Combinatorial Allocation Bandits with Nonlinear Arm Utility

Yuki Shibukawa
The University of Tokyo and RIKEN AIP
shibu-yu762@g.ecc.u-tokyo.ac.jp

Koichi Tanaka
Keio University
kouichi_1207@keio.jp

Yuta Saito
Hanuku-kaso, Co., Ltd.
saito@hanjuku-kaso.com

Shinji Ito
The University of Tokyo and RIKEN AIP
shinji@mist.i.u-tokyo.ac.jp

Abstract

A matching platform is a system that matches participants of different types, such as companies and job-seekers. In such a platform, maximizing matches may concentrate assignments on popular participants, increasing dissatisfaction among others, and eventually causing churn, which reduces the platform's profit opportunities. To address this issue, we propose a novel online learning problem, Combinatorial Allocation Bandits (CAB), which incorporates the notion of *arm satisfaction*. In CAB, at each round, the learner observes feature vectors for K arms and N users, assigns users to arms, and observes feedback following a generalized linear model (GLM). Unlike prior work, the objective is to maximize arm satisfaction rather than the number of positive feedback. For CAB, we develop an upper confidence bound algorithm that uses an approximate optimization oracle and achieves an approximate regret upper bound, whose dependence on d, T, and N matches the known lower bound for contextual combinatorial linear bandits up to logarithmic factors. We also analyze a Thompson sampling algorithm with a standard regret bound under an exact optimization oracle, and propose a cheaper one-pass variant retaining sublinear approximate regret under a self-concordance assumption. Experiments on synthetic data support the objective and show that CAB-UCB achieves higher cumulative satisfaction than baselines.

1 Introduction

Online learning is a framework in which decisions are made sequentially based on observed information. It has a wide range of potential applications, such as recommender systems, and has been studied extensively from a theoretical perspective [6, 11, 30].

Although these studies make important theoretical contributions, they mainly focus on maximizing the number of positive feedback, such as matches or clicks, which sometimes may not reflect real-world business objectives. For example, under unconstrained settings, a match-maximizing algorithm often yields an imbalanced selection of arms, leading to dissatisfaction among infrequently selected arms. In job-matching platforms that recommend companies to visiting users, the revenue model typically relies on fees paid by the companies participating in the platform to hire qualified applicants. Thus, the economic cost of company churn can outweigh raw match counts. Similar structures arise in dating apps and paper review processes. Dating apps match users with one another. When matches concentrate among a few popular users, many others receive few or no matches, which in turn reduces their incentives to remain active. This decline in active participation is undesirable for the platform.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

Similarly, paper review processes can be regarded as a match between authors and reviewers. If authors are not sufficiently satisfied with the quality of the reviews, they may submit to other journals, resulting in a loss of future submissions.

The key point in the above discussions is that companies whose satisfaction falls below a certain level are expected to have a higher probability of leaving the platform. Accordingly, the platform objective should not be to maximize raw match counts alone, but to avoid allocations in which matches are concentrated among a small subset of companies. Motivated by economic principles and by the cost of interviewing many applicants[1], we model each arm's satisfaction as a concave function of its expected number of matches. Here, satisfaction represents an arm-side evaluation of the platform determined by the quality of its allocated users. The concavity captures diminishing marginal utility and mitigates concentration without imposing explicit fairness constraints.

As illustrated in Figure 1, when one arm is substantially easier to match than the others, a match-maximizing policy concentrates all recommendations on that arm. Such concentration is undesirable because arm satisfaction typically exhibits diminishing returns and therefore does not scale linearly with the number of matches. Moreover, even for arm A, which receives many assignments, real-world constraints such as budget limits and capacity restrictions, as well as the economic principle of diminishing marginal utility, imply that the satisfaction obtained does not necessarily grow proportionally with the reward.

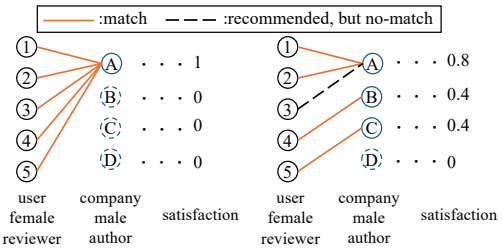


Figure 1: This figure schematically compares satisfaction-based outcomes with match-maximization outcomes. Firm popularity is assumed to decrease from Firm A to Firm D.

One way to model such limitations is to impose explicit resource constraints on arm usage. This view is closely related to bandits with knapsacks (BwK), which incorporate resource constraints into online learning [8, 4, 39]. BwK is natural when the goal is to limit arm usage via explicit budgets or capacities, but BwK does not directly capture aspects of our interest, such as arm satisfaction or the diminishing property of the utility function.

Another approach imposes fairness constraints on each arm's selection frequency [18, 29, 48]. These approaches focus on fairness in exposure or selection, rather than directly modeling arm-side utility. Consequently, they do not necessarily capture the satisfaction objective, which depends not only on how often an arm is selected but also on the quality of the assigned users and the arms' popularity.

1.1 Our contributions

Our main contributions are primarily theoretical and are twofold. First, we formulate Combinatorial Allocation Bandits (CAB), a contextual combinatorial semi-bandit problem with GLM feedback and nonlinear arm-side utility. Beyond the linear case, the available combinatorial GLM result is limited to a logistic model with binary feedback [33]. In contrast, CAB is formulated for GLM feedback with a non-negative mean and can handle non-binary feedback distributions. On the objective side, CAB uses a concave arm-side utility that aggregates the expected feedback of the users assigned to the same arm, rather than a linear or user-wise separable objective. This structure captures diminishing returns and makes the per-round optimization problem NP-hard (Theorem C.1).

Second, we develop two learning algorithms for CAB: CAB-UCB (Section 4.1), based on the upper confidence bound (UCB) principle, and CAB-TS (Section 4.2), based on Thompson sampling (TS), under different offline optimization assumptions. For CAB-UCB, we combine GLM confidence bounds with an approximate optimization oracle and prove an approximate regret bound of $\widetilde{O}(d\sqrt{NT}+dN)$ (Theorem 4.1), whose dependence on d, T, and N matches the known lower bound for contextual combinatorial linear bandits up to logarithmic factors. We also show how to implement the CAB-UCB oracle efficiently via a reduction to the submodular welfare problem (Section 4.1.2). For CAB-TS, we assume access to an exact optimization oracle, introduce a user-wise i.i.d. sampling scheme in which the sampled perturbations enter the objective function linearly as user-wise additive perturbations, and prove a standard regret bound of $\widetilde{O}(dN\sqrt{T}+dN^{3/2})$ (Theorem 4.2). Notably,

[1] Similar costs arise in dating apps through going on dates and in paper review processes through responding to reviews.

the regret analyses of CAB-UCB and CAB-TS do not require a self-concordance assumption on the link function, which is used in recent analyses [33, 49].

As a computational extension, we also present a one-pass variant (Section 4.3). At each round t, CAB-UCB and CAB-TS solve a regularized MLE, incurring $O(t)$ cost. To reduce this computational cost, we replace the MLE update with a one-pass parameter update based on online mirror descent (OMD), following [49]. Under an additional self-concordance assumption on the link function, the resulting algorithm avoids solving a regularized MLE while retaining sublinear approximate regret.

Finally, we conduct experiments on synthetic data that support the proposed objective and show that CAB-UCB achieves higher cumulative satisfaction than match-oriented and fairness-oriented baselines (Section 5).

For space constraints, we provide a discussion of the related work in Appendix B.

1.2 Technical challenges

A technical difficulty specific to CAB is that the per-round objective is no longer separable across users. Unlike standard contextual combinatorial semi-bandits with linear or arm-wise additive rewards, the benefit of assigning a user to an arm depends on whether other users are assigned to the same arm, because the arm-side utility exhibits diminishing returns. As a result, analyses for separable reward objectives do not directly apply, and even the offline allocation problem becomes NP-hard. This structural difference makes the offline allocation step nontrivial and requires oracle assumptions tailored to each algorithm.

This non-separability further complicates uncertainty assessment in the GLM setting. For CAB-UCB, our key observation is that, although the objective itself is non-separable, its estimation error can still be upper-bounded by a bonus that decomposes as a sum of user-wise terms defined in (3). This form is essential for retaining the submodular welfare structure, which allows CAB-UCB to be analyzed using an implementable approximate optimization oracle. Despite its simple form, the bonus term yields a regret bound with the desired dependence on the number of rounds and users.

CAB-TS is more delicate for several reasons. First, a common Gaussian perturbation is insufficient in CAB. With a common perturbation, user-level feature vectors are summed before their size is measured, so their directions can cancel, and we cannot derive a suitable bound on the variance of the aggregate perturbation (see Remark D.10). CAB-TS addresses this issue by using user-wise independent Gaussian perturbations, which preserve enough variance in the aggregate perturbation of the optimal allocation to support the regret analysis. Second, directly perturbing the nonlinear utility is difficult because the perturbation enters a coupled concave objective. We therefore add exploration through a separate user-wise linear perturbation. We choose the covariance surrogate by placing the regularization term inside the Hessian-weighted sum, keeping it comparable to the GLM information matrix. This design avoids perturbing the nonlinear utility directly (see Remark D.11 for details).

2 Preliminaries

In this section, we describe the setting of the GLM and the submodular welfare problem used in the implementation of our algorithms.

Notations For $n \in \mathbb{N}$, let $[n] = \{1, \dots, n\}$. For $\boldsymbol{x} \in \mathbb{R}^d$, denote the transpose by $\boldsymbol{x}^\top$. For a positive definite matrix $\boldsymbol{A}$, define $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}$, and let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote its minimum and maximum eigenvalues, respectively. Let $\mathbf{0}$ be the all-zero vector. A set function $f\colon 2^X \to \mathbb{R}$ is called *monotone* if $f(S) \le f(T)$ whenever $S \subseteq T \subseteq X$. It is called *submodular* if $f(S \cup T) + f(S \cap T) \le f(S) + f(T)$ for any $S, T \subseteq X$. We denote the Lipschitz constant of the function g by L_g. Let $\Pi = \{\pi\colon [N] \to [K]\}$.

2.1 Generalized linear models

Within the framework of GLM [36], the conditional distribution of the response variable Y given the explanatory variable X belongs to the exponential family. Formally, the probability density function parameterized by $\boldsymbol{\theta}$ is given by

$$\mathbb{P}(Y|\boldsymbol{\theta}; \boldsymbol{X}) \propto \exp\{Y \boldsymbol{X}^\top \boldsymbol{\theta} - m(\boldsymbol{X}^\top \boldsymbol{\theta})\}, \tag{1}$$

Problem 1 Combinatorial Allocation Bandits (CAB)

for $t = 1, \dots, T$ **do**
  Context: Observe $\{\boldsymbol{\phi}_t(i,a)\}_{a\in[K]}$ for each $i \in [N]$.
  Action: Choose allocation $\pi_t \in \Pi$.
  Feedback: Observe $y_t(i) \sim \mathbb{P}(\cdot|\boldsymbol{\theta}^*; \boldsymbol{\phi}_t(i, \pi_t(i)))$ for each $i \in [N]$.
  Reward (unobserved): Receive $f_t(\pi_t; \boldsymbol{\theta}^*) = \sum_{a\in[K]} r(\sum_{i\in\pi_t^{-1}(a)} \mu(\boldsymbol{\phi}_t(i,a)^\top \boldsymbol{\theta}^*))$.

where we use the unit-dispersion canonical form, since any fixed known dispersion parameter can be absorbed into the sub-Gaussian parameter and tuning constants without changing the regret rates. Under this normalization, $m\colon \mathbb{R} \to \mathbb{R}$ is a known twice differentiable function and satisfies $\dot{m}(\boldsymbol{X}^\top\boldsymbol{\theta}) = \mathbb{E}[Y|\boldsymbol{X}]$ and $\ddot{m}(\boldsymbol{X}^\top\boldsymbol{\theta}) = \mathrm{Var}(Y|\boldsymbol{X})$. In what follows, we set $\mu(\boldsymbol{X}^\top\boldsymbol{\theta}) = \dot{m}(\boldsymbol{X}^\top\boldsymbol{\theta})$. The exponential family comprises distributions such as the Gaussian and Bernoulli.

In this setting, given independent samples $Y_1, \dots, Y_n$ conditional on $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, we denote the dataset by $\mathcal{D} = \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$. Then, the negative log-likelihood function is given by $\mathcal{L}(\mathcal{D}; \boldsymbol{\theta}) = \sum_{i=1}^n [m(\boldsymbol{X}_i^\top\boldsymbol{\theta}) - Y_i\boldsymbol{X}_i^\top\boldsymbol{\theta}] + \text{constant}$. Then, since m is differentiable, the minimum likelihood estimator (MLE) is given by the minimizer of $\sum_{i=1}^n [m(\boldsymbol{X}_i^\top\boldsymbol{\theta}) - Y_i\boldsymbol{X}_i^\top\boldsymbol{\theta}]$ (see [15, 31] for details).

However, in our problem, using the MLE requires an initial exploration. To avoid this issue, we employ a regularized MLE with ridge regularization. Here, the regularized negative log-likelihood corresponding to the regularized MLE takes the form $\widetilde{\mathcal{L}}(\mathcal{D}; \boldsymbol{\theta}, \lambda) = \mathcal{L}(\mathcal{D}; \boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$.

