \right) \le \frac{1}{4}\,.
\end{aligned}$$

Also by setting $J = 2\log\left(\frac{1}{4TC_{\epsilon,2}} \frac{\sqrt{\lambda}}{TKL} \right) \frac{TKL}{C_1}$, we have

$$TC_{\epsilon,2}(1 - \eta m \lambda)^{J/2} \sqrt{TKL/\lambda} \le \frac{1}{4}$$

which follows, $T\epsilon \le 1$. Hence, $\mathcal{R}(T)$ can be bounded by

$$\mathcal{R}(T) \le \frac{4C_0\beta_T\nu}{\widetilde{p}}\left[\sqrt{T\lambda\left(2\widetilde{d}\log(1+TN/\lambda)+3\right)} + C_2\sqrt{8TLK\log T}\right] + 2C_0\sqrt{K} + \mathcal{O}(1)$$
$$+ C_0\nu(\beta_T+2)\sqrt{T\lambda\left(2\widetilde{d}\log(1+TN/\lambda)+3\right)}\,.$$

□

C. Auxiliary Lemmas

Lemma 9. *(Abramowitz & Stegun, 1964) For a Gaussian distributed random variable Z with mean μ and variance σ^2, for any $z \ge 1$,*

$$\frac{1}{2\sqrt{\pi}z}\exp(-z^2/2) \le \mathbb{P}(|Z-\mu| > z\sigma) \le \frac{1}{\sqrt{\pi}z}\exp(-z^2/2)\,.$$

Lemma 10 (Azuma-Hoeffding inequality)**.** *If a super-martingale $(Y_t, t \ge 0)$ corresponding to filtration $\mathcal{F}_t$, satisfies $|Y_t - Y_{t-1}| < \beta_t$ for some constant β_t, for all $t = 1, \dots, T$, then for any $a \ge 0$,*

$$\mathbb{P}(Y_t - Y_{t-1} \ge a) \le 2\exp\left(-\frac{a^2}{2\sum_{t=1}^{T}\beta_t^2}\right)\,.$$

D. Extensions from Neural Bandits for Single Action

In this section, we describe how the auxiliary lemmas used in the neural bandit works (Zhou et al., 2020; Zhang et al., 2021) for single action can be extended to the combinatorial action settings. The main distinction is that in single action settings, the amount of data to be trained at time t is t, whereas in combinatorial action settings, it is tK. Therefore, by properly accounting for this difference, we can obtain the following results.

Definition 3. *For simplicity, we restate some definitions used in this section.*

$$\begin{aligned}
&\bar{\mathbf{Z}}_t = \lambda\mathbf{I} + \sum_{k=1}^{t}\sum_{i\in S_k}\mathbf{g}(\mathbf{x}_{k,i};\boldsymbol{\theta}_0)\mathbf{g}(\mathbf{x}_{k,i};\boldsymbol{\theta}_0)^\top/m\,,\\
&\mathbf{Z}_t(\textit{or } \widetilde{\mathbf{Z}}_t) = \lambda\mathbf{I} + \sum_{k=1}^{t}\sum_{i\in S_k}\mathbf{g}(\mathbf{x}_{k,i};\boldsymbol{\theta}_{k-1})\mathbf{g}(\mathbf{x}_{k,i};\boldsymbol{\theta}_{k-1})^\top/m\,,\\
&\bar{\sigma}_{t,i}^2 = \lambda\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)^\top\bar{\mathbf{Z}}_{t-1}\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_0)/m\,,\\
&\sigma_{t,i}^2 = \lambda\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})^\top\widetilde{\mathbf{Z}}_{t-1}\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/m\,,\\
&\bar{\mathbf{J}}_t = [\mathbf{g}(\mathbf{x}_{1,a_{11}};\boldsymbol{\theta}_0),\dots,\mathbf{g}(\mathbf{x}_{1,a_{1K}};\boldsymbol{\theta}_0),\dots\mathbf{g}(\mathbf{x}_{t,a_{tK}};\boldsymbol{\theta}_0)] \in \mathbb{R}^{p\times tK}\,,\\
&\mathbf{J}_t = [\mathbf{g}(\mathbf{x}_{1,a_{11}};\boldsymbol{\theta}_{t-1}),\dots,\mathbf{g}(\mathbf{x}_{1,a_{1K}};\boldsymbol{\theta}_{t-1}),\dots\mathbf{g}(\mathbf{x}_{t,a_{tK}};\boldsymbol{\theta}_{t-1})] \in \mathbb{R}^{p\times tK}\,,\\
&\mathbf{y}_t = [v_{1,a_{11}},\dots,v_{1,a_{1K}},\dots,v_{t,a_{tK}}]^\top \in \mathbb{R}^{tK}\,,
\end{aligned}$$

where a_{tk} is the k-th action in the super arm S_t at time t, i.e., $S_t := \{a_{t1}, \cdots, a_{tK}\}$.

