

---

# Differentiable Linear Bandit Algorithm

---

**Kaige Yang**  
 University College London  
 kaige.yang.11@ucl.ac.uk

**Laura Toni**  
 University College London  
 l.toni@ucl.ac.uk

## Abstract

Upper Confidence Bound (UCB) is among the most commonly used methods for linear multi-arm bandit problems. While conceptually and computationally simple, this method highly relies on the confidence bounds, failing to strike the optimal exploration-exploitation if these bounds are not properly set. In the literature, confidence bounds are typically derived from concentration inequalities based on assumptions on the reward distribution, *e.g.*, sub-Gaussianity. The validity of these assumptions however is unknown in practice. In this work, we aim at learning the confidence bound in a data-driven fashion, making it adaptive to the actual problem structure. Noting that existing UCB-typed algorithms are not differentiable with respect to confidence bound, we first propose a novel differentiable linear bandit algorithm. Then, we introduce a gradient estimator, which allows permit to learn the confidence bound via iterative gradient ascent. Theoretically, we show that the proposed algorithm achieves a  $\tilde{O}(\hat{\beta}\sqrt{dT})$  upper bound of  $T$ -round regret, where  $d$  is the dimension of arm features and  $\hat{\beta}$  is the learned size of the confidence bound. Empirical results show that  $\hat{\beta}$  is significantly smaller than its theoretical upper bound and proposed algorithms outperform baseline ones on both synthetic and real-world datasets.

## 1 Introduction

Multi-Arm Bandit (MAB) [5] is an online decision making problem, in which an agent selects arms sequentially and observes stochastic rewards as feedback. The goal of the agent is to maximize the expected cumulative reward over a number of trials. The expected reward of each arm is unknown *a priori* and it is learned from experience by the agent. As a consequence, the agent needs to balance the selection of arms to improve its knowledge (exploration) and the selection of the highest rewarding arm given the knowledge acquired till thus far (exploitation). This is formalized as the so-called exploration-exploitation trade-off. Bandit algorithms are designed to strike this trade-off. One class of MAB problems is the linear MAB [8], in which each arm is described by a feature vector and the expected reward follows a linear model over its feature vector and an unknown parameter vector. Each arm's feature vector is known *a priori* by the agent and it is considered as a hint on the arm reward. The learning problem boils down to the agent inferring the unknown parameter vector, based on the history (selected arms and received rewards) and selecting arms accordingly.

One popular algorithm to solve linear MAB is the *Upper Confidence Bound* (UCB) [3] [8] [1]. Its popularity is motivated by its conceptual simplicity and strong theoretical guarantees. UCB-typed algorithms rely on the construction of an upper confidence bound, which is the estimated reward inflated based on the level of uncertainty of the estimate. At each decision opportunity, the agent selects the arm with the highest upper confidence bound. This reflects the *Optimism in Face of Uncertainty* principle. In such way, either the arm with high estimated reward (exploitation) or high uncertainty (exploration) is selected. However, to properly balance between exploration and exploitation, it is fundamental to establish a tight confidence bound [16].

In most existing works, confidence bounds are derived from concentration inequalities [1] [4] [18] given *a priori* assumptions on the reward distribution (e.g., sub-Gaussianity). These bounds achieve strong minimax theoretical guarantees, outperforming competitor algorithms such as LinTS [2]. While these bounds are essential for a theoretical analysis, they do not necessarily translate into practice. In fact, these constructed confidence bounds are typically conservative *in practice*, as noted in [19] [14]. This is because concentration inequalities are usually built based on given reward distributions instead of the actual data (or problem structure). This results in non-adaptive and potentially wide confidence bounds which in turn lead to suboptimal performance in practice.

Alternatively, in this work we aim to learn the confidence bound in a data-driven fashion making it adaptive to the actual problem structure. Inspired by [6], we aim at having a parametrized and differentiable cumulative reward function with respect to the confidence bound, which can then be optimized. The key challenge is that existing UCB-typed algorithms are non-differentiable with respect to the confidence bound, mainly due to maximization of the UCB index (i.e., due to the presence of the  $\arg \max$  operator in the OFUL [1], LinUCB [8]). To address this, we propose a novel differentiable UCB-typed linear bandit algorithm and introduce a gradient estimator which enables the confidence bound to be learned via gradient ascent.

Our proposed algorithm contains two core components. First, we consider a more informative UCB-based index than the classical UCB index used in OFUL [1], LinUCB [8], which not only summarizes the history of each arm but also differentiates arms to be suboptimal arms and non-suboptimal arms. Second, we consider a softmax function, which transforms each index into a probability distribution, where the probability for each suboptimal arm to be selected is arbitrary small. Conversely, the probability for a non-suboptimal arm to be selected is greater for arms with larger index. The key idea is that the exploration is conducted by selecting arms with large index more often than others. The exploitation is achieved by soft-eliminating suboptimal arms (arbitrary small probability to be selected). The softmax function ensures the differentiability of the reward function, paving the way to learn confidence bound via gradient ascent. Based on this, we provide two linear bandit algorithms for learning confidence bound in both offline and online settings. Theoretically, we provide a regret upper bound for the offline learning setting.

In summary, we make the following contributions:

- We propose a novel UCB-typed linear bandit algorithm where the expected cumulative reward is a differentiable function of the confidence bound.
- We introduce a gradient estimator and show how the confidence bound can be learned via gradient ascent both in offline/online settings.
- Theoretically, we prove a  $\tilde{O}(\hat{\beta}\sqrt{dT})$  upper bound of  $T$ -rounds regret where  $\hat{\beta}$  is the learned size of the confidence bound in the offline setting.
- Empirically, we show  $\hat{\beta}$  is significantly smaller than its theoretical upper bound, leading to substantially lower cumulative regrets with respect to state-of-the-art baselines on synthetic and real-world datasets.

**Notation:**  $[K]$  mean the set  $\{1, 2, \dots, K\}$ . Arm is indexed by  $i, j \in \mathcal{A}$ . We use boldface lower letter, e.g.,  $\mathbf{x}$ , to denote vector and boldface upper letter, e.g.,  $\mathbf{M}$ , to denote matrix. For a positive definite matrix  $\mathbf{M} \in \mathbb{R}^{d \times d}$  and a vector  $\mathbf{x} \in \mathbb{R}^d$ , we denote the weighted 2-norm by  $\|\mathbf{x}\|_{\mathbf{M}} = \sqrt{\mathbf{x}^T \mathbf{M} \mathbf{x}}$ . Each arm  $k$  is represented by the feature vector  $\mathbf{x}_k \in \mathbb{R}^d$ . We denote by  $\mathbb{P}$  and  $\mathbb{E}$  the probability distribution and the expectation operator, respectively.

## 2 Related work

Our work is inspired by [6], which was the first attempt in addressing policy-gradient optimization of bandit policies via differentiable bandit algorithm. However, there are fundamental differences between [6] and our work. First, authors proposed a differential bandit framework for Bayesian MAB problem, which is not directly applicable to linear MAB problems. Conversely, we propose a differentiable UCB-typed linear bandit algorithm. Second, the main goal of [6] is to learn the learning rate (coldness-parameter) of the softmax function, while our algorithm aims at learning the size of the confidence bound. Third, we propose algorithms for both offline and online settings, while [6] covered the offline setting only. Moreover, in [6] a regret analysis was provided for MAB with two

arms. In contrast, we provide a regret analysis for linear MAB with arbitrary finite number of arms in offline setting.

Another work focused on data-dependent UCB is [12]. Authors proposed an algorithm called `bootstrappedUCB`. In [12], the stochastic reward is assumed to be sub-Weibull random variable. Multiplier bootstrap was employed to approximate the reward distribution. The bootstrapped quantile acted as UCB to facilitate exploration. Their algorithm was deployed on both MAB and linear MAB problems, while regret analysis covered MAB only. Similar to this work, other bootstrap techniques were employed [7] [9] [22]. Although aiming to the same goal (data-dependent UCB), these works are fundamentally different from our approach. Our algorithm is a differentiable bandit algorithm where we rely on gradient estimator to learn UCB. Their algorithm is non-differentiable, relying on the bootstrapped quantile of the assumed reward distribution to construct UCB.