2.2 Submodular welfare problem

The submodular welfare problem was first studied by Lehmann et al. [28] and is defined as follows: *the submodular welfare problem, given m items and n players with submodular utility functions $w_i\colon 2^{[m]} \to \mathbb{R}_{\geq 0}$, is the problem of maximizing $\sum_{i=1}^n w_i(S_i)$, where $S_1, \dots, S_n$ are disjoint subsets of the item set.* A limitation for the submodular welfare problem is that no approximation better than $1 - 1/e$ can be achieved unless $P = NP$ [21]. There are two commonly considered oracle models, the value oracle model and the demand oracle model. The former model returns the value of utility $w_i(S)$, and the latter model returns the set S which maximizes $w_i(S) - \sum_{j\in S} p_j$ given an assignment of prices to items $p\colon [m] \to \mathbb{R}$. We call an algorithm ε-approximate algorithm if the value obtained by the algorithm (we denote it ALG) satisfies the inequality $\varepsilon\mathrm{OPT}_{sub} \leq \mathrm{ALG}$, where OPT_{sub} denotes the optimal value. Under the value oracle model, there exists an approximate algorithm that achieves the following approximation:

Lemma 2.1 (45, Section 5)**.** *The submodular welfare problem admits a $1 - 1/e$-approximation in the value-oracle model when the utility functions are monotone submodular.*

3 Combinatorial allocation bandits

This section introduces a setting of Combinatorial Allocation Bandits (CAB) and our intention for constructing the problem.

3.1 Problem setting

We introduce CAB (Problem 1), a novel online learning problem. At round t, for each user $i \in [N]$, the learner obtains a context set $\{\boldsymbol{\phi}_t(i,a)\}_{a\in[K]}$ with $\|\boldsymbol{\phi}_t(i,a)\| \leq 1$, where the contexts are chosen by an oblivious adversary before the learning process begins. Note that i is not associated with a particular user, but instead denotes the index reflecting the order of observation. Given the observations, the learner determines the allocation $\pi_t \in \Pi$ at round t. Subsequently, based on π_t, the learner observes the feedback $y_t(i) = y_t(i, \pi_t(i))$ for each $i \in [N]$. Let $\mathcal{F}_t = \sigma(y_1(1), \dots, y_1(N), \dots, y_t(1), \dots, y_t(N), \pi_1, \dots, \pi_t)$, where $\sigma(\mathcal{A})$ is the smallest σ-algebra containing $\mathcal{A}$. Then, $(\mathcal{F}_t)_t$ is a filtration. Denote the conditional probability and expectation given the history of observations by $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid \mathcal{F}_{t-1})$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$.

We consider that the feedback $y_t(i)$ observed by the learner follows a GLM with an unknown parameter $\boldsymbol{\theta}^* \in \mathbb{R}^d$ (Section 2.1). We assume that $\{y_t(i)\}_{i\in[N]}$ are conditionally independent given $\{\boldsymbol{\phi}_t(i, \pi_t(i))\}_{i\in[N]}$, and that $\|\boldsymbol{\theta}^*\|_2 \leq D$ for some constant $D > 0$. Let $\Theta = \left\{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \leq D\right\}$.

Accordingly, the probability density function (or the probability mass function) of $y_t(i)$ can be expressed using (1) as $\mathbb{P}(y_t(i)|\boldsymbol{\theta}^*; \boldsymbol{\phi}_t(i, \pi_t(i)))$, whose mean is given by $\mu(\boldsymbol{\phi}_t(i, \pi_t(i))^\top \boldsymbol{\theta}^*)$. In the target applications, the observed feedback $y_t(i)$ is used to assess outcomes such as matching success or other positive feedback. Accordingly, its conditional mean $\mu(\boldsymbol{\phi}_t(i, a)^\top \boldsymbol{\theta})$ represents a non-negative latent quantity. In this problem, we assume that the deviation between the observation and its mean, $y_t(i) - \mu(\boldsymbol{\phi}_t(i, \pi_t(i))^\top \boldsymbol{\theta}^*)$, is sub-Gaussian with parameter $\sigma > 0$. That is, for any $t \in [T]$, $i \in [N]$, and $\xi \in \mathbb{R}$ it holds that $\mathbb{E}_t[e^{\xi(y_t(i) - \mu(\boldsymbol{\phi}_t(i, \pi_t(i))^\top \boldsymbol{\theta}^*))}] \leq e^{\xi^2 \sigma^2 / 2}$. Furthermore, motivated by prior GLM bandit analyses [31, 20, 33], we impose the following assumption on μ.

Assumption 3.1. Let $\mathcal{B} = \{(\boldsymbol{x}, \boldsymbol{\theta}) : \|\boldsymbol{x}\|_2 \leq 1,\ \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq \max\{D + 1, 2D\}\}$. We assume that μ is first-order differentiable and Lipschitz continuous, that $\mu(\boldsymbol{x}^\top \boldsymbol{\theta}) \geq 0$ for all $(\boldsymbol{x}, \boldsymbol{\theta}) \in \mathcal{B}$, and that there exists a known constant $\kappa_\mu > 0$ such that $\kappa_\mu \leq \inf_{(\boldsymbol{x}, \boldsymbol{\theta}) \in \mathcal{B}} \dot{\mu}(\boldsymbol{x}^\top \boldsymbol{\theta})$.

At the end of each round, the allocation π_t induces arm-side satisfaction for each arm. Specifically, we consider the satisfaction $r(\sum_{i \in \pi_t^{-1}(a)} \mu(\boldsymbol{\phi}_t(i, a)^\top \boldsymbol{\theta}^*))$, where $r \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a known concave and monotone increasing function bounded by M. r is a platform-specified model of arm-side satisfaction, rather than an additional unknown function to be learned online. The uncertainty lies in the unknown parameter $\boldsymbol{\theta}^*$. We assume that r is Lipschitz continuous. For convenience, we define

$$f_t(\pi; \boldsymbol{\theta}) = \sum_{a \in [K]} r\Big(\sum_{i \in \pi^{-1}(a)} \mu(\boldsymbol{\phi}_t(i, a)^\top \boldsymbol{\theta}) \Big). \tag{2}$$

In CAB, the goal of the learner is to maximize the cumulative satisfaction $\sum_{t=1}^T f_t(\pi_t; \boldsymbol{\theta}^*)$. We measure its performance by the standard regret, the cumulative gap to the per-round optimum, $\mathcal{R}_T = \sum_{t=1}^T (f_t(\pi_t^*; \boldsymbol{\theta}^*) - f_t(\pi_t; \boldsymbol{\theta}^*))$, where $\pi_t^* = \arg\max_{\pi \in \Pi} f_t(\pi; \boldsymbol{\theta}^*)$ is the optimal allocation at round t. However, in general, maximizing f_t is NP-hard even when $\boldsymbol{\theta}^*$ is known (Theorem C.1). Thus, some algorithms below use an α-approximate optimization oracle ($0 < \alpha \leq 1$). For such algorithms, we also use the α-approximate regret $\mathcal{R}_T^\alpha = \sum_{t=1}^T (\alpha f_t(\pi_t^*; \boldsymbol{\theta}^*) - f_t(\pi_t; \boldsymbol{\theta}^*))$.

Next, we explain the modeling intuition behind the above formulation. In many applications, the observed feedback $y_t(i)$ is a realization of the underlying match quality between user i and arm a. For example, in a job-matching platform, $y_t(i) \in \{0, 1\}$ indicates whether a match occurs, and $\mu(\boldsymbol{\phi}_t(i, a)^\top \boldsymbol{\theta}^*)$ can be interpreted as the match probability.

The observed feedback $y_t(i)$ depends not only on the match quality between user i and arm a, but also on random user-side factors after the assignment, such as the user's acceptance decision or availability constraints. For this reason, we do not model an arm's satisfaction using the realized outcomes $y_t(i)$. Instead, we use the latent match probability $\mu(\boldsymbol{\phi}_t(i, a)^\top \boldsymbol{\theta}^*)$ between user i and arm a as the basic quantity for evaluating the users assigned to each arm. Thus, for each arm a, we use the total expected number of matches $\sum_{i \in \pi_t^{-1}(a)} \mu(\boldsymbol{\phi}_t(i, a)^\top \boldsymbol{\theta}^*)$, as the input to the satisfaction function r. The arm-side satisfaction in round t is therefore modeled as $r(\sum_{i \in \pi_t^{-1}(a)} \mu(\boldsymbol{\phi}_t(i, a)^\top \boldsymbol{\theta}^*))$.

We consider a concave and nondecreasing satisfaction function r. Monotonicity means that a greater expected number of successful matches should not reduce the arm's utility. Concavity captures diminishing marginal value under constraints such as interview capacity, screening costs, or budget constraints [37, 34]. Thus, the objective discourages excessive concentration on a small number of arms and can induce more balanced allocations without explicit fairness constraints.

4 Algorithm and theoretical results

This section presents UCB and TS algorithms for CAB and their regret analyses.

4.1 Upper confidence bound algorithm

Our proposed method, CAB-UCB, follows the UCB principle, a standard approach in bandit algorithm design [25, 7, 38, 31]. CAB-UCB has two parameters, $\lambda_0 > 0$ and $c_1 > 0$. The parameter λ_0 determines the regularization strength for MLE. In addition, λ_0 plays the role of ensuring that $\lambda_{\min}(\boldsymbol{V}_t)$ is strictly positive. The parameter c_1 controls exploration, and a larger value results in a greater degree of exploration.

In each round, we compute $\overline{\boldsymbol{\theta}}_t$, the regularized MLE of $\boldsymbol{\theta}^*$, using the set of observations $\mathcal{D}_t = \{(\boldsymbol{x}_s(i), y_s(i))\}_{i \in [N], s < t}$, where $\boldsymbol{x}_s(i) = \boldsymbol{\phi}_s(i, \pi_s(i))$. In calculating π_t, we balance exploitation

Algorithm 1 CAB-UCB

Input: The total rounds T, the number of users N, tuning parameter λ_0 and c_1, and access to an α-approximate optimization oracle.

$\mathcal{D}_1 \leftarrow \varnothing$ and $\boldsymbol{V}_1 \leftarrow \lambda_0 \boldsymbol{I}$.
for $t = 1, \ldots, T$ **do**
  $\overline{\boldsymbol{\theta}}_t \leftarrow \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \widetilde{\mathcal{L}}(\mathcal{D}_t; \boldsymbol{\theta}, \kappa_\mu \lambda_0)$.
  Call an α-approximate optimization oracle for $f_t(\pi; \overline{\boldsymbol{\theta}}_t) + g_t(\pi)$ and let π_t denote its output.
  Observe $y_t(i)$ for any $i \in [N]$.
  $\boldsymbol{x}_t(i) \leftarrow \boldsymbol{\phi}_t(i, \pi_t(i))$ for any $i \in [N]$ and $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(\boldsymbol{x}_t(i), y_t(i))\}_{i \in [N]}$.
  $\boldsymbol{V}_{t+1} \leftarrow \lambda_0 \boldsymbol{I} + \sum_{s=1}^{t} \sum_{i=1}^{N} \boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^\top$.

and exploration by maximizing the estimated total satisfaction $f_t(\pi; \overline{\boldsymbol{\theta}}_t)$ and the bonus term

$$g_t(\pi) = c_1 \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i, \pi(i))\|_{\boldsymbol{V}_t^{-1}}, \tag{3}$$

where $\boldsymbol{V}_t = \lambda_0 \boldsymbol{I} + \sum_{s=1}^{t-1} \sum_{i=1}^{N} \boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^\top$. The bonus term is related to the width of the confidence interval.

4.1.1 Regret analysis

Algorithm 1 achieves the following regret:

Theorem 4.1. *Fix any $\delta \in (0, 1)$. Algorithm 1, with tuning parameters chosen as in Appendix D.2, achieves with probability at least $1 - 2\delta$ the regret bound $\mathcal{R}_T^\alpha = \widetilde{O}(\kappa_\mu^{-1} L_r L_\mu D(d\sqrt{NT} + dN))$.*

This bound matches the known lower bound for contextual combinatorial linear bandits [43, Theorem 7] in its dependence on d, T, and N, up to logarithmic factors. If $f(\pi; \boldsymbol{\theta}_1) - f(\pi; \boldsymbol{\theta}_2) \leq C \sum_i |\mu(\boldsymbol{x}_t(i)^\top \boldsymbol{\theta}_1) - \mu(\boldsymbol{x}_t(i)^\top \boldsymbol{\theta}_2)|$ holds for any $\pi \in \Pi$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, where $C > 0$ is a constant, then we can derive a similar bound for a general CCGLS as well. While we use the regularization in Algorithm 1, a similar bound can be obtained via an initial exploration. Using the initial exploration, however, introduces an additional regret term and requires assumptions on $\boldsymbol{\phi}_t$ [2]. The full statement and proof of Theorem 4.1 are given in Appendix D.2.

4.1.2 Approximate optimization oracle construction

We next describe one concrete way to instantiate the α-approximate optimization oracle used by CAB-UCB. For a parameter $\boldsymbol{\theta}$, define $g_a(S; \boldsymbol{\theta}) = r(\sum_{i \in S} \mu(\boldsymbol{\phi}_t(i, a)^\top \boldsymbol{\theta}))$. By the concavity and monotonicity of r, each g_a is monotone submodular. Hence, maximizing $f_t(\pi; \boldsymbol{\theta})$ is an instance of the submodular welfare problem. The UCB bonus $g_t(\pi)$ is additive, so $f_t(\pi; \boldsymbol{\theta}) + g_t(\pi)$ remains within the same problem class. Therefore, Lemma 2.1 provides a concrete α-approximate optimization oracle for the CAB-UCB allocation step.

4.2 Thompson sampling algorithm

Here, we introduce CAB-TS, which is based on the TS method [44]. The TS algorithm has been proposed for various bandits problems. Theoretically, in these problems, the TS algorithm often has worse regret upper bounds than the UCB algorithms [5, 42, 2]. Empirically, however, TS has often been found to perform comparably to, and sometimes better than, UCB algorithms [12, 35, 47].