Lemma 11 (Lemma 5.2 in Zhou et al. (2020))**.** *Suppose that there exist some positive constants $\bar{C}_1, \bar{C}_2 > 0$ such that for any $\delta \in (0,1)$, $\eta \le \bar{C}_1(TKmL + m\lambda)^{-1}$ and*

$$\begin{aligned}
&m \ge \bar{C}_2K^{-\frac{1}{2}}L^{-\frac{3}{2}}\lambda^{\frac{1}{2}}\left(\log(TNL^2/\delta)\right)^{\frac{3}{2}}\,,\\
&m(\log m)^{-3} \ge \bar{C}_2\left(TKL^{12}\lambda^{-1} + T^4K^4L^{18}\lambda^{-10}(\lambda+TKL)^6 + T^7K^7L^{21}\lambda^{-7}(1+\sqrt{TK/\lambda})^6\right)\,.
\end{aligned}$$

Then, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \le 2\sqrt{tK/(m\lambda)}\,,\\
&\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|_{\mathbf{Z}_t} \le \gamma_t/\sqrt{m}\,.
\end{aligned}$$

Lemma 12 (Lemma B.2 in Zhou et al. (2020)). *There exist some constants $\{\bar{C}_i\}_{i=1}^4 > 0$ such that for any $\delta \in (0,1)$, if for any $t \in [T]$, η, m satisfy that*

$$
\begin{aligned}
&2\sqrt{tK/(m\lambda)} \geq \bar{C}_1 m^{-\frac{3}{2}} L^{-\frac{3}{2}} \left(\log(TNL^2/\delta)\right)^{\frac{3}{2}} , \\
&2\sqrt{tK/(m\lambda)} \leq \bar{C}_2 \min\left\{ L^{-6} \left(\log m\right)^{-\frac{3}{2}} , \left(m\lambda^2\eta^2 L^{-6} (\log m)^{-1}\right)^{\frac{3}{8}} \right\} , \\
&\eta \leq \bar{C}_3 (m\lambda + tKmL)^{-1} , \\
&m^{\frac{1}{6}} \geq \bar{C}_4 t^{\frac{7}{6}} K^{\frac{7}{6}} L^{\frac{7}{2}} \lambda^{-\frac{7}{6}} \sqrt{\log m}(1 + \sqrt{tK/\lambda}) ,
\end{aligned}
$$

then, with probability at least $1-\delta$, we have $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\| \leq 2\sqrt{tK/(m\lambda)}$ and

$$
\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{J}}_t \mathbf{y}_t / m\|_2 \leq (1-\eta m\lambda)^{\frac{J}{2}} \sqrt{tK/(m\lambda)} + \bar{C}_5 t^{\frac{7}{6}} K^{\frac{7}{6}} L^{\frac{7}{2}} \lambda^{-\frac{7}{6}} m^{-\frac{2}{3}} \sqrt{\log m}(1 + \sqrt{tK/\lambda}) ,
$$

where $\bar{C}_5 > 0$ is an absolute constant.

Lemma 13 (Lemma B.3 in Zhou et al. (2020)). *There exist some constants $\{\bar{C}_i\}_{i=1}^5 > 0$ such that for any $\delta \in (0,1)$, if for any $t \in [T]$, m satisfies that*

$$
\bar{C}_1 m^{-\frac{3}{2}} L^{-\frac{3}{2}} \left(log(TNL^2/\delta)\right)^{\frac{3}{2}} \leq 2\sqrt{tK/(m\lambda)} \leq \bar{C}_2 L^{-6} (\log m)^{-\frac{3}{2}} ,
$$

then, with probability at least $1-\delta$, for any $t \in [T]$ we have

$$
\begin{aligned}
&\|\mathbf{Z}_t\|_2 \leq \lambda + \bar{C}_3 tKL , \\
&\|\bar{\mathbf{Z}}_t - \mathbf{Z}_t\|_F \leq \bar{C}_4 t^{\frac{7}{6}} K^{\frac{7}{6}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} , \\
&\left| \log \frac{\det \bar{\mathbf{Z}}_t}{\det \lambda \mathbf{I}} - \log \frac{\det \mathbf{Z}_t}{\det \lambda \mathbf{I}} \right| \leq \bar{C}_5 t^{\frac{5}{3}} K^{\frac{5}{3}} L^4 \lambda^{-\frac{1}{6}} m^{-\frac{1}{6}} \sqrt{\log m} ,
\end{aligned}
$$

where $\bar{C}_3, \bar{C}_4, \bar{C}_5 > 0$ are some absolute constants, and $\bar{\mathbf{Z}}_t = \lambda \mathbf{I} + \sum_{k=1}^{t-1} \sum_{i \in S_k} \mathbf{g}(\mathbf{x}_{k,i}; \boldsymbol{\theta}_0) \mathbf{g}(\mathbf{x}_{k,i}; \boldsymbol{\theta}_0)^\top$.

Lemma 14 (Lemma C.2 in Zhou et al. (2020)). *For any $\delta \in (0,1)$, $\bar{C}_1, \bar{C}_2 > 0$, suppose that τ satisfies*

$$
\bar{C}_1 m^{-\frac{3}{2}} L^{-\frac{3}{2}} \left(\log(TNL^2/\delta)\right)^{\frac{3}{2}} \leq \tau \leq \bar{C}_2 L^{-6} (\log m)^{-\frac{3}{2}} ,
$$

Then, with probability at least $1-\delta$, if for any $j \in [J]$, $\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(0)}\|_2 \leq \tau$, we have the following results for any $j, s \in [J]$,

$$
\begin{aligned}
&\|\mathbf{J}^{(j)}\|_F \leq \bar{C}_3 \sqrt{tKmL} , \\
&\|\mathbf{J}^{(j)} - \mathbf{J}^{(0)}\|_F \leq \bar{C}_4 \tau^{\frac{1}{3}} L^{\frac{7}{2}} \sqrt{tKm \log m} , \\
&\|\mathbf{f}^{(s)} - \mathbf{f}^{(j)} - (\mathbf{J}^{(j)})^\top (\boldsymbol{\theta}^{(s)} - \boldsymbol{\theta}^{(j)})\|_2 \leq \bar{C}_5 \tau^{\frac{4}{3}} L^3 \sqrt{tKm \log m} , \\
&\|\mathbf{y}\|_2 \leq \sqrt{tK} ,
\end{aligned}
$$

where $\bar{C}_3, \bar{C}_4, \bar{C}_5 > 0$ are some absolute constants.