Bootstrap techniques were used also for Thompson Sampling exploration in [19], in which author proposed the `BootstrapThompson` algorithm for MAB. Bootstrap techniques were used to sample observations from historical and pseudo observations to approximate the posterior distribution which was then used to encourage exploration. As an extension, [23] generalized this technique to Gaussian reward MAB, while [13] and [14] proposed an extension to contextual linear bandit, achieving the same regret bound of `LinTS` [2]. The problem they aimed to address was the computational infeasibility of inferring posterior distribution when reward follows nonlinear models. This departs from our goal, which is rather learning the confidence bound from data.

Our work can be viewed as a subtle combination of `EXP3` [5] and `Phased Elimination`<sup>1</sup> [17]. `EXP3` was designed for MAB, where arms with higher empirical averaged reward are signed with larger probability by softmax function. The coldness-parameter of softmax function is a tunable hyper-parameter chosen by the user. In our work, we propose a novel scheme to set this parameter automatically in a data-driven fashion. Moreover, although `Exp3` is a differentiable bandit algorithm, it is not a UCB-typed algorithm. `Phased Elimination` eliminates suboptimal arms based on the same index as ours and selects non-suboptimal arms uniformly (pure exploration). There are several fundamental differences between this approach and our work: *i*) the confidence bound in our work is learned from data and not from concentration inequalities – leading to a less conservative bound; *ii*) `Phased Elimination` is a non-differentiable algorithm; *iii*) `Phased Elimination` achieves optimality in a worst case scenario (minmax regret) while our algorithm get an empirical gain being data dependent.

In summary, to the best of our knowledge, our work is the first differentiable UCB-typed linear bandit algorithm which enables confidence bound to be learned purely from data without relying on concentration inequalities and assumptions on the form of reward distribution.

### 3 Problem setting

We consider the stochastic linear bandit with an arm set  $\mathcal{A}$  and a time horizon of  $T$ -rounds. The arm set contains  $K$  arms, i.e.,  $|\mathcal{A}| = K$ , where  $K$  could be large. Each arm  $i \in \mathcal{A}$  is associated with a known feature vector  $\mathbf{x}_i \in \mathbb{R}^d$ . The expected reward of each arm  $\mu_i = \mathbf{x}_i^T \boldsymbol{\theta}$  follows a linear relationship over  $\mathbf{x}_i$  and an unknown parameter vector  $\boldsymbol{\theta}$ . Similarly to other works in the bandit literature, we assume that arm feature and parameter vector are bounded  $\|\mathbf{x}_i\|_2 \leq L$  and  $\|\boldsymbol{\theta}\|_2 \leq C$ , where  $L > 0$  and  $C > 0$ . At the beginning of each decision opportunity  $t \in [T]$ , the learning agent selects one arm  $i \in \mathcal{A}$  within the arm set  $\mathcal{A}$ . Upon this selection, the agent observes the instantaneous reward  $y_t \in [0, 1]$ , which is drawn independently from a distribution with unknown mean  $\mu_i = \mathbf{x}_i^T \boldsymbol{\theta}$ . The agent aims to maximize the expected cumulative reward over the time horizon  $T$ . Namely,

$$Y_T = \sum_{t=1}^T \mathbb{E}[y_t] \quad (1)$$

This is equivalent to minimize the expected cumulative regret which measures the difference between the expected cumulative reward if the optimal arm were always selected and the agent's expected cumulative reward. Denoting by  $\mu_* = \max_{i \in \mathcal{A}} \mathbf{x}_i^T \boldsymbol{\theta}$  the expected reward of the optimal arm, we get

$$R_T = T\mu_* - \sum_{t=1}^T \mathbb{E}[y_t]. \quad (2)$$

---

<sup>1</sup>Algorithm: `Phased elimination with G-optimal exploration` page. 258 [17]

---

**Algorithm 1: SoftUCB**

---

**Input** :  $\beta, \mathcal{A}, K, T, \alpha$ .**Initialization** :  $\mathbf{V}_0 = \alpha \mathbf{I} \in \mathbb{R}^{d \times d}$ ,  $\mathbf{b}_0 = \mathbf{0} \in \mathbb{R}^d$ ,  $\hat{\boldsymbol{\theta}}_0 = \mathbf{0} \in \mathbb{R}^d$ ,  $\gamma_0 = 0$ .**for**  $t \in [1, T]$  **do**

1. Find  $S_{i,t}, \forall i \in \mathcal{A}$  via Eq. 7 with  $\beta$ .
2. Find  $\pi_t$  via Eq. 8 with  $\gamma_{t-1}$ .
3. Select arm  $i_t \in \mathcal{A}$  randomly following  $\pi_t$  and receive payoff  $y_t$ .
4. Update  $\mathbf{V}_t \leftarrow \mathbf{V}_t + \mathbf{x}_t \mathbf{x}_t^T$ ,  $\mathbf{b}_t \leftarrow \mathbf{b}_{t-1} + \mathbf{x}_t y_t$  and  $\hat{\boldsymbol{\theta}}_t = \mathbf{V}_t^{-1} \mathbf{b}_t$ .
5. Update  $\gamma_t$  via Eq. 9.

**end**

---

**Upper Confidence Bound (UCB).** The upper confidence bound algorithm, e.g., **OFUL** [1], is designed based on the *Optimism in Face of Uncertainty* principle. The key aspect is to construct a confidence bound of the estimated reward of each arm. Formally, at each round  $t$ , the confidence bound is defined as

$$|\hat{\mu}_{i,t} - \mu_i| \leq \beta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}, \quad \forall i \in \mathcal{A} \quad (3)$$

where  $\hat{\mu}_{i,t}$  is the estimate of the reward of arm  $i$  at round  $t$  and  $\mathbf{V}_t = \sum_{s=1}^t \mathbf{x}_s \mathbf{x}_s^T$  is the Gram matrix up to round  $t$ . Then, the agent selects the arm with the highest upper confidence bound as follows

$$i_t = \arg \max_{i \in \mathcal{A}} \hat{\mu}_{i,t} + \beta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \quad (4)$$

It is well known that the tighter the bound in Eq. 3, the better the balance between exploration and exploitation [16]. Most existing confidence bounds are established based on concentration inequalities. e.g., Hoeffding inequality [4], self-normalized [1], Azuma Inequality [17], Bernstein inequality [18]. As a specific example, under the assumption of the stochastic reward to be a  $R$ -sub-Gaussian variable, one of the state-of-the-art high probability upper bound of  $\beta$ , derived based on properties of self-normalized martingale, was given by [1]:

$$\beta \leq R \sqrt{2 \log \left( \frac{1}{\delta} \right) + d \log \left( 1 + \frac{T}{d} \right)} + \sqrt{\alpha} C \quad (5)$$

where  $\alpha$  is a regularizer parameter of least-square estimator,  $1 - \delta$  is the probability of which Eq. 3 holds and  $\|\boldsymbol{\theta}\|_2 \leq C$ . The tightness of this (and other bounds) relies on the validity of assumptions on the reward distribution, which is unfortunately unknown in practice. Alternatively, we aim at learning the confidence bound, i.e.,  $\beta$ , in a data-driven fashion without any *a priori* assumption on the unknown reward distribution except the linearity function of the mean reward, i.e., is  $\mu_i = \mathbf{x}_i^T \boldsymbol{\theta}, \forall i \in \mathcal{A}$ .

## 4 Algorithms

In this section, we first present our proposed algorithm whose expected cumulative reward is a differentiable function of the confidence bound. Then, we provide a gradient estimator which enables confidence bound to be learned via gradient ascent. Next, we propose two algorithms to learn the confidence bound in offline and online settings, respectively. Finally, we prove a regret upper bound for offline learning setting.