CAB-TS has parameters $\lambda_0 > 0$ and $a > 0$, which control the regularization strength and the degree of exploration, respectively. Up to the step of computing $\overline{\boldsymbol{\theta}}_t$ via the regularized MLE, the procedure is identical to Algorithm 1. However, the subsequent method for computing π_t differs. In CAB-TS, after computing $\overline{\boldsymbol{\theta}}_t$, for each $i \in [N]$, we independently sample $\widetilde{\boldsymbol{\varepsilon}}_t(i)$ from $\mathcal{N}(\boldsymbol{0}, a^2 \boldsymbol{H}_t^{-1})$, where $\boldsymbol{H}_1 = L_\mu \lambda_0 \boldsymbol{I}$ and $\boldsymbol{H}_t = \sum_{s=1}^{t-1} \sum_{i=1}^{N} \dot{\mu}(\boldsymbol{x}_s(i)^\top \overline{\boldsymbol{\theta}}_t) (\boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^\top + \frac{\lambda_0}{N(t-1)} \boldsymbol{I})$ for $t \geq 2$. The isotropic regularization term is intentionally placed inside the weighted sum. With the scaling

[2] *E.g.,* [31] study generalized linear contextual bandits using initial exploration, where $\boldsymbol{\phi}_t$ is generated in a stochastic manner and additional regularity assumptions are imposed.

$\lambda_0/N(t-1)$, the matrices inside this weighted sum add up to $\boldsymbol{V}_t$, because the $(t-1)$ copies of $\lambda_0\boldsymbol{I}/(t-1)$ add to $\lambda_0\boldsymbol{I}$. Since $\dot{\mu}(z) \in [\kappa_\mu, L_\mu]$, this gives $\kappa_\mu \boldsymbol{V}_t \preceq \boldsymbol{H}_t \preceq L_\mu \boldsymbol{V}_t$. In what follows, we collectively denote these samples by $\widetilde{\mathcal{E}}_t = \{\widetilde{\boldsymbol{\varepsilon}}_t(1), \ldots, \widetilde{\boldsymbol{\varepsilon}}_t(N)\}$. We choose the allocation π_t to maximize the objective function $f_t(\pi; \overline{\boldsymbol{\theta}}_t) + h_t(\pi; \widetilde{\mathcal{E}}_t)$, where

$$h_t(\pi; \widetilde{\mathcal{E}}) = \sum_{i=1}^{N} \boldsymbol{\phi}_t(i, \pi(i))^\top \widetilde{\boldsymbol{\varepsilon}}(i). \tag{4}$$

We approximate the posterior of $\boldsymbol{\theta}^*$ by the Laplace approximation. In this setting, we sample $\widetilde{\boldsymbol{\varepsilon}}_t(i)$ i.i.d. across users, because this independence is needed to preserve enough variance in the aggregate perturbation, although using a common $\widetilde{\boldsymbol{\varepsilon}}_t$ is also a natural idea (see Remark D.10 for details).

4.2.1 Regret analysis

Unlike CAB-UCB, our analysis of CAB-TS assumes access to an exact optimization oracle for $f_t(\pi; \overline{\boldsymbol{\theta}}_t) + h_t(\pi; \widetilde{\mathcal{E}}_t)$. Accordingly, the regret guarantee for CAB-TS is stated in terms of the standard regret $\mathcal{R}_T$. CAB-TS achieves the following regret bound:

Theorem 4.2. *Fix any $\delta \in (0, 1/T)$. CAB-TS, with tuning parameters chosen as in Appendix D.3, achieves the regret upper bound $\mathbb{E}[\mathcal{R}_T] = \widetilde{O}(\kappa_\mu^{-1} L_r L_\mu D(dN\sqrt{T} + dN^{3/2}))$.*

This regret upper bound has an extra $\sqrt{N}$ factor compared with CAB-UCB. The proof of Theorem 4.2 partitions Π into a "good" subset and its complement and lower-bounds the probability that π_t lies in the good subset. This argument uses the exact optimality of π_t. However, the approximate optimization oracle may return a near-optimal allocation outside this subset, so the same probability lower bound is unavailable (see Remark D.7 for details). The full statement and proof of Theorem 4.2 are given in Appendix D.3. Although the objective function $f_t(\pi; \overline{\boldsymbol{\theta}}_t) + h_t(\pi; \widetilde{\mathcal{E}}_t)$ is designed so that $\widetilde{\mathcal{E}}_t$ enters linearly, one can also consider a heuristic variant of CAB-TS that instead maximizes $\sum_{a \in [K]} r\big(\sum_{i \in \pi^{-1}(a)} \mu\big(\boldsymbol{\phi}_t(i,a)^\top (\overline{\boldsymbol{\theta}}_t + \widetilde{\boldsymbol{\varepsilon}}_t(i))\big)\big)$. We describe this heuristic in Appendix D.4 and include it only for empirical comparison. Similar to CAB-UCB, the analysis for CAB-TS can also be extended to a general CCGLS under the appropriate assumption.

4.3 One-pass update algorithm

CAB-UCB and CAB-TS solve a regularized MLE at round t using all prior observations. For simplicity, throughout this runtime discussion, we suppress the dependence on the d and N. With full-history computation, if the solver uses I_t iterations, the update cost grows linearly with the number of rounds, namely $O(tI_t)$. To avoid this linear growth, we replace the MLE update with a one-pass parameter update based on OMD [49]. However, this variant requires an additional self-concordance assumption on the link function. This assumption is used in prior work [33, 49], and holds for commonly used link functions such as the logistic and Poisson link functions.

Assumption 4.3. There exists a constant $R > 0$, such that for any $z \in \mathbb{R}$ $|\ddot{\mu}(z)| \le R\dot{\mu}(z)$.

We call the resulting procedure CAB-OFU with one-pass OMD update, where OFU denotes the principle of optimism in the face of uncertainty, which uses a confidence set constructed from past observations [3]. For each user, define the negative log-likelihood $\ell_{t,i}(\boldsymbol{\theta}) = -y_t(i)\boldsymbol{x}_t(i)^\top\boldsymbol{\theta} + m(\boldsymbol{x}_t(i)^\top\boldsymbol{\theta})$, where $\dot{m} = \mu$. We form the quadratic surrogate $\widetilde{\ell}_t(\boldsymbol{\theta}) = \sum_{i=1}^{N}(\langle \nabla_{\boldsymbol{\theta}} \ell_{t,i}(\boldsymbol{\theta}_t), \boldsymbol{\theta} - \boldsymbol{\theta}_t\rangle + \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|^2_{\nabla^2_{\boldsymbol{\theta}}\ell_{t,i}(\boldsymbol{\theta}_t)})$, and update the parameter by

$$\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta} \in \Theta}\big(\widetilde{\ell}_t(\boldsymbol{\theta}) + \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|^2_{\boldsymbol{Q}_t}\big), \tag{5}$$

and set $\boldsymbol{Q}_t = \lambda_{\mathrm{op}}\boldsymbol{I} + \sum_{s=1}^{t-1}\sum_{i=1}^{N} \nabla^2_{\boldsymbol{\theta}} \ell_{s,i}(\boldsymbol{\theta}_{s+1})$. Once $\boldsymbol{Q}_t$ is maintained incrementally, this subproblem uses only the current-round surrogate and $\boldsymbol{Q}_t$. If (5) is solved in $\widetilde{I}_t$ iterations, the update cost is $O(\widetilde{I}_t)$. We use the confidence set $C_t(\delta) = \{\boldsymbol{\theta} \in \Theta \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|_{\boldsymbol{Q}_t} \le \beta_t(\delta)\}$, with $\beta_t(\delta) = \widetilde{O}(D\sqrt{\lambda_{\mathrm{op}}} + \sqrt{d})$, suppressing logarithmic factors. Using this confidence set, we choose π_t satisfying $\max_{\boldsymbol{\theta} \in C_t(\delta)} f_t(\pi_t; \boldsymbol{\theta}) \ge \alpha \max_{\pi \in \Pi} \max_{\boldsymbol{\theta} \in C_t(\delta)} f_t(\pi; \boldsymbol{\theta})$.

By this update, we achieve the following regret bound:

Theorem 4.4. *The above update with suitable tuning parameters achieves* $\mathcal{R}_T^{\alpha} = \widetilde{O}(\kappa_{\mu}^{-1} \max\{d, N\}(\sqrt{dNT} + dN))$*, where we display only the dependence on* κ_{μ}*,* d*,* N*, and* T*.*

The full statement and proof are given in Appendix E.

5 Experiments

This section empirically evaluates CAB-UCB and CAB-TS using synthetic data. Our code to reproduce the experimental results is shared as Supplementary Material.

5.1 Setting

In synthetic experiments, we define the five-dimensional feature vector $\boldsymbol{\phi}(i, a) = \lambda \boldsymbol{\phi}_{pop}(i, a) + (1 - \lambda)\boldsymbol{\phi}_{base}(i, a)$, where $\boldsymbol{\phi}_{pop}$ and $\boldsymbol{\phi}_{base}$ are sampled from the standard normal distribution. We impose $\boldsymbol{\phi}_{pop}(i, a) < \boldsymbol{\phi}_{pop}(i, a + 1)$ component-wise for all users i and all $a \in [K - 1]$. The parameter λ controls arm-popularity strength. A large λ makes all users prefer arms in a similar order, making it difficult to jointly maximize matches and arm satisfaction. We use $\mu(x) = 1/(1 + \exp(-x))$ and $r(x) = \min\{x, \beta\}$ as feedback mean and satisfaction functions, respectively. Thus, matches beyond β have no additional effect on satisfaction. Smaller β yields faster saturation.

We compare CAB-UCB (Algorithm 1), CAB-TS (ε) (Algorithm 2), the heuristic variant CAB-TS (θ) (Algorithm 3), and One-pass OMD (Algorithm 4) against three baselines, "Random", "Max match", and "FairX". Random selects arms uniformly at random. Max match is a UCB algorithm that aims to maximize cumulative expected matches. FairX, a UCB-based fairness algorithm proposed by Wang et al. [46], ensures that each arm receives a share of exposure that is proportional to its expected match, aiming to reduce over-selection of specific arms. Full baseline specifications are given in Appendix F.1. The optimization routines used in the experiments are practical proxies for the offline allocation steps appearing in the theory. In particular, CAB-TS (θ) is included only as an empirical heuristic variant and is not covered by our theory, while the one-pass variant is an empirical proxy for the theory-side oracle-based procedure.

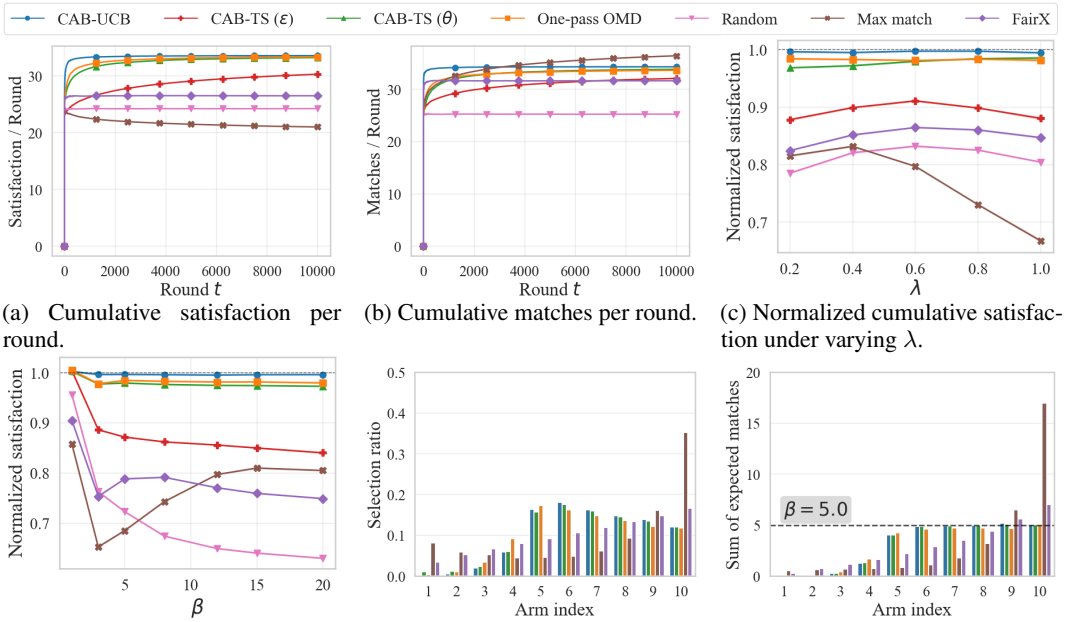
5.2 Results

We consider four settings. Fix $N = 50$ and $K = 10$ for all settings. The default comparisons use $T = 10000$, $\lambda = 0.5$, and $\beta = 5.0$ over 10 runs. The λ- and β-sweeps vary the respective parameter with $T = 5000$ over 5 runs. Histogram analyses use $T = 5000$, $\lambda = 1$, and $\beta = 5.0$ with 5 runs. We compare overall performance under the default setting, examine the effect of popularity concentration and satisfaction saturation by varying λ and β, and inspect the learned allocations.

We evaluate the learning performance of the proposed CAB algorithms in terms of cumulative satisfaction and matches. Figure 2a compares average cumulative satisfaction per round, and Figure 2b reports average cumulative matches per round. As shown in Figure 2a, all CAB variants substantially outperform the Max match and FairX, indicating that they successfully learn to optimize the satisfaction objective. CAB-UCB achieves the highest average cumulative satisfaction throughout the horizon. As intended, Max match yields the most matches in Figure 2b. However, this does not translate into maximizing satisfaction, since satisfaction is a concave function of the assigned users' expected matches. Under this fixed configuration, CAB-UCB achieves high cumulative satisfaction earlier than CAB-TS and the one-pass variant. This pattern is at least consistent with its sharper dependence on d and N in the regret bound, although the theory does not directly predict finite-horizon transients and the guarantees are stated under different oracle assumptions.

We vary λ and β to test whether the proposed methods remain effective when arm popularity is concentrated and when satisfaction strongly saturates. The λ-sweep tests robustness to preference concentration across users, while the β-sweep tests robustness to different concavity levels in the arm-side satisfaction function. In Figures 2c and 2d, each point reports the cumulative satisfaction under the corresponding parameter setting, normalized by that of a reference allocation computed using the true parameter and the same allocation routine as CAB-UCB.

The λ-sweep in Figure 2c examines robustness to preference concentration across users. As λ increases, users rank arms more similarly, causing match-oriented algorithms to over-concentrate assignments on a few popular arms, even after their satisfaction is already close to the saturation level



(a) Cumulative satisfaction per round.