Lemma 15 (Lemma C.3 in Zhou et al. (2020)). *For any $\delta \in (0,1)$ and $\{\bar{C}_i\}_{i=1}^4 > 0$, suppose that τ, η satisfy*

$$
\begin{aligned}
&\bar{C}_1 m^{-\frac{3}{2}} L^{-\frac{3}{2}} \left(\log(TNL^2/\delta)\right)^{\frac{3}{2}} \leq \bar{C}_2 L^{-6} (\log m)^{-\frac{3}{2}} , \\
&\eta \leq \bar{C}_3 (m\lambda + tKmL)^{-1} , \\
&\tau^{\frac{8}{3}} \leq \bar{C}_4 m \lambda^2 \eta^2 L^{-6} (\log m)^{-1} .
\end{aligned}
$$

Then, with probability at least $1-\delta$, if for any $j \in [J]$, $\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(0)}\|_2 \leq \tau$, we have that for any $j \in [J]$, $\|\mathbf{f}^{(j)} - \mathbf{y}\|_2 \leq 2\sqrt{tK}$.

Lemma 16 (Lemma C.4 in Zhou et al. (2020)). *For any $\delta \in (0,1)$ and $\{\bar{C}_i\}_{i=1}^3 > 0$, suppose that τ, η satisfy*

$$
\begin{aligned}
&\bar{C}_1 m^{-\frac{3}{2}} L^{-\frac{3}{2}} \left(\log(TNL^2/\delta)\right)^{\frac{3}{2}} \leq \bar{C}_2 L^{-6} (\log m)^{-\frac{3}{2}} , \\
&\eta \leq \bar{C}_3 (m\lambda + tKmL)^{-1} .
\end{aligned}
$$

Then, with probability at least $1-\delta$, we have for any $j \in [J]$,

$$
\begin{aligned}
&\|\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(0)}\|_2 \leq \sqrt{tK/(m\lambda)} , \\
&\|\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(0)} - \bar{\mathbf{Z}}^{-1} \bar{\mathbf{J}} \mathbf{y} / m\|_2 \leq (1-\eta m\lambda)^{\frac{j}{2}} \sqrt{tK/(m\lambda)} .
\end{aligned}
$$

E. Extension of Regret Analysis to α-approximation Oracle

In this section, we extend our regret analysis to the case when the agent only has access to an α-approximation oracle, $\mathbb{O}_{\mathcal{S}}^{\alpha}$. First, we replace S_t with $S_t^{\alpha} = \mathbb{O}_{\mathcal{S}}^{\alpha}(\mathbf{u}_t + \mathbf{e}_t)$ for `CN-UCB` (Algorithm 1) and $S_t^{\alpha} = \mathbb{O}_{\mathcal{S}}^{\alpha}(\widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})$ for `CN-TS` (Algorithm 2). The total regret $\mathcal{R}(T)$ is replaced with an α-regret defined as:

$$\mathcal{R}^{\alpha}(T) = \sum_{t=1}^{T} \mathbb{E}\left[\alpha R(S_t^*, \mathbf{v}_t^*) - R(S_t^{\alpha}, \mathbf{v}_t^*)\right]$$

.

For `CN-UCB`, note that $\alpha R(S_t, \mathbf{u}_t + \mathbf{e}_t) \le R(S_t^{\alpha}, \mathbf{u}_t + \mathbf{e}_t)$. Also, $\alpha R(S_t^*, \mathbf{v}_t^*) \le \alpha R(S_t^*, \mathbf{u}_t + \mathbf{e}_t) \le \alpha R(S_t, \mathbf{u}_t + \mathbf{e}_t) \le R(S_t^{\alpha}, \mathbf{u}_t + \mathbf{e}_t)$. We can derive that the α-regret bound of `CN-UCB` is $\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{T})$ or $\widetilde{\mathcal{O}}(\sqrt{\widetilde{d}TK})$, whichever is higher, by substituting the following notations in Appendix A.4:

$$\begin{aligned} R(S_t^*, \mathbf{v}_t^*) &\to \alpha R(S_t^*, \mathbf{v}_t^*) \\ R(S_t^*, \mathbf{u}_t + \mathbf{e}_t) &\to \alpha R(S_t^*, \mathbf{u}_t + \mathbf{e}_t) \\ S_t &\to S_t^{\alpha} \, . \end{aligned}$$