### 4.1 Differentiable Algorithm

Our proposed algorithm named **SoftUCB** is shown in Algorithm 2. **SoftUCB** contains two core components: an UCB-based index  $S_{i,t}$  and an arm selection policy  $\pi_t$ . Formally, for  $i \in \mathcal{A}$ ,  $\hat{\mu}_{i,t} = \mathbf{x}_i^T \hat{\boldsymbol{\theta}}_t$  where  $\hat{\boldsymbol{\theta}}_t = \mathbf{V}_t^{-1} \sum_{s=1}^t \mathbf{x}_s y_s$  is the least-square estimator and  $\mathbf{V}_t^{-1} = \sum_{s=1}^t \mathbf{x}_s \mathbf{x}_s^T$  is the Gram matrix up to round  $t$ . Let denote by  $i_* = \arg \max_{i \in \mathcal{A}} \hat{\mu}_{i,t} - \beta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}$  the arm with the largest lower confidence bound at round  $t$ . Let us also define

$$\phi_{i,t} = \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} + \|\mathbf{x}_{i_*}\|_{\mathbf{V}_t^{-1}} \quad \text{and} \quad \hat{\Delta}_{i,t} = \hat{\mu}_{i_*,t} - \hat{\mu}_{i,t} \quad (6)$$

where  $\beta$  is the confidence bound defined in Eq. 3 and  $\hat{\Delta}_{i,t}$  is the estimated reward gap between  $i_*$  and  $i$ . Equipped with the above notations, we are now ready to introduce the UCB-based index  $S_{i,t}$  defined as

$$S_{i,t} = \beta\phi_{i,t} - \hat{\Delta}_{i,t}. \quad (7)$$

It is worth noting that  $S_{i,t}$  is more informative than classical UCB index provided Eq. 4, because of the following two key properties: *i*),  $S_{i,t}$  differentiates arms into suboptimal arms and non-suboptimal arms. Specifically,  $S_{i,t} < 0$  identifies arms which are suboptimal,  $\Delta_i = \mu_* - \mu_i > 0$ , and therefore could be eliminated (i.e., not selected by the agent); *ii*),  $S_{i,t} \geq S_{j,t} \geq 0$  implies that the upper confidence bound  $\hat{\mu}_{i,t} + \beta\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \geq \hat{\mu}_{j,t} + \beta\|\mathbf{x}_j\|_{\mathbf{V}_t^{-1}}$  and therefore arm  $i$  is more likely to be selected, in line with the *Optimism in Face of Uncertainty* principle. These two properties are stated formally in Lemma 1.

**Lemma 1.** *If  $S_{i,t} < 0$ , arm  $i$  is a suboptimal arm, i.e.,  $\mu_* - \mu_i > 0$ . If  $S_{i,t} \geq S_{j,t} \geq 0$ , then the upper confidence bound  $\hat{\mu}_{i,t} + \beta\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \geq \hat{\mu}_{j,t} + \beta\|\mathbf{x}_j\|_{\mathbf{V}_t^{-1}}$ . The proof is provided in Appendix A.*

We now describe the arm selection strategy. At each round  $t \in [T]$ , the probability for arm  $i$  to be selected is defined as

$$p_{i,t} = \frac{\exp(\gamma_t S_{i,t})}{\sum_{j=1}^K \exp(\gamma_t S_{j,t})} \quad (8)$$

where  $\gamma_t > 0$  is the coldness-parameter controlling the concentration of the distribution (policy)  $\boldsymbol{\pi}_t = [p_{1,t}, p_{2,t}, \dots, p_{K,t}]$ , and it is set as

$$\gamma_t = \frac{\log\left(\frac{\delta|\mathcal{L}_t|}{1-\delta}\right)}{\tilde{S}_{\max,t}} \quad (9)$$

where at each round  $t$ , the arm set  $\mathcal{A}$  is divided into two subsets  $\mathcal{U}_t$  and  $\mathcal{L}_t$  with  $\mathcal{U}_t \cup \mathcal{L}_t = \mathcal{A}$  and  $\mathcal{U}_t \cap \mathcal{L}_t = \emptyset$ . Namely,  $\mathcal{L}_t$  is the set of suboptimal arms (i.e.,  $i \in \mathcal{L}_t$  if  $S_{i,t} < 0$ ) and  $\mathcal{U}_t$  is the set of non-suboptimal arms (i.e.,  $i \in \mathcal{U}_t$  if  $S_{i,t} \geq 0$ ).  $\tilde{S}_{\max,t} = \max_{i \in \mathcal{U}_t} S_{i,t}$ ,  $|\mathcal{L}_t|$  is the cardinality of  $\mathcal{L}_t$  and  $\delta$  is a probability hyper-parameter explained in the following Lemma.

**Lemma 2.** *At any round  $t \in [T]$ , for any  $\delta \in (0, 1)$ , setting  $\gamma_t \geq \log\left(\frac{\delta|\mathcal{L}_t|}{1-\delta}\right)/\tilde{S}_{\max,t}$  guarantees that  $p_{\mathcal{U}_t} = \sum_{i \in \mathcal{U}_t} p_{i,t} \geq \delta$  and  $p_{\mathcal{L}_t} = \sum_{i \in \mathcal{L}_t} p_{i,t} < 1 - \delta$ . The proof is provided in Appendix B.*

According to Lemma 2, Eq. 9 guarantees that suboptimal arms ( $i \in \mathcal{L}_t$ ) are selected with an arbitrary small probability (i.e.,  $p_{\mathcal{L}_t} < 1 - \delta \approx 0$  when  $\delta \approx 1$ ). This leads to a soft-elimination of suboptimal arms. Furthermore, a positive  $\gamma_t$  guarantees  $p_{i,t} \geq p_{j,t}$  if  $S_{i,t} \geq S_{j,t} \geq 0$ ,  $\forall i, j \in \mathcal{U}_t$  which obeys the *Optimism in Face of Uncertainty* principle.

Overall, **SoftUCB** (soft-) eliminates suboptimal arms and selects non-suboptimal arms according to the index in Eq. 7 which favors the selection of arms with either high estimated reward or high uncertainty.

## 4.2 Gradient Estimator of $\beta$

We now show that the expected cumulative reward of **SoftUCB** is a differentiable function over  $\beta$  and introduce a gradient estimator. Formally, given the expected cumulative reward defined in Eq. 1 and **SoftUCB** described above, we have the optimization objective defined as

$$\max_{\beta} Y_T = \max_{\beta} \sum_{t=1}^T \mathbb{E}[y_t] = \max_{\beta} \sum_{t=1}^T \sum_{i=1}^K p_{i,t} \mu_i, \quad s.t. \quad |\mu_i - \hat{\mu}_{i,t}| \leq \beta\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}, \quad \forall i \in \mathcal{A}, t \in [T] \quad (10)$$

The imposed constraint ensures that  $\beta\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}$  is indeed an actual upper confidence bound (UCB) at any round  $t \in [T]$  for any arm  $i \in \mathcal{A}$ . Applying the Lagrange multipliers gives the new objective:

$$\max_{\beta} \sum_{t=1}^T \sum_{i=1}^K p_{i,t} \mu_i - \eta \left( |\mu_i - \hat{\mu}_{i,t}| - \beta\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \right), \quad s.t. \quad \eta > 0 \quad (11)$$

The gradient of  $\beta$ , denoted as  $g(\beta)$ , can be derived as (proof in Appendix C):

$$g(\beta) = \sum_{t=1}^T \sum_{i=1}^K p_{i,t} \mu_i \left( \gamma_t \phi_{i,t} - \frac{\sum_{j=1}^K \gamma_t \phi_{j,t} \exp(\gamma_t S_{j,t})}{\sum_{j=1}^K \exp(\gamma_t S_{j,t})} \right) + \eta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \quad (12)$$

Note that  $\mu_i$  is unknown in practice and it is therefore replaced by its empirical estimate  $\hat{\mu}_{i,t}$ , leading to the following gradient estimator

$$\hat{g}(\beta) = \sum_{t=1}^T \sum_{i=1}^K p_{i,t} \hat{\mu}_{i,t} \left( \gamma_t \phi_{i,t} - \frac{\sum_{j=1}^K \gamma_t \phi_{j,t} \exp(\gamma_t S_{j,t})}{\sum_{j=1}^K \exp(\gamma_t S_{j,t})} \right) + \eta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \quad (13)$$

The gradient estimator  $\hat{g}(\beta)$  in Eq. 13 enables  $\beta$  to be learned via gradient ascent. As a stochastic gradient method, under standard condition of learning rate, e.g., RM [20], it is expected that  $\hat{\beta}$  converges to local optimum.