(b) Cumulative matches per round.

(c) Normalized cumulative satisfaction under varying λ.

(d) Normalized cumulative satisfaction under varying β.

(e) Selection share over the entire horizon.

(f) Average expected matches of each arm.

Figure 2: Figures (a) and (b) use the default setting ($N = 50$, $K = 10$, $T = 10000$, $\lambda = 0.5$, $\beta = 5.0$). Figures (c) and (d) vary λ and β, respectively, with all other parameters fixed. Figures (e) and (f) show learned allocations for $\lambda = 1$ and $\beta = 5.0$: selection shares over the full horizon and arm-wise average expected matches over the last 100 steps. Solid lines show run means, and shaded regions indicate 95% confidence intervals where applicable.

β. This increases the gap between maximizing matches and maximizing satisfaction, since additional matches on well-served arms yield little marginal satisfaction. Consistent with this mechanism, the normalized cumulative satisfaction of Max match decreases with λ, whereas CAB-UCB remains nearly flat and stays close to the reference value across the entire sweep. The gap to FairX further suggests that reducing exposure imbalance alone is insufficient unless the allocation is explicitly aligned with the saturation structure of satisfaction.

The β-sweep in Figure 2d tests robustness to satisfaction saturation. As β increases, $r(x) = \min\{x, \beta\}$ becomes closer to linear over a wider range, and the satisfaction objective becomes less sensitive to over-allocation to already well-served arms. Accordingly, Max match improves with β, while the proposed methods maintain strong performance across the entire sweep. These results support the claim that explicitly modeling arm-side satisfaction is particularly important when utility saturates strongly, while CAB-UCB remains robust across different saturation regimes.

The allocation histograms in Figures 2e and 2f examine whether the proposed methods avoid excessive arm concentration. Figure 2e shows selection shares over the full horizon. When $\lambda = 1$, Max match heavily favors the most popular arms because all users share the same arm ranking. In contrast, the proposed methods select arms at more balanced rates, except for the least popular ones. FairX also reduces concentration, but allocates to unpopular arms rarely selected by the proposed methods, as it targets fairness in expected matches rather than the saturation structure of arm-side utility. Figure 2f reports arm-wise total expected matches averaged over the last 100 steps. Max match assigns high expected matches mainly to the most popular arms, often exceeding the saturation threshold $\beta = 5.0$. The proposed methods instead allocate not only to the most popular arms but also to several moderately popular arms, keeping expected matches closer to the saturation threshold. These results support the claim that CAB improves satisfaction by aligning allocations with diminishing returns, not merely by reducing imbalance.

Additional experimental results, including runtime comparisons, sweeps over other parameters, and further details of the experimental setup, are provided in Appendix F.

6 Conclusions

We proposed CAB, developed its algorithms, established regret upper bounds, and conducted experimental evaluations of its performance. We conclude with several future research directions. One possible direction is to improve the dependence on κ_μ. In more specific settings, recent analyses have reduced such dependence under a self-concordance assumption [33, 49]. Second, the current CAB-TS analysis assumes an exact optimization oracle, whereas CAB-UCB only requires an approximate one. Relaxing this assumption would require overcoming the proof obstacle that an approximate optimization oracle may return an unfavorable allocation among near-optimal allocations.

Disclosure of Funding

SI is supported by JSPS KAKENHI Grant Number JP25K03184 and by JST PRESTO, Japan, Grant Number JPMJPR2511.

References

[1] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[2] Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR, 2017.

[3] Rajeev Agrawal. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

[4] Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in neural information processing systems*, 29, 2016.

[5] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 127–135. PMLR, 2013.

[6] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[7] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[8] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, volume 54, 2013.

[9] Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134. PMLR, 2014.

[10] Semih Cayci, Swati Gupta, and Atilla Eryilmaz. Group-fair online allocation in continuous time. In *Advances in Neural Information Processing Systems*, volume 33, pages 13750–13761. Curran Associates, Inc., 2020.

[11] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge university press, 2006.

[12] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[13] Kani Chen, Inchi Hu, and Zhiliang Ying. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4):1155 – 1163, 1999.

[14] Qin Ding, Cho-Jui Hsieh, and James Sharpnack. An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and thompson sampling. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1585–1593. PMLR, 2021.

[15] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[16] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. Freeman, USA, 1979.

[17] Zhiming Huang, Yifan Xu, Bingshan Hu, Qipeng Wang, and Jianping Pan. Thompson sampling for combinatorial semi-bandits with sleeping arms and long-term fairness constraints. *arXiv preprint arXiv:2005.06725*, 2020.

[18] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016.

[19] Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5148–5157. PMLR, 2021.

[21] Subhash Khot, Richard J. Lipton, Evangelos Markakis, and Aranyak Mehta. Inapproximability results for combinatorial auctions with submodular utility functions. In *Algorithmica*, 52, pages 3–18. Springer-Verlag, 2008.

[22] Wonyoung Kim, Kyungbok Lee, and Myunghee Cho Paik. Double doubly robust thompson sampling for generalized linear contextual bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8300–8307, 2023.

[23] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR, 2015.

[24] Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 2066–2076. PMLR, 2020.

[25] Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3):1091 – 1114, 1987.

[26] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[27] Jungyhun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024.

[28] Benny Lehmann, Daniel Lehmann, and Noam Nisan. Combinatorial auctions with decreasing marginal utilities. In *Proceedings of the 3rd ACM Conference on Electronic Commerce*, EC '01, page 18–28. Association for Computing Machinery, 2001.

[29] Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813, 2019.

[30] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[31] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2071–2080. PMLR, 2017.

[32] Qingsong Liu, Weihang Xu, Siwei Wang, and Zhixuan Fang. Combinatorial bandits with linear constraints: Beyond knapsacks and fairness. *Advances in Neural Information Processing Systems*, 35:2997–3010, 2022.

[33] Xutong Liu, Xiangxiang Dai, Xuchuang Wang, Mohammad Hajiesmaili, and John C.S. Lui. Combinatorial logistic bandits. In *Abstracts of the 2025 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '25, page 112–114, New York, NY, USA, 2025. Association for Computing Machinery.

[34] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Number 9780195102680 in OUP Catalogue. Oxford University Press, 1995.

[35] Benedict C. May, Nathan Korda, Anthony Lee, and David S. Leslie. Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, 13(67): 2069–2106, 2012.

[36] Peter McCullagh and John A. Nelder. *Generalized linear models*. Chapman & Hall, London; New York, 1983.

[37] John W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1-2):122–136, 1964. ISSN 00129682, 14680262.

[38] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM, 2014.

[39] Karthik Abinav Sankararaman and Aleksandrs Slivkins. Combinatorial semi-bandits with knapsacks. In *International Conference on Artificial Intelligence and Statistics*, pages 1760–1770. PMLR, 2018.

[40] Tareq Si Salem, Georgios Iosifidis, and Giovanni Neglia. Enabling long-term fairness in dynamic resource allocation. *Proc. ACM Meas. Anal. Comput. Syst.*, 6(3), 2022.

[41] Abhishek Sinha, Ativ Joshi, Rajarshi Bhattacharjee, Cameron Musco, and Mohammad Hajiesmaili. No-regret algorithms for fair resource allocation. In *Advances in Neural Information Processing Systems*, volume 36, pages 48083–48109. Curran Associates, Inc., 2023.

[42] Kei Takemura and Shinji Ito. An arm-wise randomization approach to combinatorial linear semi-bandits. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1318–1323, 2019.

[43] Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. Near-optimal regret bounds for contextual combinatorial semi-bandits with linear payoff functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9791–9798, 2021.

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

[45] Jan Vondrak. Optimal approximation for the submodular welfare problem in the value oracle model. *Proceedings of 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 67–74, 2008.

[46] Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10686–10696. PMLR, 2021.

[47] Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5114–5122. PMLR, 2018.

[48] Huanle Xu, Yang Liu, Wing Cheong Lau, and Rui Li. Combinatorial multi-armed bandits with concave rewards and fairness constraints. In *IJCAI*, pages 2554–2560, 2020.

[49] Yu-Jie Zhang, Sheng-An Xu, Peng Zhao, and Masashi Sugiyama. Generalized linear bandits: Almost optimal regret with one-pass update, 2025. URL https://arxiv.org/abs/2507.11847.

A Notations

Table 1 summarizes the symbols used in this paper.

Table 1: Notation

Symbol	Meaning
$T \in \mathbb{N}$	time horizon
$N \in \mathbb{N}$	the number of users
$K \in \mathbb{N}$	the number of arms
$d \in \mathbb{N}$	dimension of feature vector
σ	sub-Gaussian parameter
κ_μ	parameter satisfying Assumption 3.1
L_μ	Lipschitz constant of function μ
L_r	Lipschitz constant of function r
$D > 0$	Upper bound on $\|\boldsymbol{\theta}^*\|_2$
$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \leq D\}$	bounded parameter space
$\mu \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$	expectation of feedback
$r \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$	satisfaction function
$\boldsymbol{\phi}_t(i, a)$	feature vector according to user i and arm a at round t
$y_t(i)$	feedback for i at round t
$\Pi = \{\pi : [N] \to [K]\}$	set of all functions from $[N]$ to $[K]$
$f_t(\pi; \boldsymbol{\theta}) = \sum_{a \in [K]} r(\sum_{i \in \pi^{-1}(a)} \mu(\boldsymbol{\phi}_t(i, a)^\top \boldsymbol{\theta}))$	cumulative expected satisfaction at round t
$\pi_t \in \Pi$	chosen allocation at round t
$\pi_t^* = \arg\max_{\pi \in \Pi} f_t(\pi; \boldsymbol{\theta}^*)$	optimal allocation at round t
$\boldsymbol{x}_t(i) = \boldsymbol{\phi}_t(i, \pi_t(i))$	chosen feature vector for user i at round t
$\boldsymbol{V}_t = \sum_{s=1}^{t-1} \sum_{i=1}^{N} \boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^\top + \lambda_0 \boldsymbol{I}$	information matrix at round t
$\boldsymbol{\theta}^*$	unknown parameter
$\overline{\boldsymbol{\theta}}_t$	MLE of $\boldsymbol{\theta}^*$ at round t
$\mathcal{R}_T^\alpha = \sum_{t=1}^{T} (\alpha f_t(\pi_t^*; \boldsymbol{\theta}^*) - f_t(\pi_t; \boldsymbol{\theta}^*))$	approximate regret
$\mathcal{R}_T = \sum_{t=1}^{T} (f_t(\pi_t^*; \boldsymbol{\theta}^*) - f_t(\pi_t; \boldsymbol{\theta}^*))$	standard regret

B Related work

We introduce several related works.

Contextual Combinatorial Semi-bandits and Generalized Linear Models From a technical perspective, a closely related problem setting is that of contextual combinatorial semi-bandits (CCS). CCS was first studied by Qin et al. [38]. CCS are problems in which, at each round, one first observes the arms together with their associated contexts, then selects a combination of arms based on these observations and past outcomes, and finally observes the reward, which depends on the chosen contexts and an unknown parameter. They consider a general framework that includes nonlinear reward functions and propose a UCB algorithm, while assuming a linear model for the feedback. In the linear-feedback setting, their algorithm achieves a regret upper bound of $\widetilde{O}(\max\{\sqrt{d}, \sqrt{N}\}\sqrt{dNT})$. To the best of our knowledge, the best known bound is $\widetilde{O}(d\sqrt{NT} + dN)$, which is achieved by the UCB algorithm of Takemura et al. [43]. In Takemura and Ito [42], in addition to UCB algorithms, a TS algorithm was studied under the setting where the reward is linear in the feedback, and it was shown that a regret upper bound of $\mathcal{R}_T = \widetilde{O}(\max\{d, \sqrt{dN}\}\sqrt{dNT})$ can be achieved. For CCS with linear reward functions, the existing works have investigated lower bounds in addition to upper bounds.[3] Kveton et al. [23] established a lower bound for non-contextual combinatorial semi-bandits in terms of the number of base arms and the action size. Applying their bound to an instance with d

[3] Lower bounds for general reward functions are not meaningful. Indeed, if the reward is constant, the regret can always be reduced to 0.

base arms and action size N yields $\Omega(\min\{\sqrt{dNT}, NT\})$. For CCS, an improved lower bound of $\Omega(\min\{d\sqrt{NT} + dN, NT\})$ was derived by Takemura et al. [43].

As a related research direction, generalized linear (contextual) bandits have also been studied. In this framework, a generalized linear model is adopted as the feedback model, rather than a linear model, and the contextual bandits setting was first investigated by Filippi et al. [15]. In the non-combinatorial case, existing work on GLM has primarily focused on UCB algorithms [31, 19, 27, 49], while TS algorithms have also been developed [19, 14, 22]. On the other hand, to the best of our knowledge, there are only a few studies that consider combinatorial settings. One notable example is Liu et al. [33], which uses the UCB algorithm and considers a contextual setting in which the feedback is sampled from a Bernoulli distribution with its mean specified by a logistic model.

Fair Allocation One line of research with a closely related idea is fair resource allocation, although the technical relevance is limited. Among them, some studies use an evaluation metric called α-fairness, namely $fair_\alpha(x) = \frac{x^{1-\alpha}-1}{1-\alpha}$ $(0 \leq \alpha < 1)$. The function $fair_\alpha(x)$ is concave, and its use is close in spirit to the idea of this work [10, 40, 41]. However, they consider an objective function of the form $\sum_i fair(\sum_{t=1}^T w_t(i))$, where $w_t(i)$ denotes the utility for i at round t, which evaluates the overall fairness across the entire horizon. This differs from our setting. If one emphasizes the final fairness of the allocation over the whole period, their objective function is more appropriate. In contrast, we consider that dissatisfaction over a short period may also lead to churn. Therefore, as an objective function in CAB, we argue that $\sum_{t=1}^T f_t$ is appropriate.