For `CN-TS`, note that $\alpha R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) \le R(S_t^{\alpha}, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})$. We split α-regret as follows:

$$\begin{aligned} \mathcal{R}^{\alpha}(T) &= \mathcal{R}_1^{\alpha}(T) + \mathcal{R}_2^{\alpha}(T) \\ &= \sum_{t=1}^{T} \mathbb{E}\left[\alpha R(S_t^*, \mathbf{v}_t^*) - \alpha R(S_t^{\alpha}, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})\right] \\ &+ \sum_{t=1}^{T} \mathbb{E}\left[\alpha R(S_t^{\alpha}, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) - R(S_t^{\alpha}, \mathbf{v}_t^*)\right] \, . \end{aligned}$$

By replacing S_t with S_t^{α} in Appendix B.2, we can get the α-regret bound of $\mathcal{R}_2^{\alpha}(T)$. For $\mathcal{R}_1^{\alpha}(T)$, since $\alpha R(S_t^*, \mathbf{v}_t^*) - \alpha R(S_t^{\alpha}, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) \le \alpha R(S_t^*, \mathbf{v}_t^*) - \alpha R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon}) \le R(S_t^*, \mathbf{v}_t^*) - R(S_t, \widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})$, we know that $\mathcal{R}_1^{\alpha}(T) \le \mathcal{R}_1(T)$. By combining the results, we can conclude that the α-regret bound of `CN-TS` is $\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{TK})$.

F. When Time Horizon T Is Unknown

For Theorems 1 and 2, we assumed that T is known for the sake of clear exposition for our proposed algorithms and their regret analysis. However, the knowledge of T is not essential both for the algorithms and their analysis. With slight modifications, our proposed algorithms can be applied to the settings where T is unknown. In this section, we propose the variants of `CN-UCB` and `CN-TS`: `CN-UCB` with `doubling` and `CN-TS` with `doubling`, and show that their regret upper bounds are of the same order of regret as those of `CN-UCB` and `CN-TS` up to logarithmic factors.

F.1. Algorithms

`CN-UCB` with `doubling` and `CN-TS` with `doubling` utilize a doubling technique (Besson & Kaufmann, 2018) in which the network size stays fixed during each epoch but is updated after the end of each epoch whose length τ doubles the length of a previous epoch. This way, even when T is unknown, the networks size can be set adaptively over epochs.

The algorithms first initialize the variables related to τ, especially the hidden layer width m_τ and the number of parameters of the neural network $p(\tau)$. For each round, after playing super arm S_t and observing the scores $\{v_{t,i}\}_{i \in S_t}$, `CN-UCB` with `doubling` and `CN-TS` with `doubling` call the `Update` algorithm. Until τ, `Update` algorithm updates $\boldsymbol{\theta}_t$ and $\mathbf{Z}_t$ or $\widetilde{\mathbf{Z}}_t$ as if τ is the time horizon. If t reaches τ, `Update` algorithm *doubles* the value of τ. After reinitializing the variables related to the *doubled* τ, which includes reconstructing the neural network to have a larger hidden layer width m_τ, the algorithm updates all of the $\boldsymbol{\theta}_{t'}$ and $\mathbf{Z}_{t'}$ or $\widetilde{\mathbf{Z}}_{t'}$ for $t' = 0, \cdots, t$. `Update` algorithm returns $\boldsymbol{\theta}_t$ and $\mathbf{Z}_t$ or $\widetilde{\mathbf{Z}}_t$ to `CN-UCB` with `doubling` or `CN-TS` with `doubling`. This process continues until t reaches T.

Note that the computation complexity of each round of `CN-UCB` and `CN-UCB` with `doubling` heavily depends on how quickly they can compute the inverse of the gram matrix $\mathbf{Z}$. Since $\mathbf{Z} \in \mathbb{M}_p(\mathbb{R})$, and p depends on m, the computation speed of each round in CN-UCB is relatively slow as m is a large constant. On the other hand, CN-UCB with doubling can show faster computation speed, especially at the beginning rounds, where m is kept relatively small. The same argument can be applied to CN-TS and CN-TS with doubling.

CN-UCB with doubling is summarized in Algorithm 3. CN-TS with doubling is summarized in Algorithm 4. The Update algorithm is summarized in Algorithm 5.

Algorithm 3 CN-UCB with doubling

Input: Epoch period τ, network depth L.
Initialization: Initialize {network width m_τ, regularization parameter λ_τ, norm parameter B_τ, step size η_τ, number of gradient descent steps J_τ} with respect to τ, set number of parameters of neural network $p(\tau) = m_\tau d + m_\tau^2(L-2) + m_\tau$, $\mathbf{Z}_0 = \lambda_\tau \mathbf{I}_{p(\tau)}$, randomly initialize $\boldsymbol{\theta}_0$ as described in Section 3.1
while $t \neq T$ **do**
  Observe $\{\mathbf{x}_{t,i}\}_{i\in[N]}$
  Compute $\hat{v}_{t,i} = f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})$ and $u_{t,i} = \hat{v}_{t,i} + \gamma_{t-1}\left\|\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/\sqrt{m_\tau}\right\|_{\mathbf{Z}_{t-1}^{-1}}$ for $i \in [N]$
  Let $S_t = \mathbb{O}_{\mathcal{S}}(\mathbf{u}_t + \mathbf{e}_t)$
  Play super arm S_t and observe $\{v_{t,i}\}_{i\in S_t}$
  Update(t, τ)
  Compute γ_t and e_{t+1} described in lemma 1 (replace $\{\lambda, m, \mathbf{I}, B, \eta, J\}$ with $\{\lambda_\tau, m_\tau, \mathbf{I}_{p(\tau)}, B_\tau, \eta_\tau, J_\tau\}$)
end while