### 4.3 Training Settings

Equipped with the gradient estimator  $\hat{g}(\beta)$  (Eq. 13), we now show how to learn  $\beta$  in offline and online settings. The corresponding algorithms named **SoftUCB offline** and **SoftUCB online** are presented in Appendix E.

**Offline setting.** In this setting, multiple  $T$ -rounds trajectories of the bandit problem with the same arm set  $\mathcal{A}$  are used to train  $\beta$ , which is refined after each  $T$ -rounds trajectory. The key steps are to initialize  $\hat{\beta}_0$  and run **SoftUCB** on  $\mathcal{A}$  for  $N$  training trajectories – each trajectory containing  $T$ -rounds. After each trajectory  $n \in [N]$ , update  $\hat{\beta}_n \leftarrow \hat{\beta}_{n-1} + \lambda \hat{g}(\beta)$  via Eq. 13 where  $\lambda$  is the learning step. At the end of the training, run **SoftUCB** on  $\mathcal{A}$  with  $\hat{\beta} = \hat{\beta}_N$ .

As a result of the training, the value of  $\hat{\beta}$  is optimized in such a way that it maximizes the expected cumulative reward of arm set  $\mathcal{A}$ . Empirically, the  $\hat{\beta}$  to which the algorithm converges is substantial less than its theoretical upper bound Eq. 5. This translates into a significant regret reduction. In the following subsection we provide a theoretical regret upper bound of **SoftUCB offline**

While the above method is fully adaptive to the structure of  $\mathcal{A}$ , it provides a burden on the computational complexity. Specifically, the computational complexity **SoftUCB offline** is  $\mathcal{O}(NKT)$ , since we run **SoftUCB**  $N$  trajectories with  $K$  arms and  $T$  rounds in each trajectory. This is much higher than other linear algorithms such as **LinUCB** [1] and **LinTS** [2]. To mitigate this issue, we propose **SoftUCB online** which learns  $\beta$  within one trajectory in an online fashion.

**Online setting.** In this setting,  $\hat{\beta}$  is updated online during one  $T$ -rounds trajectory. Specifically,  $\hat{\beta}_0$  is initialized and **SoftUCB** on  $\mathcal{A}$  is run for  $T$  rounds. At the end of each round  $t \in [T]$ , update  $\hat{\beta}_t \leftarrow \hat{\beta}_{t-1} + \lambda \hat{g}_t(\beta)$  where  $\lambda$  is the learning step and  $\hat{g}_t(\beta)$  is the gradient estimator (Eq. 15 defined below). This reduces the computational complexity to  $\mathcal{O}(KT)$  since it does not require the  $N$ -training trajectories, which is at the same level of **OFUL** [1], **LinUCB** [8] and **LinTS** [2].

In this setting,  $Y_T = \sum_{t=1}^T \mathbb{E}[y_t]$ , the objective function we aim at maximizing, is not available before the end of the trajectory. To obviate to this problem, similarly to policy gradient methods for non-episodic reinforcement learning problems [21], we update  $\hat{\beta}$  to maximizes the average reward per round  $\hat{Y}_t$ . Formally, at each round  $t$ ,  $\hat{Y}_t$  consists of two parts: the observed cumulative reward up to round  $t$  and bootstrapped future reward under the current policy  $\pi_t = [p_{1,t}, p_{2,t}, \dots, p_{K,t}]$ . This translates in the following problem formulation

$$\begin{aligned} \max_{\beta} \hat{Y}_t &= \max_{\beta} \left( \sum_{s=1}^t \sum_{i=1}^K p_{i,s} \hat{\mu}_{i,s} + (T-t) \sum_{i=1}^K p_{i,t} \hat{\mu}_{i,t} \right) / T \\ \text{s.t. } & |\hat{\mu}_{i,t} - \mu_{i,t}| \leq \beta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}, \forall i \in \mathcal{A} \end{aligned} \quad (14)$$

The gradient estimator  $\hat{g}_t(\beta)$  at round  $t$  can be derived as

$$\hat{g}_t(\beta) = \frac{1}{T} \left( \sum_{s=1}^t \sum_{i=1}^K \hat{\mu}_{i,s} \nabla_{\beta} p_{i,s} + (T-t) \sum_{i=1}^K \hat{\mu}_{i,t} \nabla_{\beta} p_{i,t} + \eta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \right) \quad (15)$$

It is worth noting that, at the end of trajectory  $t = T$ , the  $\hat{Y}_t$  converges to  $Y_T$  in the offline setting.

Table 1: The comparison between  $\hat{\beta}$  (offline) and theoretical bound  $\tilde{\beta}$

$d = 5, T = 2^8$	$d = 5, T = 2^9$	$d = 5, T = 2^{10}$	$d = 10, T = 2^{10}$	$d = 15, T = 2^{10}$
$\hat{\beta} = \mathbf{0.5}$	$\hat{\beta} = \mathbf{0.6}$	$\hat{\beta} = \mathbf{0.9}$	$\hat{\beta} = \mathbf{1.1}$	$\hat{\beta} = \mathbf{1.2}$
$\tilde{\beta} = 2.56$	$\tilde{\beta} = 2.66$	$\tilde{\beta} = 2.76$	$\tilde{\beta} = 3.25$	$\tilde{\beta} = 3.61$

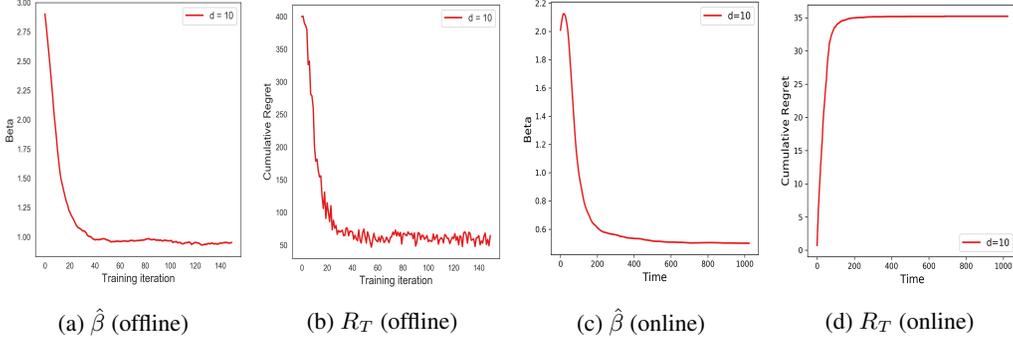


Figure 1: Learning curves of **SoftUCB offline** and **SoftUCB online**

#### 4.4 Theoretical Analysis

**Theorem 1.** Define  $\mathbb{E}[r_t] = \mathbb{E}[\mu_* - \sum_{i=1}^K p_{i,t}\mu_i]$  be the expected regret at round  $t \in [T]$ . Let  $\hat{\beta} = \beta_N$  be the confidence bound learned from the offline training setting after  $N$   $T$ -rounds trajectories. Let assume that  $\gamma_t$  follows Lemma 2 and  $\delta \approx 1$ . The cumulative regret of **SoftUCB** is bounded as

$$R_T = \sum_{t=1}^T \mathbb{E}[r_t] \leq 4\sqrt{2}\hat{\beta}\delta\sqrt{Td\log\left(\alpha + \frac{T}{d}\right)} = \tilde{O}\left(\hat{\beta}\sqrt{dT\log\left(1 + \frac{T}{d}\right)}\right) \quad (16)$$

where  $\tilde{O}(\cdot)$  hides absolute constant. The proof is contained in Appendix D.