Moreover, there are technical differences. Representative aspects are the feedback model and the action space. Since the true utility value or the reward vector is observed at the end of each round in their works, their feedback assumption is stronger than the setting we consider (*e.g.,*the platform observes only whether a match occurred). Regarding the action space, while Sinha et al. [41, Section 4] discusses integrality constraints, their formulation essentially considers fractional decisions rather than combinatorial ones. Thus, although the high-level idea is similar to our work, the technical setting is entirely different, and a direct comparison of theoretical results is impossible.

Bandits With Fairness Constraints One line of research addressing fairness in the context of the bandit problem is Joseph et al. [18]. In their work, fairness is defined as allocating arms without favoring any particular arm, based on their expected rewards. On the other hand, a problem that deals with a concept of fairness similar to that considered in this study is the combinatorial sleeping bandits with fairness constraints proposed by Li et al. [29]. In this problem, for each arm a, a minimum selection count n_a is specified, and the objective is to maximize the cumulative expected reward under this constraint. Li et al. [29] provided a UCB algorithm, while Huang et al. [17] later proposed a TS algorithm. Xu et al. [48] focused on the objective function, considering a setting where the total reward R, obtained as a linear combination of the rewards from the arms, is transformed by a strictly concave function f, resulting in the objective function $f(R)$. More recently, studies have considered constraints that combine knapsack constraints with fairness constraints [32].

Bandits With Knapsacks The Bandits with Knapsacks (BwK) problem extends the standard bandit setting by introducing knapsack-type resource constraints. Given budget limitations on multiple resources, the learner can no longer obtain rewards once any resource budget is depleted. The goal is to maximize the cumulative expected reward. As a method of introducing constraints into the multi-armed bandits, Badanidiyuru et al. [8] proposed the BwK framework, which incorporates budget constraints. Building on this framework, subsequent studies have extended the knapsack constraints to linear contextual bandits [9, 4] and combinatorial semi-bandits [39].

C NP-hardness

We discuss the computational complexity of this problem. As the following theorem shows, computing π_t^* and π_t is NP-hard.

Theorem C.1. *Consider a set of N items and K players, and let $r\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a monotone concave function. We consider the problem of maximizing $\sum_{i\in[K]} v_i(S_i)$ subject to $\bigcup_{i\in[K]} S_i = [N]$ and $S_i \cap S_j = \emptyset$ for $i \neq j$, where $v_a\colon 2^{[N]} \to \mathbb{R}_{\geq 0}$ is defined by $v_a(S) = r\left(\sum_{i\in S} w_{i,a}\right)$, and*

$w_{i,a} > 0$ denotes the value of item i for agent a. Then, there exists a concave and monotone increasing function r for which the above problem is NP-hard.

Our proof uses an approach similar to that of Lehmann et al. [28, Theorem 10].

Proof. We will perform a reduction from the well-known NP-complete problem "Subset Sum" (as detailed in Garey and Johnson [16]). The problem is as follows: given a sequence of integers $a_1, \ldots, a_m$ and a target total t, the objective is to determine if a subset S of these integers exists such that the sum of the elements in S equals t (i.e., $\sum_{i \in S} a_i = t$). Based on this input, we will construct two valuations for the m items.

We consider the decision problem of determining whether $\max f(\mathcal{F}) = Vt + V - t$ holds for $\mathcal{F} = \{U_1, U_2\}$ with $U_1 \cap U_2 = \emptyset$, $U_1 \cup U_2 = U$, $f(\mathcal{F}) = v_1(U_1) + v_2(U_2)$, $r(x) = \min\{Vt, x\}$, and $V = \sum_{i \in U} a_i$. Let $w_{i,1} = Va_i, w_{i,2} = a_i$. We allocate S to valuation 1 and S^c to valuation 2. We examine three cases: $\sum_{i \in S} a_i = t$, $\sum_{i \in S} a_i < t$, and $\sum_{i \in S} a_i > t$.

- If $\sum_{i \in S} a_i = t$, then $v_1(S) + v_2(S^c) = Vt + V - t$.
- If $\sum_{i \in S} a_i < t$, then $v_1(S) + v_2(S^c) = V \sum_{i \in S} a_i + V - \sum_{i \in S} a_i < Vt + V - t$.
- If $\sum_{i \in S} a_i > t$, then $v_1(S) + v_2(S^c) = Vt + V - \sum_{i \in S} a_i < Vt + V - t$.

Therefore, the instance of "Subset Sum" is a Yes-instance precisely when the proposition is satisfied, and a No-instance otherwise. □

D Details of Sections 4.1 and 4.2

In this section, we provide the omitted proofs in Sections 4.1 and 4.2.

D.1 Technical lemmas

Here, in preparation for the proofs of the theorems, we introduce technical lemmas that are commonly used in the proofs of Theorems 4.1 and 4.2.

The following lemma guarantees that the regularized MLE $\overline{\boldsymbol{\theta}}_t$ remains close to the true parameter with high probability. While Kveton et al. [24, Lemma 9] relies on an initial exploration phase to establish this property, an analogous result can be derived by employing regularization instead.

Lemma D.1. *Assume that $\lambda_0 \geq \frac{\sigma^2}{\kappa_\mu^2}\big(d \log\big(1 + \frac{NT}{d}\big) + 2 \log \frac{1}{\delta}\big)$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all $t \in [T]$,*

$$\|\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 \leq D + 1.$$

Proof. Let $S_t = \sum_{s=1}^{t-1} \sum_{i=1}^{N} \big(y_s(i) - \mu(\boldsymbol{x}_s(i)^\top \boldsymbol{\theta}^*)\big)\boldsymbol{x}_s(i)$. First, using Chen et al. [13, Lemma A], we show that the following proposition holds:

$$\|S_t - \kappa_\mu \lambda_0 \boldsymbol{\theta}^*\|_{\boldsymbol{V}_t^{-1}} \leq \kappa_\mu (D + 1)\sqrt{\lambda_0} \Rightarrow \|\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 \leq D + 1.$$

Define the map $\boldsymbol{G}_t(\boldsymbol{\theta}) = \sum_{s=1}^{t-1} \sum_{i=1}^{N} \mu(\boldsymbol{x}_s(i)^\top \boldsymbol{\theta})\boldsymbol{x}_s(i) + \kappa_\mu \lambda_0 \boldsymbol{\theta}$. From the definition of $\overline{\boldsymbol{\theta}}_t$, we have $\boldsymbol{G}_t(\overline{\boldsymbol{\theta}}_t) = \sum_{s=1}^{t-1} \sum_{i=1}^{N} y_s(i)\boldsymbol{x}_s(i)$. Thus, it holds that

$$\boldsymbol{G}_t(\overline{\boldsymbol{\theta}}_t) - \boldsymbol{G}_t(\boldsymbol{\theta}^*) = \sum_{s=1}^{t-1} \sum_{i=1}^{N} (y_s(i) - \mu(\boldsymbol{x}_s(i)^\top \boldsymbol{\theta}^*))\boldsymbol{x}_s(i) - \kappa_\mu \lambda_0 \boldsymbol{\theta}^* = S_t - \kappa_\mu \lambda_0 \boldsymbol{\theta}^*.$$

Moreover, for any $\boldsymbol{\theta} \in \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq D + 1\}$, from Assumption 3.1, we have

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{G}_t(\boldsymbol{\theta}) = \sum_{s=1}^{t-1} \sum_{i=1}^{N} \dot{\mu}(\boldsymbol{x}_s(i)^\top \boldsymbol{\theta})\boldsymbol{x}_s(i)\boldsymbol{x}_s(i)^\top + \kappa_\mu \lambda_0 \boldsymbol{I}$$

$$\succeq \kappa_\mu \sum_{s=1}^{t-1}\sum_{i=1}^{N} \boldsymbol{x}_s(i)\boldsymbol{x}_s(i)^\top + \kappa_\mu\lambda_0\boldsymbol{I} = \kappa_\mu \boldsymbol{V}_t.$$

Therefore, applying Chen et al. [13, Lemma A] to the map $\boldsymbol{\theta} \mapsto {\boldsymbol{V}_t}^{-1/2}(\boldsymbol{G}_t(\boldsymbol{\theta}) - \boldsymbol{G}_t(\boldsymbol{\theta}^*))$ yields the proposition

$$\|S_t - \kappa_\mu\lambda_0\boldsymbol{\theta}^*\|_{\boldsymbol{V}_t^{-1}} \le \kappa_\mu(D+1)\sqrt{\lambda_{\min}(\boldsymbol{V}_t)} \Rightarrow \|\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 \le D+1.$$

Since $\lambda_{\min}(\boldsymbol{V}_t) \ge \lambda_0$, we obtain

$$\|S_t - \kappa_\mu\lambda_0\boldsymbol{\theta}^*\|_{\boldsymbol{V}_t^{-1}} \le \kappa_\mu(D+1)\sqrt{\lambda_0} \Rightarrow \|\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 \le D+1.$$

Next, we upper bound $\|S_t - \kappa_\mu\lambda_0\boldsymbol{\theta}^*\|_{\boldsymbol{V}_t^{-1}}$. By the triangle inequality and $\boldsymbol{V}_t \succeq \lambda_0\boldsymbol{I}$, we have

$$\begin{aligned}\|S_t - \kappa_\mu\lambda_0\boldsymbol{\theta}^*\|_{\boldsymbol{V}_t^{-1}} &\le \|S_t\|_{\boldsymbol{V}_t^{-1}} + \kappa_\mu\lambda_0\|\boldsymbol{\theta}^*\|_{\boldsymbol{V}_t^{-1}} \le \|S_t\|_{\boldsymbol{V}_t^{-1}} + \kappa_\mu\sqrt{\lambda_0}\|\boldsymbol{\theta}^*\|_2 \\ &\le \|S_t\|_{\boldsymbol{V}_t^{-1}} + \kappa_\mu D\sqrt{\lambda_0},\end{aligned}$$

where the last inequality follows from $\|\boldsymbol{\theta}^*\|_2 \le D$. Thus, we have

$$\|S_t\|_{\boldsymbol{V}_t^{-1}} \le \kappa_\mu\sqrt{\lambda_0} \Rightarrow \|\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 \le D+1. \tag{6}$$

Here, from Abbasi-yadkori et al. [1, Theorem 1], with probability at least $1-\delta$, for all $t \ge 1$, it holds that

$$\|S_t\|^2_{\boldsymbol{V}_t^{-1}} \le 2\sigma^2 \log\left(\frac{\det(\boldsymbol{V}_t)^{1/2}\det(\lambda_0\boldsymbol{I})^{-1/2}}{\delta}\right) \le \sigma^2\left(d\log\left(1+\frac{NT}{d\lambda_0}\right) + 2\log\frac{1}{\delta}\right), \tag{7}$$

where the second inequality follows from $\det(\boldsymbol{V}_t) \le ((\mathrm{tr}(\lambda_0\boldsymbol{I}) + Nt)/d)^d = (\lambda_0 + Nt/d)^d$. Thus, using (7) and the definition of λ_0, it holds that $\|S_t\|_{\boldsymbol{V}_t^{-1}} \le \kappa_\mu\sqrt{\lambda_0}$. Therefore, by (6), we have $\|\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 \le D+1$ for any t with probability at least $1-\delta$. □

The following lemma shows that the regularized MLE also controls the error of the objective function.

Lemma D.2. *Assume that it holds that $\|\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 \le D+1$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $t \in [T]$,*

$$|f_t(\pi;\boldsymbol{\theta}^*) - f_t(\pi;\overline{\boldsymbol{\theta}}_t)| \le c_1 \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i,\pi(i))\|_{\boldsymbol{V}_t^{-1}},$$

where $c_1 = \kappa_\mu^{-1}L_rL_\mu\left(\sigma\sqrt{d\log\left(1+\frac{NT}{d\lambda_0}\right) + 2\log\frac{1}{\delta}} + \kappa_\mu D\sqrt{\lambda_0}\right)$.

Proof. Let $S_t = \sum_{s=1}^{t-1}\sum_{i=1}^{N}(y_s(i) - \mu(\boldsymbol{x}_s(i)^\top\boldsymbol{\theta}^*))\boldsymbol{x}_s(i)$, $\mathcal{D}_1 = \{\boldsymbol{x}_s(i), y_s(i)\}_{i\in[N],s<t}$, and $\mathcal{D}_2 = \{\boldsymbol{x}_s(i), \mu(\boldsymbol{x}_s(i)^\top\boldsymbol{\theta}^*)\}_{i\in[N],s<t}$.

First, we expand Kveton et al. [24, Lemma 1] to the regularized MLE. It holds that

$$\begin{aligned}S_t &= \nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathcal{D}_2;\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathcal{D}_1;\boldsymbol{\theta}^*) = \nabla_{\boldsymbol{\theta}}\widetilde{\mathcal{L}}(\mathcal{D}_1;\overline{\boldsymbol{\theta}}_t,\kappa_\mu\lambda_0) - \nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathcal{D}_1;\boldsymbol{\theta}^*) \\ &= \nabla_{\boldsymbol{\theta}}\widetilde{\mathcal{L}}(\mathcal{D}_1;\overline{\boldsymbol{\theta}}_t,\kappa_\mu\lambda_0) - \nabla_{\boldsymbol{\theta}}\widetilde{\mathcal{L}}(\mathcal{D}_1;\boldsymbol{\theta}^*,\kappa_\mu\lambda_0) + \kappa_\mu\lambda_0\boldsymbol{\theta}^* \\ &= \nabla^2_{\boldsymbol{\theta}}\widetilde{\mathcal{L}}(\mathcal{D}_1;\boldsymbol{\theta}',\kappa_\mu\lambda_0)(\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*) + \kappa_\mu\lambda_0\boldsymbol{\theta}^* \\ &= \boldsymbol{V}(\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*) + \kappa_\mu\lambda_0\boldsymbol{\theta}^*,\end{aligned}$$

where $\boldsymbol{\theta}'$ is a convex combination of $\overline{\boldsymbol{\theta}}_t$ and $\boldsymbol{\theta}^*$, and $\boldsymbol{V} = \nabla^2_{\boldsymbol{\theta}}\widetilde{\mathcal{L}}(\mathcal{D}_1;\boldsymbol{\theta}',\kappa_\mu\lambda_0)$. We used $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathcal{D}_2;\boldsymbol{\theta}^*) = \nabla_{\boldsymbol{\theta}}\widetilde{\mathcal{L}}(\mathcal{D}_1;\overline{\boldsymbol{\theta}}_t,\kappa_\mu\lambda_0) = 0$ in the second equality.