Algorithm 4 CN-TS with doubling

Input: Epoch period τ, network depth L, sample size M
Initialization: Initialize {network width m_τ, regularization parameter λ_τ, exploration variance ν_τ, step size η_τ, number of gradient descent steps J_τ} with respect to τ, set number of parameters of neural network $p(\tau) = m_\tau d + m_\tau^2(L-2) + m_\tau$, $\widetilde{\mathbf{Z}}_0 = \lambda_\tau \mathbf{I}_{p(\tau)}$, randomly initialize $\boldsymbol{\theta}_0$ as described in Section 3.1
while $t \neq T$ **do**
  Observe $\{\mathbf{x}_{t,i}\}_{i\in[N]}$.
  Compute $\sigma_{t,i}^2 = \lambda_\tau \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})^\top \widetilde{\mathbf{Z}}_{t-1}^{-1}\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/m_\tau$ for each $i \in [N]$
  Sample $\{\widetilde{v}_{t,i}^{(j)}\}_{j=1}^M$ independently from $\mathcal{N}(f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}), \nu_\tau^2\sigma_{t,i}^2)$ for each $i \in [N]$
  Compute $\widetilde{v}_{t,i} = \max_j \widetilde{v}_{t,i}^{(j)}$ for each $i \in [N]$
  Let $S_t = \mathbb{O}_{\mathcal{S}}(\widetilde{\mathbf{v}}_t + \boldsymbol{\epsilon})$
  Play super arm S_t and observe $\{v_{t,i}\}_{i\in S_t}$
  Update(t, τ)
end while

F.2. Regret Analysis

The regret upper bounds of CN-UCB with doubling and CN-UCB (or CN-TS with doubling and CN-TS) have the same rate up to logarithmic factors. We provide the sketch of proof.

By modifying Definitions 1 and 2 with respect to τ, the effective dimension $\widetilde{d}_\tau$ can be written as $\widetilde{d}_\tau = \log\det(\mathbf{I} + \mathbf{H}_\tau/\lambda_\tau)/\log(1+\tau N/\lambda_\tau)$. Denote the epoch periods as $\tau_n = 2^n\tau_0$, where $n \in \mathbb{Z}_{\geq 0}$ and τ_0 is the initial epoch period.

If $T < \tau_0$, CN-UCB with doubling and CN-TS with doubling are equivalent to CN-UCB and CN-TS respectively. In this case, there is no change in the regret upper bounds. Meanwhile, if $T \geq \tau_0$, there exists $\hat{n} \in \mathbb{Z}_+$ such that $\tau_{\hat{n}-1} \leq T < \tau_{\hat{n}}$. Denote the instantaneous regret as Reg_t. Define $\sum_{t=a}^b \mathrm{Reg}_t := 0$ if $a > b$. Then the regret can be written as

$$\mathcal{R}(T) = \sum_{t=1}^{\tau_0}\mathrm{Reg}_t + \sum_{t=\tau_0+1}^{\tau_1}\mathrm{Reg}_t + \cdots + \sum_{\tau_{\hat{n}-2}+1}^{\tau_{\hat{n}-1}}\mathrm{Reg}_t + \sum_{t=\tau_{\hat{n}-1}+1}^{T}\mathrm{Reg}_t\,.$$

Let $\widetilde{d} := \max\{\widetilde{d}_{\tau_0}, \ldots, \widetilde{d}_{\tau_{\hat{n}}}\}$. For CN-UCB with doubling, each sum has an upper bound $\widetilde{\mathcal{O}}(\max\{\widetilde{d}_{\tau_n}, \sqrt{\widetilde{d}_{\tau_n}K}\}\sqrt{\tau_n})$.

Algorithm 5 $\texttt{Update}(t, \tau)$

Input: Epoch period τ, round t
if $t < \tau$ **then**
  Update $\mathbf{Z}_t = \mathbf{Z}_{t-1} + \sum_{i \in S_t} \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1}) \mathbf{g}(\mathbf{x}_{t,i}; \boldsymbol{\theta}_{t-1})^\top / m_\tau$
  Update $\boldsymbol{\theta}_t$ to minimize the loss Eq.(4) using gradient descent with η_τ for J_τ times
else
  $\tau \leftarrow 2\tau$
  Reinitialize $\{m_\tau, \lambda_\tau, \eta_\tau, J_\tau\}$ with respect to τ, set $p(\tau) = m_\tau d + m_\tau^2 (L-2) + m_\tau$, randomly reinitialize $\boldsymbol{\theta}_0$ as described in Section 3.1.
  For `CN-UCB` with `doubling`, reinitialize B_τ with respect to τ, $\mathbf{Z}_0 = \lambda_\tau \mathbf{I}_{p(\tau)}$
  For `CN-TS` with `doubling`, reinitialize ν_τ with respect to τ, $\widetilde{\mathbf{Z}}_0 = \lambda_\tau \mathbf{I}_{p(\tau)}$
  for $t' = 1, \cdots, t$ **do**
    For `CN-UCB` with `doubling`, $\mathbf{Z}_{t'} = \mathbf{Z}_{t'-1} + \sum_{i \in S_{t'}} \mathbf{g}(\mathbf{x}_{t',i}; \boldsymbol{\theta}_{t'-1}) \mathbf{g}(\mathbf{x}_{t',i}; \boldsymbol{\theta}_{t'-1})^\top / m_\tau$
    For `CN-TS` with `doubling`, $\widetilde{\mathbf{Z}}_{t'} = \widetilde{\mathbf{Z}}_{t'-1} + \sum_{i \in S_{t'}} \mathbf{g}(\mathbf{x}_{t',i}; \boldsymbol{\theta}_{t'-1}) \mathbf{g}(\mathbf{x}_{t',i}; \boldsymbol{\theta}_{t'-1})^\top / m_\tau$
    Update $\boldsymbol{\theta}_{t'}$ to minimize the loss (4) using gradient descent with η_τ for J_τ times
  end for
end if
Return: $\boldsymbol{\theta}_t$, $\mathbf{Z}_t$ or $\widetilde{\mathbf{Z}}_t$