Theorem 1 provides a regret upper bound of **SoftUCB** in the offline setting. To compare the regret bound with that of other algorithms, we show  $\hat{\beta}$  explicitly in the upper bound. Our regret bound scales with  $d$  and  $T$  as the regret bound  $\mathcal{O}(\beta\sqrt{dT})$  of existing UCB-typed algorithms, e.g., **OFUL** [1], **LinUCB** [8], **Giro** [13]. Since we make no assumption on the reward distribution, we can not derive a theoretical upper bound on  $\hat{\beta}$ . However, it is worth to noting that empirical results (in next section) show that  $\hat{\beta}$  is significantly smaller than its theoretical upper bound Eq. 5. The theoretical analysis for the online setting is left for future works.

## 5 Experiments

Our experimental evaluation aims to answer the following questions: (1) Does the learning curve of  $\hat{\beta}$  converge in offline and online settings? (2) Is  $\hat{\beta}$  lower than its theoretical counterpart? (3) How do our proposed algorithms perform compare to baseline ones?

In synthetic datasets, there are  $K = 50$  arms with feature vector drawn uniformly from  $[-1, 1]$ . The dimension of arm feature is set as  $d = 10, 20$ . Arm feature vectors are normalized to be unit vectors. The parameter vector  $\theta$  is generated as a random unit vector. The noise level is set as 0.5 and the regularizer parameter is  $\alpha = 1$ . We use two real-world datasets: **Jester** [11] and **Movielens** [15] (see Appendix G for more details). We compare the proposed algorithms with baseline ones, namely **LinUCB** [1], **LinTS** [2] and  $\epsilon$ -**greedy** [21]. The  $\beta$  in **LinUCB** is set as Eq. 5, **LinTS** follows [2], and  $\epsilon = 0.05$  in  $\epsilon$ -**greedy**.

Fig. 1 depicts the learning curves of  $\hat{\beta}$  and the corresponding  $R_T$  in both offline and online settings for the synthetic datasets. The feature dimension  $d = 10$ . In both settings,  $\hat{\beta}$  and  $R_T$  achieve

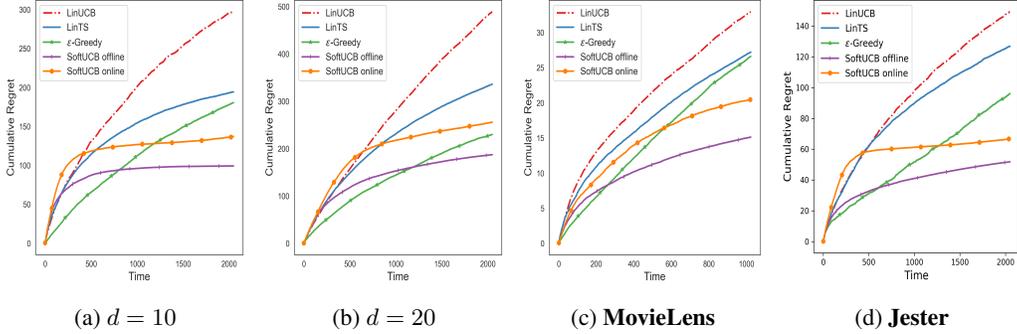


Figure 2: Performance of algorithms on synthetic and real-world datasets

convergence. Note that in offline setting,  $\hat{\beta}$  is optimized to maximize the expected cumulative reward Eq. 10, while in online setting,  $\hat{\beta}$  is optimized to maximize the average reward per round Eq. 14.

In Table 1, we compare  $\hat{\beta}$  obtained from offline training and its theoretical suggested  $\tilde{\beta}$  given by Eq. 5. Clearly,  $\hat{\beta}$  is significantly less than  $\tilde{\beta}$  consistently in all cases. This is because  $\hat{\beta}$  is adaptive to the structure of  $\mathcal{A}$ , while  $\tilde{\beta}$  is derived based on worst-case (minimax analysis). Note that the value of  $\hat{\beta}$  is highly data-dependent. The value we report here only valid for our experimental data. However, it is reasonable to expect  $\hat{\beta}$  less than  $\tilde{\beta}$  in general. The corresponding learning curves are shown in Appendix F.

In Fig. 2, the proposed algorithms converge to lower cumulative regret comparing with baselines. There are two reasons: First, the confidence bound  $\hat{\beta}$  is optimized. Second, the proposed algorithm eliminates (softly) suboptimal arms which accelerates the rate of convergence. It is worth noting that the regret of **SoftUCB online** is large at the initial phase. This is because at the beginning, when  $\gamma_0 = 0$ ,  $|\mathcal{L}_t| = 0$ , **SoftUCB online** selects arms uniformly which results in large regret. Later, when suboptimal arms are identified,  $|\mathcal{L}_t| > 0$ ,  $\gamma_t > 0$  according to Eq. 9. Suboptimal arms are soft-eliminated and non-suboptimal arms are selected following index Eq. 7 which controls the regret.

Finally, during our experiments, we noticed that the convergence of  $\hat{\beta}$  in both offline and online setting is sensitive to the Lagrange multiplier  $\eta$ . With large  $\eta$ , the gradient ascent algorithm fails in converging, this is because the gradient estimator Eq. 15 is dominated by  $\eta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}$ . On the other hand, too small  $\eta$  does not ensure the key constraint  $|\hat{\mu}_{i,t} - \mu_i| \leq \beta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}$ . This can lead to erroneously eliminating the optimal arm. Therefore, the hyper-parameter  $\eta$  needs to be tuned carefully during experiments.

## 6 Conclusion

We propose **SoftUCB**, a novel UCB-typed linear bandit algorithm based on an *adaptive* confidence bound, resulting in a less conservative algorithm respect to UCB-typed algorithms with *constructed* confidence bounds. The key novelty is to propose an expected cumulative reward which is a differentiable function of the confidence bound, and derive a gradient estimator, which enables confidence bound to be learned via gradient ascent. The estimated confidence bound  $\hat{\beta}$  can be updated under offline/online training settings with the proposed **SoftUCB offline** and **SoftUCB online**, respectively. Theoretically, we provide a  $\tilde{\mathcal{O}}(\hat{\beta}\sqrt{dT})$  regret upper bound of **SoftUCB** in the offline setting. Empirically, we show that  $\hat{\beta}$  is significantly less than its theoretical counterpart leading to a reduction of the cumulative regret compared to state-of-the-art baselines.

There are several directions for future work. First, our work can be combined with meta-learning algorithms, e.g., **MAML** [10], to learn a confidence bound which is adaptive to the common structure of a set of bandit tasks. Second, we believe our work can be generalized to reinforcement learning (RL) tasks where exploration and exploitation trade-off is a long standing challenge.

## 7 Broader Impact Discussion

Our work is an algorithm for multi-arm bandit (MAB) problem. On the novelty side, our work automates the exploration in bandit problems. Such algorithm could be used in recommendation system and clinic trials. On the positive side, our work could balance the exploration and exploitation trade-off in a problem dependently way, which might improve the customer satisfaction or patient's health care. On the negative side, depending to the deployed application, the recommended contents might be unsuitable for some users. To mitigate this issue, domain knowledge might be required to filter the recommended contents before releasing to users. Regarding the health care application, expert's supervision is essential to avoid any potential hazard.