Next, we bound $|f_t(\pi;\boldsymbol{\theta}^*) - f_t(\pi;\overline{\boldsymbol{\theta}}_t)|$.

$$|f_t(\pi;\boldsymbol{\theta}^*) - f_t(\pi;\overline{\boldsymbol{\theta}}_t)| \le L_rL_\mu \sum_{i=1}^{N} |\boldsymbol{x}_t(i)^\top(\boldsymbol{\theta}^* - \overline{\boldsymbol{\theta}}_t)|$$

$$
\begin{aligned}
&\leq L_r L_\mu \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} \|\boldsymbol{\theta}^* - \overline{\boldsymbol{\theta}}_t\|_{\boldsymbol{V}_t} \\
&= L_r L_\mu \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} \|\boldsymbol{V}^{-1} S_t - \kappa_\mu \lambda_0 \boldsymbol{V}^{-1} \boldsymbol{\theta}^*\|_{\boldsymbol{V}_t} \\
&\leq L_r L_\mu \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} (\|\boldsymbol{V}^{-1} S_t\|_{\boldsymbol{V}_t} + \kappa_\mu \lambda_0 \|\boldsymbol{V}^{-1} \boldsymbol{\theta}^*\|_{\boldsymbol{V}_t}) \\
&\leq L_r L_\mu \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} (\sqrt{S_t^\top \boldsymbol{V}^{-1} \boldsymbol{V}_t \boldsymbol{V}^{-1} S_t} + \kappa_\mu \lambda_0 \sqrt{\boldsymbol{\theta}^{*\top} \boldsymbol{V}^{-1} \boldsymbol{V}_t \boldsymbol{V}^{-1} \boldsymbol{\theta}^*}) \\
&\leq \kappa_\mu^{-1} L_r L_\mu \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} (\|S_t\|_{\boldsymbol{V}_t^{-1}} + \kappa_\mu D \sqrt{\lambda_0}),
\end{aligned}
$$

where the second inequality follows from the Cauchy–Schwarz inequality, and the last inequality follows from the $\kappa_\mu \boldsymbol{V}_t \preceq \boldsymbol{V}$ on $\|\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 \leq D + 1$.

Therefore, combining the above inequality with (7) in the proof of Lemma D.1, we have that, with probability at least $1 - \delta$, for all $t \geq 1$,

$$
|f_t(\pi; \boldsymbol{\theta}^*) - f_t(\pi; \overline{\boldsymbol{\theta}}_t)| \leq \kappa_\mu^{-1} L_r L_\mu \left(\sigma \sqrt{d \log\left(1 + \frac{NT}{d\lambda_0}\right) + 2 \log \frac{1}{\delta}} + \kappa_\mu D \sqrt{\lambda_0} \right) \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}}.
$$

□

We can bound $\sum_{t=1}^{T} \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}}$ using the following lemma:

Lemma D.3 (Takemura et al. 43, Lemma 2)**.** *Let $\{\boldsymbol{x}_t(i)\}_{(i,t) \in [N] \times \mathbb{N}}$ be a sequence in $\mathbb{R}^d$ satisfying $\|\boldsymbol{x}_t(i)\|_2 \leq 1$. For all $t \geq 1$, define $\boldsymbol{V}_t = \lambda_0 \boldsymbol{I} + \sum_{s=1}^{t-1} \sum_{i=1}^{N} \boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^\top$, where $\lambda_0 \geq 0$. It holds that*

$$
\begin{aligned}
&\sum_{t=1}^{T} \sum_{i=1}^{N} \min\left\{ \frac{1}{\sqrt{N}}, \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} \right\} \leq \sqrt{2dNT \log\left(1 + \frac{NT}{d\lambda_0}\right)}, \quad \textit{and} \\
&\sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{1}\left[\|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} > \frac{1}{\sqrt{N}} \right] < 2dN \log\left(1 + \frac{NT}{d\lambda_0}\right).
\end{aligned}
$$

This lemma implies that

$$
\begin{aligned}
\sum_{t=1}^{T} \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} &= \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{1}\left[\|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} > \frac{1}{\sqrt{N}} \right] \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} \\
&\quad + \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{1}\left[\|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} \leq \frac{1}{\sqrt{N}} \right] \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} \\
&\leq \frac{1}{\sqrt{\lambda_0}} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{1}\left[\|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} > \frac{1}{\sqrt{N}} \right] + \sum_{t=1}^{T} \sum_{i=1}^{N} \min\left\{ \frac{1}{\sqrt{N}}, \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} \right\} \\
&\leq \sqrt{2dNT \log\left(1 + \frac{NT}{d\lambda_0}\right)} + \frac{2dN}{\sqrt{\lambda_0}} \log\left(1 + \frac{NT}{d\lambda_0}\right) \qquad (8)
\end{aligned}
$$

D.2 Details of Section 4.1

We provide the full version of Theorem 4.1.

Theorem D.4. *Fix any* $\delta \in (0,1)$. *If we run Algorithm 1 with* $c_1 = \frac{L_r L_\mu}{\kappa_\mu}\big(\sigma\sqrt{d\log\Big(1+\frac{NT}{d\lambda_0}\Big)+2\log\frac{1}{\delta}} + \kappa_\mu D\sqrt{\lambda_0}\big)$ *and* $\lambda_0 \geq \frac{\sigma^2}{\kappa_\mu^2}\big(d\log\big(1+\frac{NT}{d}\big)+2\log\frac{1}{\delta}\big)$, *then, with probability at least* $1-2\delta$, *the regret of algorithm is upper bounded by*

$$\mathcal{R}_T^\alpha \leq 2c_1\left(\sqrt{2dNT\log\left(1+\frac{NT}{d\lambda_0}\right)}+\frac{2dN}{\sqrt{\lambda_0}}\log\left(1+\frac{NT}{d\lambda_0}\right)\right).$$

If we set $\lambda_0 = \kappa_\mu^{-2}\sigma^2\big(d\log\big(1+\frac{NT}{d}\big)+2\log\frac{1}{\delta}\big)$, we have

$$\mathcal{R}_T^\alpha = \widetilde{O}(d\sqrt{NT}+dN).$$

Proof. We define $\boldsymbol{x}_t{}^*(i) = \boldsymbol{\phi}_t(i,\pi_t^*(i))$, $\Delta_t = \overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*$. First, we bound the one-step regret, $f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*)$. Here, we want to bound the following terms:

$$\begin{aligned}\alpha f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*) = {}& (\alpha f_t(\pi_t^*;\overline{\boldsymbol{\theta}}_t) - f_t(\pi_t;\overline{\boldsymbol{\theta}}_t)) \\ & + \alpha(f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t^*;\overline{\boldsymbol{\theta}}_t)) + (f_t(\pi_t;\overline{\boldsymbol{\theta}}_t) - f_t(\pi_t;\boldsymbol{\theta}^*)).\end{aligned}$$

From the definition of π_t, the first term is bounded as

$$\alpha f_t(\pi_t^*;\overline{\boldsymbol{\theta}}_t) - f_t(\pi_t;\overline{\boldsymbol{\theta}}_t) \leq g_t(\pi_t) - \alpha g_t(\pi_t^*) = c_1\sum_{i=1}^N(\|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} - \alpha\|\boldsymbol{x}_t{}^*(i)\|_{\boldsymbol{V}_t^{-1}}). \quad (9)$$

The second and third terms are bounded as

$$\alpha(f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t^*;\overline{\boldsymbol{\theta}}_t)) \leq \alpha c_1\sum_{i=1}^N\|\boldsymbol{x}_t{}^*(i)\|_{\boldsymbol{V}_t^{-1}}, \quad (10)$$

$$f_t(\pi_t;\overline{\boldsymbol{\theta}}_t) - f_t(\pi_t;\boldsymbol{\theta}^*) \leq c_1\sum_{i=1}^N\|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} \quad (11)$$

from Lemma D.2.

Combining (9), (10), and (11), with probability at least $1-2\delta$, it holds that

$$\alpha f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*) \leq 2c_1\sum_{i=1}^N\|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}}.$$

Since $\|\boldsymbol{x}_t(i)\|_2 \leq 1$, we have

$$\begin{aligned}\sum_{t=1}^T(\alpha f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*)) &\leq 2c_1\sum_{t=1}^T\sum_{i=1}^N\|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} \\ &\leq 2c_1\left(\sqrt{2dNT\log\left(1+\frac{NT}{d\lambda_0}\right)}+\frac{2dN}{\sqrt{\lambda_0}}\log\left(1+\frac{NT}{d\lambda_0}\right)\right), \quad (12)\end{aligned}$$

where the last inequality follows from (8). This is the desired upper bound. □

D.3 Details of Section 4.2.1

Algorithm 2 CAB-TS

Input: The total rounds T, the number of users N, tuning parameter λ_0 and a, and access to an exact optimization oracle.

$\mathcal{D}_1 \leftarrow \varnothing$, $\boldsymbol{V}_1 \leftarrow \lambda_0 \boldsymbol{I}$, and $\boldsymbol{H}_1 \leftarrow L_\mu \lambda_0 \boldsymbol{I}$.
for $t = 1, \ldots, T$ **do**
  $\boldsymbol{V}_t \leftarrow \lambda_0 \boldsymbol{I} + \sum_{s=1}^{t-1} \sum_{i=1}^{N} \boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^\top$.
  $\overline{\boldsymbol{\theta}}_t \leftarrow \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \widetilde{\mathcal{L}}(\mathcal{D}_t; \boldsymbol{\theta}, \kappa_\mu \lambda_0)$.
  If $t \geq 2$, $\boldsymbol{H}_t \leftarrow \sum_{s=1}^{t-1} \sum_{i=1}^{N} \dot{\mu}(\boldsymbol{x}_s(i)^\top \overline{\boldsymbol{\theta}}_t)(\boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^\top + \frac{\lambda_0}{N(t-1)} \boldsymbol{I})$.
  for $i = 1, \ldots, N$ **do**
    $\widetilde{\varepsilon}_t(i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, a^2 \boldsymbol{H}_t^{-1})$.
  Call an exact optimization oracle for $f_t(\pi; \overline{\boldsymbol{\theta}}_t) + h_t(\pi; \widetilde{\mathcal{E}}_t)$ and let π_t denote its output .
  Observe $y_t(i)$ for any $i \in [N]$.
  $\boldsymbol{x}_t(i) \leftarrow \boldsymbol{\phi}_t(i, \pi_t(i))$ for any $i \in [N]$ and $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(\boldsymbol{x}_t(i), y_t(i))\}_{i \in [N]}$.

For completeness, we provide the pseudo-code of CAB-TS (Algorithm 2) and the full statement of Theorem 4.2.

Theorem D.5. *Algorithm 2 with* $a = c_1 \sqrt{L_\mu N}$ *and* $\lambda_0 \geq \kappa_\mu^{-2} \sigma^2 \big(d \log\big(1 + \frac{NT}{d}\big) + 2 \log \frac{1}{\delta}\big)$ *achieves the following regret bound for any* $\delta \in (0, 1/T)$*:*

$$\mathbb{E}[\mathcal{R}_T] \leq (c_1 + c_2)\left(1 + \frac{2}{0.15 - 2\delta}\right)\left(\sqrt{2dNT \log\left(1 + \frac{NT}{d\lambda_0}\right)} + \frac{2dN}{\sqrt{\lambda_0}} \log\left(1 + \frac{NT}{d\lambda_0}\right)\right) + 4\delta KMT,$$

where $c_1 = \kappa_\mu^{-1} L_r L_\mu \left(\sigma \sqrt{d \log\big(1 + \frac{NT}{d\lambda_0}\big) + 2 \log \frac{1}{\delta}} + \kappa_\mu D \sqrt{\lambda_0}\right)$ *and* $c_2 = c_1 \sqrt{2 \kappa_\mu^{-1} L_\mu N \log \frac{KN}{\delta}}$.

If we set $\lambda_0 = \kappa_\mu^{-2} \sigma^2 \big(d \log\big(1 + \frac{NT}{d}\big) + 2 \log \frac{1}{\delta}\big)$, we have

$$\mathcal{R}_T = \widetilde{O}(dN\sqrt{T} + dN^{3/2}).$$

To prove Theorem D.5, we first bound the one-step regret. For convenience, we define

$$E_{1,t} = \{^\forall \pi \in \Pi \mid |f_t(\pi; \boldsymbol{\theta}^*) - f_t(\pi; \overline{\boldsymbol{\theta}}_t)| \leq c_1 \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i, \pi(i))\|_{\boldsymbol{V}_t^{-1}}\}. \tag{13}$$

Lemma D.6. *Define events* $E_{2,t}$ *and* $E_{3,t}$ *as*

$$E_{2,t} = \left\{^\forall \pi \in \Pi \mid |h_t(\pi; \widetilde{\mathcal{E}}_t)| \leq c_2 \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i, \pi(i))\|_{\boldsymbol{V}_t^{-1}}\right\}, \text{ and}$$

$$E_{3,t} = \left\{h_t(\pi_t^*; \widetilde{\mathcal{E}}_t) \geq c_1 \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i, \pi_t^*(i))\|_{\boldsymbol{V}_t^{-1}}\right\}.$$

Let $\mathbb{P}_t(E_{2,t}) \geq 1 - p_2$ *and* $\mathbb{P}_t(E_{3,t}) \geq p_3$ *(*$p_3 > p_2$*). If* $E_{1,t}$ *holds, then we have*

$$\mathbb{E}_t[f_t(\pi_t^*; \boldsymbol{\theta}^*) - f_t(\pi_t; \boldsymbol{\theta}^*)] \leq (c_1 + c_2)\left(1 + \frac{2}{p_3 - p_2}\right) \mathbb{E}_t\left[\sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}}\right] + p_2 KM$$

Proof. Let $c = c_1 + c_2$, $\mathcal{S}_t = \{\pi \in \Pi \mid c \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i, \pi(i))\|_{\boldsymbol{V}_t^{-1}} < f_t(\pi_t^*; \boldsymbol{\theta}^*) - f_t(\pi; \boldsymbol{\theta}^*)\}$, $\overline{\mathcal{S}}_t = \Pi / \mathcal{S}_t$, and $\pi_t' = \arg\min_{\pi \in \overline{\mathcal{S}}_t} \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i, \pi(i))\|_{\boldsymbol{V}_t^{-1}}$.