Thus, the regret is bounded by $\widetilde{\mathcal{O}}(\max\{\widetilde{d}, \sqrt{\widetilde{d}K}\}\sqrt{2T})$. Similarly, for `CN-TS` with `doubling`, each sum has upper bound $\widetilde{\mathcal{O}}(\widetilde{d}_{\tau_n}\sqrt{\tau_n K})$ and the regret has upper bound $\widetilde{\mathcal{O}}(\widetilde{d}\sqrt{2TK})$.

G. Specific Examples of Combinaotrial Feedback Models

As mentioned in Remark 1, algorithms having a reward function satisfying Assumptions 1 and 2 encompasses various combinatorial feedback models, suggesting that these assumptions are not restrictive. In this section, we provide specific examples.

G.1. Semi-bandit Model

In the semi-bandit setting, after choosing a superarm, the agent observes all of the scores (or feedback) associated with the superarm and receives a reward as a function of the scores. The main text of this paper describes how our algorithms cover semi-bandit feedback models. Recall that in semi-bandit setting, if the feature vectors are independent then the score of each arm is independent. Meanwhile, in ranking models (or click models), chosen arms may have a position within the superarm, and the scores of arms may depend on its own attractiveness as well as its position.

G.2. Document-based Model

The document-based model is a click model that assumes the scores of an arm are identical to its attractiveness. The attractiveness of an arm is determined by the context of arm. Formally, for each arm $i \in [N]$, let $\alpha(\mathbf{x}_{t,i}) \in [0,1]$ be the attractiveness of arm i at time t. Then the document-based model assumes that the score function of $\mathbf{x}_{t,i}$ in the k-th position is defined as

$$h(\mathbf{x}_{t,i}, k) = \alpha(\mathbf{x}_{t,i}) \, \mathbb{1}(k \leq K) \,. \tag{21}$$

Note that h in Eq.(21) is bounded in $[0,1]$. Since a neural network is a universal approximator, we can utilize neural networks to estimate the score of arm i in position k as follows:

$$\hat{h}(\mathbf{x}_{t,i}, k) = f(\mathbf{x}_{t,i}, k; \boldsymbol{\theta}_{t-1}) \,.$$

Note that for any $k \in [K]$, the score of an arm only depends on the attractiveness of the arm. Hence, our algorithms can be directly applicable to the document-based model without any modification.

G.3. Position-based Model

In the document-based model, the score of an arm is invariant to the position within the super arm. However, in the position-based model, the score of a chosen arm varies depending on its position. Let $\chi : [K] \to [0,1]$ be a function that measures the quality of a position within the super arm. The position-based model assumes that the score function of a chosen arm associated to $\mathbf{x}_{t,i}$ and located in the k-th position is defined as

$$h(\mathbf{x}_{t,i}, k) = \alpha(\mathbf{x}_{t,i})\chi(k)\,. \tag{22}$$

Note that the score of an arm can change as its position moves within the superarm. We can slightly modify our suggested algorithms to reflect this. First, we introduce a modified neural network $\dot{f}(\mathbf{x}_{t,i}, k; \boldsymbol{\theta}_{t-1})$ that estimates the score of each arm at every available position. By this, the action space of each round increases from N to NK. The regret bound only changes as much as the action space increases. Denote the gradient of $\dot{f}(\mathbf{x}_{t,i}, k; \boldsymbol{\theta}_{t-1})$ as $\dot{\mathbf{g}}(\mathbf{x}_{t,i}, k; \boldsymbol{\theta}_{t-1})$.

Furthermore, we replace the oracle to $\dot{\mathbb{O}}_{\mathcal{S}}\left(\{u_{t,i}(k) + e_t\}_{i\in[N],k\in[K]}\right)$ that considers the position of the arms. The oracle $\dot{\mathbb{O}}_{\mathcal{S}}$ chooses only one arm for one position. Also, an arm that has been chosen for a certain position cannot be chosen for another position. As an optimization problem having the above constraints can be solved with linear programming, $\dot{\mathbb{O}}_{\mathcal{S}}\left(\{u_{t,i}(k) + e_t\}_{i\in[N],k\in[K]}\right)$ can compute exact optimization within polynomial time. Modified algorithm for a position-based model is described in Algorithm 6.