## References

- [1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.
- [2] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.
- [3] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
- [4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.
- [5] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331. IEEE, 1995.
- [6] Craig Boutilier, Chih-Wei Hsu, Branislav Kveton, Martin Mladenov, Csaba Szepesvari, and Manzil Zaheer. Differentiable bandit exploration. *arXiv preprint arXiv:2002.06772*, 2020.
- [7] Richard Y Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.
- [8] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.
- [9] Adam N Elmachtoub, Ryan McNellis, Sechan Oh, and Marek Petrik. A practical method for solving contextual bandit problems using decision trees. *arXiv preprint arXiv:1706.04687*, 2017.
- [10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.
- [11] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *information retrieval*, 4(2):133–151, 2001.
- [12] Botao Hao, Yasin Abbasi Yadkori, Zheng Wen, and Guang Cheng. Bootstrapping upper confidence bound. In *Advances in Neural Information Processing Systems*, pages 12123–12133, 2019.
- [13] Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. *arXiv preprint arXiv:1902.10089*, 2019.
- [14] Branislav Kveton, Csaba Szepesvari, Zheng Wen, Mohammad Ghavamzadeh, and Tor Lattimore. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. *arXiv preprint arXiv:1811.05154*, 2018.
- [15] Shyong Lam and Jon Herlocker. MovieLens data sets. *Department of Computer Science and Engineering at the University of Minnesota*, 2006.
- [16] Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. *arXiv preprint arXiv:1610.04491*, 2016.
- [17] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

- [18] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, pages 672–679, 2008.
- [19] Ian Osband and Benjamin Van Roy. Bootstrapped thompson sampling and deep exploration. *arXiv preprint arXiv:1507.00300*, 2015.
- [20] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- [21] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [22] Liang Tang, Yexi Jiang, Lei Li, Chunqiu Zeng, and Tao Li. Personalized recommendation via parameter-free contextual bandits. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 323–332, 2015.
- [23] Sharan Vaswani, Branislav Kveton, Zheng Wen, Anup Rao, Mark Schmidt, and Yasin Abbasi-Yadkori. New insights into bootstrapping for bandits. *arXiv preprint arXiv:1805.09793*, 2018.

## Appendix A

This section contains the proof of Lemma 1.

*Proof.* Suppose  $S_{i,t} < 0$ , that is

$$\beta(\|\mathbf{x}_{i_*}\|_{\mathbf{V}_t^{-1}} + \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}) < \hat{\mu}_{i_*,t} - \hat{\mu}_{i,t} \quad (17)$$

Rearrange terms gives

$$\hat{\mu}_{i,t} + \beta\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \leq \hat{\mu}_{i_*,t} - \beta\|\mathbf{x}_{i_*}\|_{\mathbf{V}_t^{-1}} \quad (18)$$

Note that  $|\mu_i - \hat{\mu}_{i,t}| \leq \beta\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}$ ,  $\forall i \in \mathcal{A}$ . Then,

$$\hat{\mu}_{i_*,t} - \beta\|\mathbf{x}_{i_*}\|_{\mathbf{V}_t^{-1}} \leq \mu_{i_*} \quad (19)$$

and

$$\mu_i \leq \hat{\mu}_{i,t} + \beta\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \quad (20)$$

Combine together we have

$$\mu_i \leq \mu_{i_*} \leq \mu_* \quad (21)$$

Recall by definition  $i_* = \arg \max_{i \in \mathcal{A}} \hat{\mu}_{i,t} - \beta\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}$  is the arm with largest lower upper bound at round  $t$ . Therefore,  $\Delta_i = \mu_* - \mu_i > 0$ . In words, arm  $i$  is suboptimal.

Suppose  $S_{j,t} \geq S_{i,t} \geq 0$

$$\beta(\|\mathbf{x}_{j_*}\|_{\mathbf{V}_t^{-1}} + \|\mathbf{x}_j\|_{\mathbf{V}_t^{-1}}) - (\hat{\mu}_{j_*,t} - \hat{\mu}_{j,t}) \leq \beta(\|\mathbf{x}_{i_*}\|_{\mathbf{V}_t^{-1}} + \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}) - (\hat{\mu}_{i_*,t} - \hat{\mu}_{i,t}) \quad (22)$$

Recall the definition of  $i_*$ ,

$$i_* = \arg \max_{j \in [K]} \hat{\mu}_{j,t} - \beta\|\mathbf{x}_j\|_{\mathbf{V}_t^{-1}} \quad (23)$$

Thus, at each time  $t$ ,  $i_* = j_*$ . Then,

$$\beta\|\mathbf{x}_j\|_{\mathbf{V}_t^{-1}} + \hat{\mu}_{j,t} \leq \beta\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} + \hat{\mu}_{i,t} \quad (24)$$

□

## Appendix B

This section contains the proof of Lemma 2.

*Proof.*

$$p_{\mathcal{U}_t} = \frac{\sum_{i \in \mathcal{U}_t} \exp(\gamma_t S_{i,t})}{\sum_{i \in \mathcal{U}_t} \exp(\gamma_t S_{i,t}) + \sum_{j \in \mathcal{L}_t} \exp(\gamma_t S_{j,t})} \quad (25)$$

By definition,  $S_{j,t} < 0$ ,  $\forall j \in \mathcal{L}$ . Thus,

$$\exp(\gamma S_{j,t}) < 1, \forall j \in \mathcal{L} \quad (26)$$

Then,

$$\sum_{j \in \mathcal{L}_t} \exp(\gamma S_{j,t}) < |\mathcal{L}_t| \quad (27)$$

Therefore,

$$p_{\mathcal{U}_t} > \frac{\sum_{i \in \mathcal{U}_t} \exp(\gamma_t S_{i,t})}{\sum_{i \in \mathcal{U}_t} \exp(\gamma_t S_{i,t}) + |\mathcal{L}_t|} \quad (28)$$

For any probability  $\delta \in (0, 1)$ , we can find a  $\gamma_t$  such that  $p_{\mathcal{U}_t} \geq \delta$ , namely

$$\frac{\sum_{i \in \mathcal{U}_t} \exp(\gamma_t S_{i,t})}{\sum_{i \in \mathcal{U}_t} \exp(\gamma_t S_{i,t}) + |\mathcal{L}_t|} \geq \delta \quad (29)$$

Rearrange terms gives

$$\sum_{i \in \mathcal{U}_t} \exp(\gamma_t S_{i,t}) \geq \frac{\delta |\mathcal{L}_t|}{1 - \delta} \quad (30)$$

Take logarithm on both sides,

$$\log \left( \sum_{i \in \mathcal{U}_t} \exp(\gamma_t S_{i,t}) \right) \geq \log \left( \frac{\delta |\mathcal{L}_t|}{1 - \delta} \right) \quad (31)$$

The left side term is LogSumExp which can be approximated by

$$\log \left( \sum_{i \in \mathcal{U}_t} \exp(\gamma_t S_{i,t}) \right) \geq \max_{i \in \mathcal{U}_t} \gamma_t S_{i,t} = \gamma_t \max_{i \in \mathcal{U}_t} S_{i,t} \quad (32)$$

Denote  $\tilde{S}_{\max,t} = \max_{i \in \mathcal{U}_t} S_{i,t}$  and let

$$\gamma_t \tilde{S}_{\max,t} \geq \log \left( \frac{\delta |\mathcal{L}_t|}{1 - \delta} \right) \quad (33)$$

we have

$$\gamma_t \geq \frac{\log \left( \frac{\delta |\mathcal{L}_t|}{1 - \delta} \right)}{\tilde{S}_{\max,t}} \quad (34)$$

Therefore, if  $\gamma_t$  satisfies Eq. 34,

$$p_{\mathcal{U}_t} \geq \delta \quad (35)$$

Clearly,  $p_{\mathcal{L}_t} < 1 - \delta$  since  $p_{\mathcal{L}_t} + p_{\mathcal{U}_t} = 1$ .  $\square$

## Appendix C

This section contains the derive of gradients.