First, we will bound $f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*)$ on $E_{2,t}$. At round t on $E_{2,t}$, we have

$$\begin{aligned}
f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*) &\le f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t';\boldsymbol{\theta}^*) + f_t(\pi_t';\overline{\boldsymbol{\theta}}_t) + h_t(\pi_t';\widetilde{\mathcal{E}}_t) \\
&\quad - (f_t(\pi_t;\overline{\boldsymbol{\theta}}_t) + h_t(\pi_t;\widetilde{\mathcal{E}}_t)) + h_t(\pi_t;\widetilde{\mathcal{E}}_t) - h_t(\pi_t';\widetilde{\mathcal{E}}_t) \\
&\quad + c_1 \sum_{i=1}^{N} (\|\boldsymbol{\phi}_t(i,\pi_t'(i))\|_{\boldsymbol{V}_t^{-1}} + \|\boldsymbol{\phi}_t(i,\pi_t(i))\|_{\boldsymbol{V}_t^{-1}}) \\
&\le (c_1 + c_2) \sum_{i=1}^{N} (2\|\boldsymbol{\phi}_t(i,\pi_t'(i))\|_{\boldsymbol{V}_t^{-1}} + \|\boldsymbol{\phi}_t(i,\pi_t(i))\|_{\boldsymbol{V}_t^{-1}}),
\end{aligned}$$

where the first inequality follows from the definition of $E_{1,t}$, and the second inequality follows from the definition of $E_{2,t}$, the optimality of π_t, and the definition of $\mathcal{S}_t$ and π_t'.

Second, we want to bound $\mathbb{E}_t\left[\sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i,\pi_t'(i))\|_{\boldsymbol{V}_t^{-1}}\right]$ by $\mathbb{E}_t\left[\sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i,\pi_t(i))\|_{\boldsymbol{V}_t^{-1}}\right]$. Note that

$$\begin{aligned}
\mathbb{E}_t\left[\sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i,\pi_t(i))\|_{\boldsymbol{V}_t^{-1}}\right] &\ge \mathbb{E}_t\left[\sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i,\pi_t'(i))\|_{\boldsymbol{V}_t^{-1}} \mid \pi_t \in \overline{\mathcal{S}}_t\right] \mathbb{P}_t(\pi_t \in \overline{\mathcal{S}}_t) \\
&\ge \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i,\pi_t'(i))\|_{\boldsymbol{V}_t^{-1}} \mathbb{P}_t(\pi_t \in \overline{\mathcal{S}}_t).
\end{aligned}$$

Thus, $\mathbb{E}_t[f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*)] \le c(1 + 2/\mathbb{P}_t(\pi_t \in \overline{\mathcal{S}}_t)) \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i,\pi_t(i))\|_{\boldsymbol{V}_t^{-1}}$ on $E_{2,t}$.

Third, we will lower-bound $\mathbb{P}_t(\pi_t \in \overline{\mathcal{S}}_t)$.

$$\begin{aligned}
\mathbb{P}_t(\pi_t \in \overline{\mathcal{S}}_t) &\ge \mathbb{P}_t\left(\exists \pi \in \overline{\mathcal{S}}_t, f_t(\pi;\overline{\boldsymbol{\theta}}_t) + h_t(\pi;\widetilde{\mathcal{E}}_t) > \max_{\pi' \in \mathcal{S}_t} f_t(\pi';\overline{\boldsymbol{\theta}}_t) + h_t(\pi';\widetilde{\mathcal{E}}_t)\right) \\
&\ge \mathbb{P}_t\left(f_t(\pi_t^*;\overline{\boldsymbol{\theta}}_t) + h_t(\pi_t^*;\widetilde{\mathcal{E}}_t) > \max_{\pi' \in \mathcal{S}_t} f_t(\pi';\overline{\boldsymbol{\theta}}_t) + h_t(\pi';\widetilde{\mathcal{E}}_t)\right) \\
&\ge \mathbb{P}_t\left(f_t(\pi_t^*;\overline{\boldsymbol{\theta}}_t) + h_t(\pi_t^*;\widetilde{\mathcal{E}}_t) > f_t(\pi_t^*;\boldsymbol{\theta}^*), E_{2,t}\right) \\
&\ge \mathbb{P}_t\left(f_t(\pi_t^*;\overline{\boldsymbol{\theta}}_t) + h_t(\pi_t^*;\widetilde{\mathcal{E}}_t) > f_t(\pi_t^*;\overline{\boldsymbol{\theta}}_t) + c_1 \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i,\pi_t^*(i))\|_{\boldsymbol{V}_t^{-1}}\right) - \mathbb{P}(\overline{E}_{2,t}) \\
&= \mathbb{P}_t\left(h_t(\pi_t^*;\widetilde{\mathcal{E}}_t) \ge c_1 \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i,\pi_t^*(i))\|_{\boldsymbol{V}_t^{-1}}\right) - \mathbb{P}(\overline{E}_{2,t}),
\end{aligned}$$

where the second inequality follows from the fact that $\pi_t^* \in \overline{\mathcal{S}}_t$, the third inequality follows from $f_t(\pi;\overline{\boldsymbol{\theta}}_t) + h_t(\pi;\widetilde{\mathcal{E}}_t) \le f_t(\pi;\boldsymbol{\theta}^*) + (c_1 + c_2) \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i,\pi(i))\|_{\boldsymbol{V}_t^{-1}} \le f_t(\pi_t^*;\boldsymbol{\theta}^*)$ for any $\pi \in \mathcal{S}_t$ on $E_{1,t}$ and $E_{2,t}$, and the fourth inequality follows from the definition of $E_{1,t}$.

Finally, we can achieve the desired bound, using the inequality,

$$\begin{aligned}
\mathbb{E}_t[f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*)] &= \mathbb{E}_t[(f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*))\mathbb{1}[E_{2,t}]] \\
&\quad + \mathbb{E}_t[(f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*))\mathbb{1}[\overline{E}_{2,t}]] \\
&\le \mathbb{E}_t[(f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*))\mathbb{1}[E_{2,t}]] + KM\mathbb{P}_t(\overline{E}_{2,t}).
\end{aligned}$$

□

Remark D.7. We explain why the analysis of CAB-TS assumes access to an exact optimization oracle. For brevity, write $F_t(\pi) = f_t(\pi;\overline{\boldsymbol{\theta}}_t) + h_t(\pi;\widetilde{\mathcal{E}}_t)$. The proof of Lemma D.6 uses the bound

$$\mathbb{P}_t(\pi_t \in \overline{\mathcal{S}}_t) \ge \mathbb{P}_t\left(\exists \pi \in \overline{\mathcal{S}}_t, F_t(\pi) > \max_{\pi' \in \mathcal{S}_t} F_t(\pi')\right).$$

This step relies on exact optimality. Since $\pi_t^* \in \overline{\mathcal{S}}_t$, the set $\overline{\mathcal{S}}_t$ is nonempty. On the event on the right-hand side, an exact maximizer of F_t cannot lie in $\mathcal{S}_t$; otherwise, the allocation in $\overline{\mathcal{S}}_t$ would have

a strictly larger value, contradicting optimality. Hence the event implies $\pi_t \in \overline{\mathcal{S}}_t$. This implication is not preserved by an α-approximate optimization oracle with $\alpha < 1$. Exact maximization only requires the separation $\max_{\pi \in \overline{\mathcal{S}}_t} F_t(\pi) > \max_{\pi \in \mathcal{S}_t} F_t(\pi)$. However, an α-approximate optimization oracle may still return an allocation in $\mathcal{S}_t$ whenever $\max_{\pi \in \mathcal{S}_t} F_t(\pi) \geq \alpha \max_{\pi \in \overline{\mathcal{S}}_t} F_t(\pi)$, because such an allocation satisfies the approximation guarantee. Thus, since the proof only ensures that a good allocation can attain the largest objective value, and does not provide the stronger margin $\max_{\pi \in \mathcal{S}_t} F_t(\pi) < \alpha \max_{\pi \in \overline{\mathcal{S}}_t} F_t(\pi)$, an α-approximate optimization oracle with $\alpha < 1$ may still return an allocation in $\mathcal{S}_t$, so the above probability lower bound is unavailable.

p_2 and p_3 in Lemma D.6 can be bounded as follows, respectively.

Lemma D.8. *For each $t \geq 1$, $E_{2,t}$ holds with probability at least $1 - 2\delta$.*

Lemma D.9. *$E_{3,t}$ holds with probability at least 0.15.*

We prove these lemmas.

Proof of Lemma D.8. From Kveton et al. [24, Lemma 4], $E'_{2,t} = \{^\forall (i,a) \in [N] \times [K] \mid |\boldsymbol{\phi}_t(i,a)^\top \widetilde{\boldsymbol{\varepsilon}}_t(i)| \leq c_2 \|\boldsymbol{\phi}_t(i,a)\|_{\boldsymbol{V}_t^{-1}}\}$ holds with probability at least $1 - 2\delta$. Here, we use the fact that $\boldsymbol{H}_t \succeq \kappa_\mu \boldsymbol{V}_t$ due to the definition of $\boldsymbol{H}_t$ and $\boldsymbol{V}_t$. Their proof uses the union bound, so we can apply that lemma to our algorithm, which samples $\widetilde{\boldsymbol{\varepsilon}}_t(i)$ independently for each $i \in [N]$. On $E'_{2,t}$, we have $\sum_{i=1}^N |\boldsymbol{\phi}_t(i, \pi(i))^\top \widetilde{\boldsymbol{\varepsilon}}_t(i)| \leq c_2 \sum_{i=1}^N \|\boldsymbol{\phi}_t(i, \pi(i))\|_{\boldsymbol{V}_t^{-1}}$ for any $\pi \in \Pi$. In addition, we have $|h_t(\pi; \widetilde{\mathcal{E}}_t)| \leq \sum_{i=1}^N |\boldsymbol{\phi}_t(i, \pi(i))^\top \widetilde{\boldsymbol{\varepsilon}}_t(i)|$ from the triangle inequality. Therefore, on $E'_{2,t}$, it holds that $|h_t(\pi; \widetilde{\mathcal{E}}_t)| \leq c_2 \sum_{i=1}^N \|\boldsymbol{\phi}_t(i, \pi(i))\|_{\boldsymbol{V}_t^{-1}}$ for any $\pi \in \Pi$. In other words, $E_{2,t}$ holds with probability at least $1 - 2\delta$. □

Proof of Lemma D.9. Using $\boldsymbol{H}_t \preceq L_\mu \boldsymbol{V}_t$, and Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\mathbb{P}_t\left(h_t(\pi_t^*; \widetilde{\mathcal{E}}_t) \geq c_1 \sum_{i=1}^N \|\boldsymbol{\phi}_t(i, \pi_t^*(i))\|_{\boldsymbol{V}_t^{-1}}\right) \\
&\geq \mathbb{P}_t\left(h_t(\pi_t^*; \widetilde{\mathcal{E}}_t) \geq c_1 \sqrt{L_\mu N \sum_{i=1}^N \|\boldsymbol{\phi}_t(i, \pi_t^*(i))\|^2_{\boldsymbol{H}_t^{-1}}}\right) \\
&= \mathbb{P}_t\left(\sum_{i=1}^N \boldsymbol{\phi}_t(i, \pi_t^*(i))^\top \widetilde{\boldsymbol{\varepsilon}}_t(i) \geq a \sqrt{\sum_{i=1}^N \|\boldsymbol{\phi}_t(i, \pi_t^*(i))\|^2_{\boldsymbol{H}_t^{-1}}}\right) \qquad (14)
\end{aligned}$$

Next, we derive a lower bound. From $\widetilde{\boldsymbol{\varepsilon}}_t(i) \sim \mathcal{N}(\boldsymbol{0}, a^2 \boldsymbol{H}_t^{-1})$, we have $\boldsymbol{\phi}_t(i, \pi_t^*(i))^\top \widetilde{\boldsymbol{\varepsilon}}_t(i) \sim \mathcal{N}(0, a^2 \|\boldsymbol{\phi}_t(i, \pi_t^*(i))\|^2_{\boldsymbol{H}_t^{-1}})$. Moreover, since $\{\widetilde{\boldsymbol{\varepsilon}}_t(i)\}_{i=1}^N$ are independent, it follows that $\sum_{i=1}^N \boldsymbol{\phi}_t(i, \pi_t^*(i))^\top \widetilde{\boldsymbol{\varepsilon}}_t(i) \sim \mathcal{N}(0, a^2 \sum_{i=1}^N \|\boldsymbol{\phi}_t(i, \pi_t^*(i))\|^2_{\boldsymbol{H}_t^{-1}})$. Consequently, we obtain

$$\mathbb{P}_t\left(\sum_{i=1}^N \boldsymbol{\phi}_t(i, \pi_t^*(i))^\top \widetilde{\boldsymbol{\varepsilon}}_t(i) \geq a \sqrt{\sum_{i=1}^N \|\boldsymbol{\phi}_t(i, \pi_t^*(i))\|^2_{\boldsymbol{H}_t^{-1}}}\right) \geq 0.15. \qquad (15)$$

Combining (14) with (15) yields the desired bound. □

We are now ready to prove Theorem D.5.