Algorithm 6 Combinatorial neural bandits for for position-based model

Initialize as Algorithm 1
for $t = 1, ..., T$ **do**
  Observe $\{\mathbf{x}_{t,i}\}_{i\in[N]}$
  if Exploration == UCB **then**
    Compute $u_{t,i}(k) = \dot{f}(\mathbf{x}_{t,i}, k; \boldsymbol{\theta}_{t-1}) + \gamma_{t-1}\|\dot{\mathbf{g}}(\mathbf{x}_{t,i}, k; \boldsymbol{\theta}_{t-1})/\sqrt{m})\|_{\mathbf{Z}_{t-1}^{-1}}$ for $i \in [N], k \in [K]$
    Let $S_t = \dot{\mathbb{O}}_{\mathcal{S}}\left(\{u_{t,i}(k) + e_t\}_{i\in[N],k\in[K]}\right)$
  else if Exploration == TS **then**
    Compute $\sigma_{t,i}^2(k) = \lambda\dot{\mathbf{g}}(\mathbf{x}_{t,i}, k; \boldsymbol{\theta}_{t-1})^\top \widetilde{\mathbf{Z}}_{t-1}^{-1}\dot{\mathbf{g}}(\mathbf{x}_{t,i}, k; \boldsymbol{\theta}_{t-1})/m$ for $i \in [N], k \in [K]$
    Sample $\{\widetilde{v}_{t,i}^{(j)}(k)\}_{j=1}^M$ independently from $\mathcal{N}(\dot{f}(\mathbf{x}_{t,i}, k; \boldsymbol{\theta}_{t-1}), \nu^2\sigma_{t,i}^2(k))$ for $i \in [N], k \in [K]$
    Compute $\widetilde{v}_{t,i}(k) = \max_j \widetilde{v}_{t,i}^{(j)}(k)$ for $i \in [N], k \in [K]$
    Let $S_t = \dot{\mathbb{O}}_{\mathcal{S}}\left(\{\widetilde{v}_{t,i}(k) + \epsilon\}_{i\in[N],k\in[K]}\right)$
  end if
  Play super arm S_t and observe $\{v_{t,i}(k_i)\}_{i\in S_t}$
  (UCB) Update $\mathbf{Z}_t = \mathbf{Z}_{t-1} + \sum_{i\in S_t}\dot{\mathbf{g}}(\mathbf{x}_{t,i}, k_i; \boldsymbol{\theta}_{t-1})\dot{\mathbf{g}}(\mathbf{x}_{t,i}, k_i; \boldsymbol{\theta}_{t-1})^\top/m$
  (TS) Update $\widetilde{\mathbf{Z}}_t = \widetilde{\mathbf{Z}}_{t-1} + \sum_{i\in S_t}\dot{\mathbf{g}}(\mathbf{x}_{t,i}, k_i; \boldsymbol{\theta}_{t-1})\dot{\mathbf{g}}(\mathbf{x}_{t,i}, k_i; \boldsymbol{\theta}_{t-1})^\top/m$
  Update $\boldsymbol{\theta}_t$ to minimize the loss in Eq.(4) using gradient descent with η for J times
end for

G.4. Cascade Model

In the cascade model, the agent suggests arms to a user one-by-one in order of the positions of the arms within the superarm. The user scans the arms one-by-one until she selects an arm that she likes, which ends the suggestion procedure. Note that the suggestion procedure potentially may end before the agent shows all the arms in the superarm to the user. Also, the user may not select any arm after she scans all the arms in the superarm. Hence, unlike the previously mentioned models, where the agent receives all of the scores of the chosen arms, in the cascade model, the agent only receives the scores of the arms observed by the user.

Let us assume that the score the agent receives when the user selects an arm in the 1-st position is 1. In case the same arm is in the k-th position, the score the agent receives when the user selects the same arm must be less than 1. To reflect this feature, we consider a position discount factor $\psi_k \in [0,1], k \leq K$ that is multiplied to the attractiveness of the arm. The observed score of an arm is determined by its attractiveness and the position discount factor that is multiplied to it. The mechanism estimating the attractiveness using a neural network is same as the one for the semi-bandits. The only difference is that the agent only receives the discounted scores of the arms observed by the user.

Algorithm 7 Combinatorial neural bandits for for cascade feedback model

Initialize as Algorithm 1, $\{\psi_k \in [0,1]\}_{k\in[K]}$: position discount factors
for $t = 1, ..., T$ **do**
  Observe $\{\mathbf{x}_{t,i}\}_{i\in[N]}$
  if Exploration == UCB **then**
    Compute $u_{t,i} = f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}) + \gamma_{t-1}\|\mathbf{g}(\mathbf{x}_{t,i}, k;\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}}$ for $i \in [N]$
    Let $S_t = \mathbb{O}_{\mathcal{S}}\left(\{u_{t,i} + e_t\}_{i\in[N]}\right)$
  else if Exploration == TS **then**
    Compute $\sigma_{t,i}^2 = \lambda \mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})^\top \widetilde{\mathbf{Z}}_{t-1}^{-1}\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})/m$ for $i \in [N]$
    Sample $\{\widetilde{v}_{t,i}^{(j)}\}_{j=1}^M$ independently from $\mathcal{N}(f(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1}), \nu^2\sigma_{t,i}^2)$ for $i \in [N]$
    Compute $\widetilde{v}_{t,i} = \max_j \widetilde{v}_{t,i}^{(j)}$ for $i \in [N]$
    Let $S_t = \mathbb{O}_{\mathcal{S}}\left(\{\widetilde{v}_{t,i} + \epsilon\}_{i\in[N]}\right)$
  end if
  Play super arm S_t and observe $\mathfrak{F}_t$, $\{\psi_k v_{t,k}\}_{k\in[\mathfrak{F}_t]}$
  (UCB) Update $\mathbf{Z}_t = \mathbf{Z}_{t-1} + \sum_{k\in[\mathfrak{F}_t]} \mathbf{g}(\mathbf{x}_{t,k};\boldsymbol{\theta}_{t-1})\mathbf{g}(\mathbf{x}_{t,i};\boldsymbol{\theta}_{t-1})^\top/m$
  (TS) Update $\widetilde{\mathbf{Z}}_t = \widetilde{\mathbf{Z}}_{t-1} + \sum_{k\in[\mathfrak{F}_t]} \mathbf{g}(\mathbf{x}_{t,k};\boldsymbol{\theta}_{t-1})\mathbf{g}(\mathbf{x}_{t,k};\boldsymbol{\theta}_{t-1})^\top/m$
  Update $\boldsymbol{\theta}_t$ to minimize the loss in Eq.(4) using gradient descent with η for J times
end for