*Proof.*

$$\begin{aligned} \max_{\beta} Y(T) &= \max_{\beta} \sum_{t=1}^T \mathbb{E}[y_t] = \max_{\beta, \gamma} \sum_{t=1}^T \sum_{i=1}^K p_{i,t} \mu_i \\ \text{s.t. } &|\mu_i - \hat{\mu}_{i,t}| - \beta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \leq 0, \quad \forall i \in \mathcal{A}, \quad \forall t \in [T] \end{aligned} \quad (36)$$

Apply the Lagrange multipliers, the optimization objective is

$$\max_{\beta} \sum_{t=1}^T \sum_{i=1}^K p_{i,t} \mu_i - \eta (|\mu_i - \hat{\mu}_{i,t}| - \beta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}) \quad \text{s.t. } \eta > 0 \quad (37)$$

Apply the score function  $\nabla_{\theta} f(\theta) = f(\theta) \nabla_{\theta} \log f(\theta)$  to  $p_{i,t}$

$$\log p_{i,t} = \gamma S_{i,t} - \log \sum_{j=1}^K \exp \gamma S_{j,t} \quad (38)$$

$$\nabla_{\beta} \log p_{i,t} = \gamma_t \phi_{i,t} - \frac{\sum_{j=1}^K \gamma_t \phi_{j,t} \exp \gamma_t S_{j,t}}{\sum_{j=1}^K \exp \gamma_t S_{j,t}} \quad (39)$$

Then, the gradient  $g(\beta)$  is

$$g(\beta) = \sum_{t=1}^T \sum_{i=1}^K \mu_i p_{i,t} \left( \gamma_t \phi_{i,t} - \frac{\sum_{j=1}^K \gamma_t \phi_{j,t} \exp \gamma_t S_{j,t}}{\sum_{j=1}^K \exp \gamma_t S_{j,t}} \right) + \eta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \quad (40)$$

The gradient estimator  $\hat{g}(\beta)$  is obtained by repalcing  $\mu_i$  with  $\hat{\mu}_{i,t} = \mathbf{x}_i^T \hat{\boldsymbol{\theta}}_t$  where  $\hat{\boldsymbol{\theta}}_t = \mathbf{V}_t^{-1} \sum_{s=1}^t \mathbf{x}_s y_s$  is obtained via least-square estimator.

$$\hat{g}(\beta) = \sum_{t=1}^T \sum_{i=1}^K \hat{\mu}_{i,t} p_{i,t} \left( \gamma_t \phi_{i,t} - \frac{\sum_{j=1}^K \gamma_t \phi_{j,t} \exp \gamma_t S_{j,t}}{\sum_{j=1}^K \exp \gamma_t S_{j,t}} \right) + \eta \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \quad (41)$$

$\square$

## Appendix D

This section contains the proof of Theorem 1.

*Proof.* The probability of each arm is defined as

$$p_{i,t} = \frac{\exp(\gamma_t S_{i,t})}{\sum_{j=1}^K \exp(\gamma_t S_{j,t})} \quad (42)$$

$S_{i,t}$  is defined as

$$S_{i,t} = \hat{\beta} \phi_{i,t} - \hat{\Delta}_{i,t} = \hat{\beta} (\|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} + \|\mathbf{x}_{i^*}\|_{\mathbf{V}_t^{-1}}) - (\hat{\mu}_{i^*,t} - \hat{\mu}_{i,t}) \quad (43)$$

The cumulative regret to be minimized is defined as

$$\begin{aligned} R_T &= \sum_{t=1}^T \mathbb{E}[r_t] = \sum_{t=1}^T \mu_* - \mathbb{E}[y_t] = \sum_{t=1}^T (\mu_* - \sum_{i=1}^K p_{i,t} \mu_i) \\ &= \sum_{t=1}^T \sum_{i=1}^K p_{i,t} (\mu_* - \mu_i) = \sum_{t=1}^T \sum_{i=1}^K p_{i,t} \Delta_i \end{aligned} \quad (44)$$

where we use  $\sum_{i=1}^K p_{i,t} = 1$ .

At each time  $t$ , arm set  $\mathcal{A}$  is divided into two subsets  $\mathcal{U}_t$  and  $\mathcal{L}_t$  with  $\mathcal{U}_t \cup \mathcal{L}_t = \mathcal{A}$ . Arm  $i \in \mathcal{U}_t$  if  $S_{i,t} \geq 0$  and arm  $i \in \mathcal{L}_t$  if  $S_{i,t} < 0$ .

$$\mathbb{E}[r_t] = \sum_{i=1}^K p_{i,t} \Delta_i = \sum_{i \in \mathcal{U}_t} p_{i,t} \Delta_i + \sum_{i \in \mathcal{L}_t} p_{i,t} \Delta_i \quad (45)$$

Suppose  $\gamma_t$  follows Lemma 2, then  $\sum_{i \in \mathcal{L}_t} p_{i,t} < 1 - \delta$ . Assume  $\Delta_i \leq 1, \forall i \in \mathcal{A}$ . Then,

$$\mathbb{E}[r_t] = \sum_{i \in \mathcal{U}_t} p_{i,t} \Delta_i + \sum_{i \in \mathcal{L}_t} p_{i,t} \leq \sum_{i \in \mathcal{U}_t} p_{i,t} \Delta_i + (1 - \delta) \quad (46)$$

By setting  $\delta \approx 1$ , we have  $1 - \delta \approx 0$ . It means arms in  $\mathcal{L}_t$  are unlikely to be selected. So, the second term can be dropped. Therefore,

$$\mathbb{E}[r_t] \leq \sum_{i \in \mathcal{U}_t} p_{i,t} \Delta_i \quad (47)$$

Thus,

$$\mathbb{E}[r_t] \leq \sum_{i \in \mathcal{U}_t} p_{i,t} \Delta_i = \sum_{i \in \mathcal{U}_t} p_{i,t} (\mu_* - \mu_i) \quad (48)$$

Note that at each time  $t$ ,  $|\hat{\mu}_{i,t} - \mu_i| \leq \hat{\beta} \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}, \forall i \in [K]$ . Then

$$\mu_* \leq \hat{\mu}_{i^*,t} + \hat{\beta} \|\mathbf{x}_{i^*}\|_{\mathbf{V}_t^{-1}} \quad (49)$$

and

$$\mu_i \geq \hat{\mu}_{i,t} - \hat{\beta} \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}} \quad (50)$$

Thus,

$$\mu_* - \mu_i \leq \hat{\beta} (\|\mathbf{x}_{i^*}\|_{\mathbf{V}_t^{-1}} + \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}) + (\hat{\mu}_{i^*,t} - \hat{\mu}_{i,t}) \quad (51)$$

Note that  $\hat{\mu}_{i^*,t} - \hat{\mu}_{i,t} \leq \hat{\mu}_{i^*,t} - \hat{\mu}_{i^*,t} + \hat{\mu}_{i^*,t} - \hat{\mu}_{i,t}$  where  $i^* = \arg \max_{j \in [K]} \hat{\mu}_{j,t} - \hat{\mu}_{i,t}$ . Therefore,

$$\mu_* - \mu_i \leq \hat{\beta} (\|\mathbf{x}_{i^*}\|_{\mathbf{V}_t^{-1}} + \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}) + (\hat{\mu}_{i^*,t} - \hat{\mu}_{i,t}) \quad (52)$$

Since  $i \in \mathcal{U}_t, S_{i,t} \geq 0$ . That is  $\hat{\mu}_{i^*,t} - \hat{\mu}_{i,t} \leq \beta (\|\mathbf{x}_{i^*}\|_{\mathbf{V}_t^{-1}} + \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}})$ . Then,

$$\begin{aligned} \mu_* - \mu_i &\leq \hat{\beta} (\|\mathbf{x}_{i^*}\|_{\mathbf{V}_t^{-1}} + \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}) + (\hat{\mu}_{i^*,t} - \hat{\mu}_{i,t}) \\ &\leq 2\hat{\beta} (\|\mathbf{x}_{i^*}\|_{\mathbf{V}_t^{-1}} + \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}) \end{aligned} \quad (53)$$