Suppose that the user selects $\mathfrak{F}_t$-th arm. Then the agent observes the discounted scores for the first $\mathfrak{F}_t$ arms in S_t. Update is based on the discounted scores, $\psi_k v_{t,k}, k \le \mathfrak{F}_t$. An adjusted Algorithm for the cascade model is described in Algorithm 7. In addition, in case we have no information of the position discount factor, we can deal with the cascade model same as the position-based model.

H. Additional Related Work

As mentioned in Section 1, the proposed methods are the first neural network-based combinatorial bandit algorithms with regret guarantees. As for the previous combinatorial TS algorithms, Wen et al. (2015) proposed a TS algorithm for a contextual combinatorial bandits with semi-bandit feedback and a linear score function. However, the regret bound for the algorithm is only analyzed in the Bayesian setting (hence establishing the Bayesian regret) which is a weaker notion of regret and much easier to control in combinatorial action settings. To our knowledge, Oh & Iyengar (2019) was the first work to establish the worst-case regret bound for a variant of contextual combinatorial bandits, multinomial logit (MNL) contextual bandits, utilizing the optimistic sampling procedure similar to `CN-TS`. Yet, our proposed algorithm differs from Oh & Iyengar (2019) in that we sample directly from the score space rather than the parameter space which avoids the computational complexity of sampling a high-dimensional network parameters. More importantly, Oh & Iyengar (2019) exploit the structure of the MNL choice feedback model to derive the regret bound whereas we address a more general semi-bandit feedback without any assumptions on the structure of the feedback model.

I. Additional Experiments

In Experiment 1, the linear combinatorial bandit algorithms perform worse than our proposed algorithms, even for the linear score function. One of the possible reasons for this is that the neural network based algorithms use much larger number of parameters than the linear model based algorithms, overparametrized for the problem setting. Overparametrized neural networks have been shown to have superior generalization performances. See Allen-Zhu et al. (2019b;a). Note that the regret performance is about the generalization to the unseen data rather than it is about the fit to the existing data. In this aspect, overparameterized neural network can show superior performance over the linear model. This is supported by Figure 3. In Figure 3, we demonstrate the empirical performances of `CN-TS` ans `CombLinTS` as the network width m decreases. We can see that by decreasing m, the results of the neural network models and linear models become more similar, i.e., the gap between the regrets reduce.

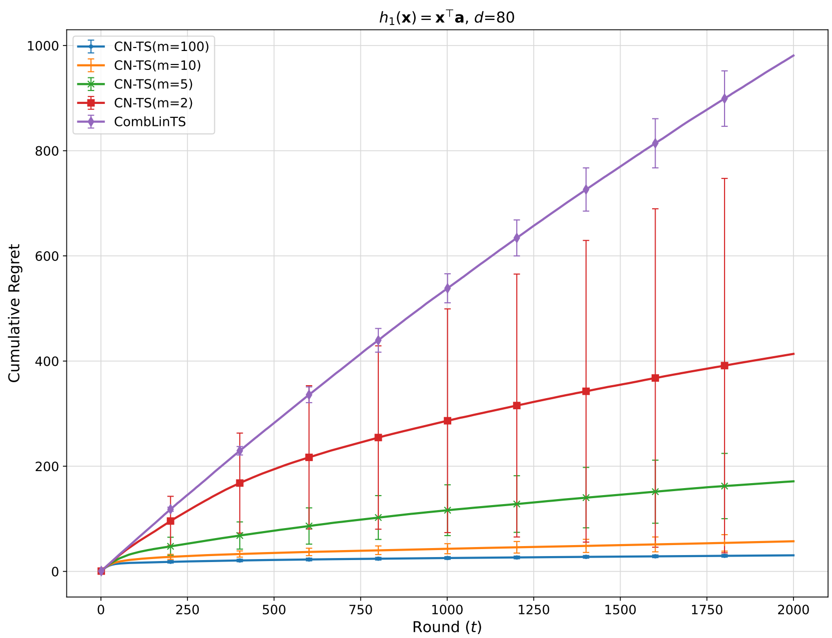

Figure 3. Cumulative regret of `CN-TS` and `CombLinTS` with respect to the network width (m).