Define  $\psi_t = \max_{i \in [K]} \|\mathbf{x}_i\|_{\mathbf{V}_t^{-1}}$ . We have

$$\mu_* - \mu_i \leq 4\hat{\beta}\psi_t \quad (54)$$

Plugging this into Eq. 48 gives

$$\mathbb{E}[r_t] \leq 4\hat{\beta} \sum_{i \in \mathcal{U}_t} p_{i,t} \psi_t \quad (55)$$

Since we assume  $\gamma_t$  follows Lemma 2, we have  $p_{\mathcal{U}_t} = \sum_{i \in \mathcal{U}_t} p_{i,t} = \delta$ . Therefore,

$$\mathbb{E}[r_t] \leq 4\hat{\beta} \sum_{i \in \mathcal{U}_t} p_{i,t} \psi_t = 4\hat{\beta} \phi_t \sum_{i \in \mathcal{U}_t} p_{i,t} = 4\hat{\beta} \psi_t p_{\mathcal{U}_t} \leq 4\hat{\beta} \delta \psi_t \quad (56)$$

Thus, the cumulative regret

$$R_T = \sum_{t=1}^T \mathbb{E}[r_t] \leq \sqrt{T \sum_{t=1}^T \mathbb{E}[r_t]^2} \leq 4\hat{\beta} \delta \sqrt{T \sum_{t=1}^T \psi_t^2} \quad (57)$$

From Lemma 3 (stated below), we have

$$\sum_{t=1}^T \psi_t^2 \leq 2d \log\left(\alpha + \frac{T}{d}\right) \quad (58)$$

Plugging in Eq. 57,

$$R_T \leq 4\hat{\beta} \delta \sqrt{2Td \log\left(\alpha + \frac{T}{d}\right)} = \tilde{O}\left(\hat{\beta} \sqrt{Td \log\left(1 + \frac{T}{d}\right)}\right) \quad (59)$$

where  $\delta$  is the probability parameter chosen by user.

**Lemma 3.** (Lemma 11 in [1])

$$\sum_{t=1}^T \|\mathbf{x}\|_{\mathbf{V}_t^{-1}}^2 \leq \log \det(\mathbf{V}_t) \leq 2d \log\left(\alpha + \frac{T}{d}\right) \quad (60)$$

□

## Appendix E

This section contains the pseudo code of **SoftUCB**, **SoftUCB offline** and **SoftUCB online**.

---

### Algorithm 2: **SoftUCB**

---

**Input** :  $\beta, \mathcal{A}, K, T, \alpha$ .

**Initialization** :  $\mathbf{V}_0 = \alpha \mathbf{I} \in \mathbb{R}^{d \times d}$ ,  $\mathbf{b}_0 = \mathbf{0} \in \mathbb{R}^d$ ,  $\hat{\boldsymbol{\theta}}_0 = \mathbf{0} \in \mathbb{R}^d$ ,  $\gamma_0 = 0$ .

**for**  $t \in [1, T]$  **do**

1. Find  $S_{i,t}, \forall i \in \mathcal{A}$  via Eq. 7 with  $\beta$ .
2. Find  $\boldsymbol{\pi}_t$  via Eq. 8 with  $\gamma_{t-1}$ .
3. Select arm  $i_t \in \mathcal{A}$  randomly following  $\boldsymbol{\pi}_t$  and receive payoff  $y_t$ .
4. Update  $\mathbf{V}_t \leftarrow \mathbf{V}_t + \mathbf{x}_t \mathbf{x}_t^T$ ,  $\mathbf{b}_t \leftarrow \mathbf{b}_{t-1} + \mathbf{x}_t y_t$  and  $\hat{\boldsymbol{\theta}}_t = \mathbf{V}_t^{-1} \mathbf{b}_t$ .
5. Update  $\gamma_t$  via Eq. 9.

**end**

---

### Algorithm 3: **SoftUCB offline**

---

**Input** :  $\mathcal{A}, K, T, \lambda, \eta$

**Initialization** :  $\beta_0 = 0, \hat{\beta} = 0$ .

**for**  $n \in [1, N]$  **do**

1. Run **SoftUCB** on  $\mathcal{A}$  rounds with  $\beta = \beta_{n-1}$ .
2. Update  $\beta_n \leftarrow \beta_{n-1} + \lambda \hat{g}(\beta)$  via Eq. 13

**end**

**Output** :  $\hat{\beta} \leftarrow \beta_N$

Run **SoftUCB** on  $\mathcal{A}$  with  $\beta = \hat{\beta}$ .

---

### Algorithm 4: **SoftUCB online**

---

**Input** :  $\mathcal{A}, K, T, \alpha, \lambda, \eta$

**Initialization** :  $\beta_0 = 0, \mathbf{V}_0 = \alpha \mathbf{I} \in \mathbb{R}^{d \times d}$ ,  $\mathbf{b}_0 = \mathbf{0} \in \mathbb{R}^d$ ,  $\hat{\boldsymbol{\theta}}_0 = \mathbf{0} \in \mathbb{R}^d$ ,  $\gamma_0 = 0$ .

**for**  $t \in [1, T]$  **do**

1. Select arm  $i_t \in [K]$  randomly following  $\boldsymbol{\pi}_t$  and receive payoff  $y_t$ .
2. Update  $\mathbf{V}_t \leftarrow \mathbf{V}_t + \mathbf{x}_t \mathbf{x}_t^T$ ,  $\mathbf{b}_t \leftarrow \mathbf{b}_{t-1} + \mathbf{x}_t y_t$  and  $\hat{\boldsymbol{\theta}}_t = \mathbf{V}_t^{-1} \mathbf{b}_t$ .
3. Update  $\beta_t \leftarrow \beta_{t-1} + \lambda \hat{g}_t(\beta)$  via Eq. 15.

**end**

---

## Appendix F

This section contains the learning curves of **SoftUCB offline**.

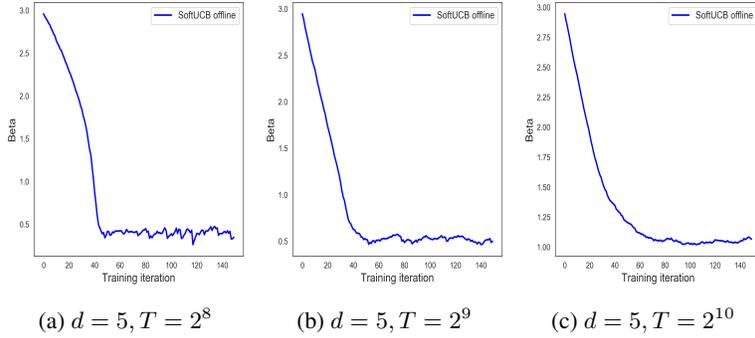


Figure 3: Learning curves of **SoftUCB offline**

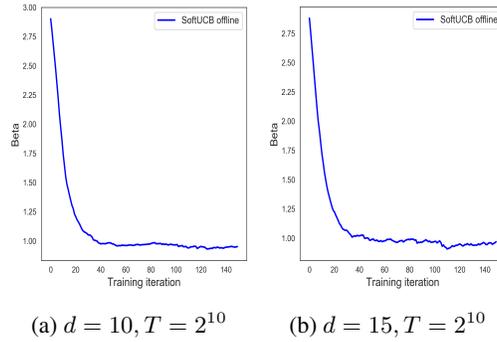


Figure 4: Learning curves of **SoftUCB offline**

## Appendix G

The dataset **Jester** contains ratings of 40 jokes from 19891 users. We sample  $K = 50$  users randomly as arms. Their rating to top 39 jokes are used as feature vector. Then, to reduce the sparsity, we apply principle component analysis algorithm to reduce the dimension  $d = 10$ . Their rating on the 40th jokes are used as rewards. At each round, the algorithm selects on user to recommend the joke and the reward is the rating given by the user. **MovieLens** contains 6k users and their ratings on 40k movies. Since not every user gives ratings on all movies, there are a large amount of missing ratings. We factorize the rating matrix to fill the missing values. The rest works the same as in **Jester**.