Proof of Theorem D.5. Let $E_{4,t} = \{\|\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_2 \leq D + 1\}$, $\mathbb{P}(E_{4,t}) \geq 1 - p_4$, $p_1 \geq \mathbb{P}(\overline{E}_{1,t} \wedge E_{4,t})$, $\mathbb{P}_t(E_{2,t}) \geq 1 - p_2$ on the event $E_{4,t}$, and $\mathbb{P}_t(E_{3,t}) \geq p_3$.

$$\mathbb{E}[\mathcal{R}_T] = \sum_{t=1}^T \mathbb{E}[f_t(\pi_t^*; \boldsymbol{\theta}^*) - f_t(\pi_t; \boldsymbol{\theta}^*)]$$

$$\leq \sum_{t=1}^{T} \mathbb{E}[(f_t(\pi_t^*; \boldsymbol{\theta}^*) - f_t(\pi_t; \boldsymbol{\theta}^*))\mathbb{1}[E_{1,t}, E_{4,t}]] + (p_1 + p_4)KMT$$
$$\leq \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}_t[(f_t(\pi_t^*; \boldsymbol{\theta}^*) - f_t(\pi_t; \boldsymbol{\theta}^*))\mathbb{1}[E_{1,t}, E_{4,t}]]] + (p_1 + p_4)KMT$$

From Lemma D.6, it holds that

$$\mathbb{E}[\mathcal{R}_T] \leq (c_1 + c_2)\left(1 + \frac{2}{p_3 - p_2}\right)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}}\right] + (p_1 + p_2 + p_4)KMT$$
$$\leq (c_1 + c_2)\left(1 + \frac{2}{p_3 - p_2}\right)\mathbb{E}\left[\sqrt{NT\sum_{t=1}^{T}\sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|^2_{\boldsymbol{V}_t^{-1}}}\right] + (p_1 + p_2 + p_4)KMT,$$

where we used the Cauchy–Schwarz inequality in the last inequality.

Next, we bound p_1, p_2, p_3, and p_4. From Lemmas D.1, D.2, D.8 and D.9, we have $p_1 \leq \delta$, $p_2 \leq 2\delta$, $p_3 \geq 0.15$, and $p_4 \leq \delta$. In addition, from the definition of λ_0, we have $p_4 \leq \delta$ from Kveton et al. [24, Lemma 9]. Therefore, using these bounds and (8), we can achieve the desired bound. □

Remark D.10. For each i, sampling i.i.d. from a Gaussian distribution is used to obtain a lower bound for (14). Indeed, $\sum_{i=1}^{N} \boldsymbol{\phi}_t(i, \pi_t^*(i))^\top \widetilde{\boldsymbol{\varepsilon}}_t(i) \sim \mathcal{N}(0, a^2 \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i, \pi_t^*(i))\|^2_{\boldsymbol{H}_t^{-1}})$, which leads to the bound in (15), is derived from the independence. On the other hand, if we sample a single $\overline{\boldsymbol{\varepsilon}}_t$ from $\mathcal{N}(0, a^2 \sum_{i=1}^{N} \|\boldsymbol{\phi}_t(i, \pi_t^*(i))\|^2_{\boldsymbol{H}_t^{-1}})$ and set $\widetilde{\boldsymbol{\varepsilon}}_t(i) = \overline{\boldsymbol{\varepsilon}}_t$ for any $i \in [N]$, then $\sum_{i=1}^{N} \boldsymbol{\phi}_t(i, \pi_t^*(i))^\top \widetilde{\boldsymbol{\varepsilon}}_t(i) \sim \mathcal{N}(0, a^2 \|\sum_{i=1}^{N} \boldsymbol{\phi}_t(i, \pi_t^*(i))\|^2_{\boldsymbol{H}_t^{-1}})$, and this prevents us from obtaining a desirable probability bound.

D.4 Variant of Algorithm 2

Algorithm 3 CAB-TS (heuristic variant)

Input: The total rounds T, the number of users N, tuning parameter λ_0 and a, and access to a practical allocation routine.

$\mathcal{D}_1 = \varnothing$, $\boldsymbol{V}_1 = \lambda_0 \boldsymbol{I}$, and $\boldsymbol{H}_1 = L_\mu \lambda_0 \boldsymbol{I}$.
for $t = 1, \ldots, T$ **do**
  $\boldsymbol{V}_t \leftarrow \lambda_0 \boldsymbol{I} + \sum_{s=1}^{t-1} \sum_{i=1}^{N} \boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^\top$.
  $\overline{\boldsymbol{\theta}}_t \leftarrow \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}(\mathcal{D}_t; \boldsymbol{\theta}, \kappa_\mu \lambda_0)$.
  If $t \geq 2$, $\boldsymbol{H}_t \leftarrow \sum_{s=1}^{t-1} \sum_{i=1}^{N} \dot{\mu}(\boldsymbol{x}_s(i)^\top \overline{\boldsymbol{\theta}}_t)(\boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^\top + \frac{\lambda_0}{N(t-1)} \boldsymbol{I})$.
  for $i = 1, \ldots, N$ **do**
    $\widetilde{\boldsymbol{\theta}}_t(i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\overline{\boldsymbol{\theta}}_t, a^2 {\boldsymbol{H}_t}^{-1})$.
  Call the practical allocation routine for $\widetilde{f}_t(\pi; \widetilde{\Theta}_t)$ and let π_t denote its output.
  Observe $y_t(i)$ for any $i \in [N]$.
  Let $\boldsymbol{x}_t(i) \leftarrow \boldsymbol{\phi}_t(i, \pi_t(i))$ for any $i \in [N]$ and $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(\boldsymbol{x}_t(i), y_t(i))\}_{i \in [N]}$.

In this section, we summarize a heuristic variant of CAB-TS related to Algorithm 2. We retain it only for the experiments and do not include it in the theoretical contribution.

In Algorithm 2, we sample $\widetilde{\boldsymbol{\varepsilon}}_t(i)$ from $\mathcal{N}(\boldsymbol{0}, a^2 {\boldsymbol{H}_t}^{-1})$ for any $i \in [N]$, and maximize $f_t(\pi; \overline{\boldsymbol{\theta}}_t) + h_t(\pi; \widetilde{\mathcal{E}}_t)$. However, it is also a natural idea to sample $\widetilde{\boldsymbol{\theta}}_t(i)$ from $\mathcal{N}(\overline{\boldsymbol{\theta}}_t, a^2 {\boldsymbol{H}_t}^{-1})$ and instead maximize $\widetilde{f}_t$, defined as follows:

$$\widetilde{f}_t(\pi; \vartheta) = \sum_{a \in [K]} r\left(\sum_{i \in \pi^{-1}(a)} \mu(\boldsymbol{\phi}_t(i, a)^\top \vartheta(i))\right). \tag{16}$$

The algorithm can be written as Algorithm 3.

Remark D.11. The heuristic variant Algorithm 3 is retained only for empirical comparison and is not supported by a theoretical guarantee in this paper. To extend the proof of Theorem D.5 by a similar argument, we would need a positive lower bound on the probability of the event

$$E_{3,t}^{\star} = \left\{ \widetilde{f}_t(\pi_t^*; \widetilde{\Theta}_t) - f_t(\pi_t^*; \overline{\boldsymbol{\theta}}_t) \geq c_1^{\star} \sum_{i \in [N]} \|\boldsymbol{\phi}_t(i, \pi_t^*(i))\|_{\boldsymbol{V}_t^{-1}} \right\}.$$

However, for the direct nonlinear objective, the concavity of r alone does not yield a useful lower bound on $\widetilde{f}_t(\pi_t^*; \widetilde{\Theta}_t) - f_t(\pi_t^*; \overline{\boldsymbol{\theta}}_t)$, and thus obtaining such a probabilistic lower bound is difficult. Even if one imposes a uniform lower-bound assumption on $\dot{r}$, analogous to the assumption on $\dot{\mu}$, this issue remains unresolved.

E Omitted details of Section 4.3

In this section, we provide the omitted details of Section 4.3. For completeness, we first restate CAB-OFU with one-pass OMD update and the quantities used in the analysis, since the main text only presents a compressed description.

E.1 Complete description of CAB-OFU with one-pass OMD update

We consider the following discussions under Assumption 4.3. The one-pass variant is summarized in Algorithm 4.

Algorithm 4 CAB-OFU with one-pass OMD update

Input: The total rounds T, the number of users N, and tuning parameters λ_{op}, η, α, and δ.

1: Initialize $\boldsymbol{\theta}_1 \in \Theta$ and $\boldsymbol{Q}_1 \leftarrow \lambda_{\mathrm{op}} \boldsymbol{I}$.
2: **for** $t = 1, \ldots, T$ **do**
3: Call an α-approximate OFU oracle satisfying (18) and obtain π_t.
4: Observe $y_t(i)$ for all $i \in [N]$.
5: Set $\boldsymbol{x}_t(i) \leftarrow \boldsymbol{\phi}_t(i, \pi_t(i))$ for all $i \in [N]$.
6: Update $\boldsymbol{\theta}_{t+1}$ by (17).

For each round t and user i, define the negative log-likelihood

$$\ell_{t,i}(\boldsymbol{\theta}) = -y_t(i)\boldsymbol{x}_t(i)^{\top}\boldsymbol{\theta} + m(\boldsymbol{x}_t(i)^{\top}\boldsymbol{\theta}),$$

where $\dot{m} = \mu$. The one-pass OMD update uses the quadratic surrogate

$$\widetilde{\ell}_t(\boldsymbol{\theta}) = \sum_{i=1}^{N} (\langle \nabla_{\boldsymbol{\theta}} \ell_{t,i}(\boldsymbol{\theta}_t), \boldsymbol{\theta} - \boldsymbol{\theta}_t \rangle + \frac{1}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|^2_{\nabla^2_{\boldsymbol{\theta}} \ell_{t,i}(\boldsymbol{\theta}_t)}).$$

The parameter is then updated by

$$\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta} \in \Theta} \left(\widetilde{\ell}_t(\boldsymbol{\theta}) + \frac{1}{2\eta} \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|^2_{\boldsymbol{Q}_t} \right), \tag{17}$$

and the matrix sequence is

$$\boldsymbol{Q}_{t+1} = \lambda_{\mathrm{op}} \boldsymbol{I} + \sum_{s=1}^{t} \sum_{i=1}^{N} \nabla^2_{\boldsymbol{\theta}} \ell_{s,i}(\boldsymbol{\theta}_{s+1}) = \lambda_{\mathrm{op}} \boldsymbol{I} + \sum_{s=1}^{t} \sum_{i=1}^{N} \dot{\mu}(\boldsymbol{x}_s(i)^{\top}\boldsymbol{\theta}_{s+1}) \boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^{\top}.$$

We use the confidence set

$$C_t(\delta) = \{\boldsymbol{\theta} \in \Theta \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|_{\boldsymbol{Q}_t} \leq \beta_t(\delta)\},$$

where

$$\beta_t(\delta) = \sqrt{4\lambda_{\mathrm{op}} D^2 + 2\eta \log(1/\delta) + d(6\eta^2 + \eta) \log(1 + L_{\mu} N t / \lambda_{\mathrm{op}})}.$$

Given the optimistic value

$$\mathrm{OPT}_t^{\mathrm{op}} = \max_{\pi\in\Pi} \max_{\boldsymbol{\theta}\in C_t(\delta)} f_t(\pi;\boldsymbol{\theta}),$$

the allocation is chosen by an α-approximate OFU rule:

$$\max_{\boldsymbol{\theta}\in C_t(\delta)} f_t(\pi_t;\boldsymbol{\theta}) \geq \alpha \mathrm{OPT}_t^{\mathrm{op}}. \tag{18}$$

Next, we compare the cost of updating the parameter-update step. If the regularized MLE at round t is solved by an iterative method with I_t optimization iterations, and one pass over the first $t-1$ rounds costs $O(t)$ when suppressing the dependence on d and N, then the MLE-based update costs $O(tI_t)$. In contrast, once $\boldsymbol{Q}_t$ is maintained incrementally, the one-pass OMD update (17) uses only the current-round surrogate and $\boldsymbol{Q}_t$. Hence, if the quadratic subproblem is solved in $\widetilde{I}_t$ optimization iterations, its update cost is $O(\widetilde{I}_t)$. The optimistic allocation step is separate and, in the practical implementation, may require multiple calls to the submodular welfare oracle.

E.2 Proof of Theorem 4.4

We provide the complete version of Theorem 4.4 and its proof.

Theorem E.1 (complete version of Theorem 4.4)**.** *Under Assumption 4.3, with probability at least* $1-\delta$*, the* α*-approximate regret of Algorithm 4 with* $\eta = 1+RD$ *and* $\lambda_{\mathrm{op}} \geq \max\{14d\eta R^2, 6\eta RDL_\mu N\}$ *is upper bounded as*

$$\begin{aligned}\mathcal{R}_T^{\alpha} \leq{}& 2\kappa_\mu^{-1/2} L_r L_\mu \beta_T(\delta)\left(\sqrt{2dNT\log\left(1+\frac{NT}{d\lambda_{\mathrm{op}}}\right)} + \frac{2dN}{\sqrt{\lambda_{\mathrm{op}}}}\log\left(1+\frac{NT}{d\lambda_{\mathrm{op}}}\right)\right)\\ &+ 16\kappa_\mu^{-1} RL_\mu L_r dN\beta_T^2(\delta)\log\left(1+\frac{NT}{d\lambda_{\mathrm{op}}}\right)\end{aligned}$$

Proof. By decreasing κ_μ if necessary, we may assume $\kappa_\mu \leq 1$. Let $(\pi_t, \widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}})$ be an output of the α-approximate optimization oracle for $\max_{\pi\in\Pi}\max_{\boldsymbol{\theta}\in C_t(\delta)} f_t(\pi;\boldsymbol{\theta})$. We work on the high-probability event in Lemma E.2, so that $\|\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}} - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_t} \leq 2\beta_t(\delta)$ for all $t\in[T]$. Let

$$\boldsymbol{V}_t = \lambda_{\mathrm{op}}\boldsymbol{I} + \sum_{s=1}^{t-1}\sum_{i=1}^{N} \boldsymbol{x}_s(i)\boldsymbol{x}_s(i)^\top.$$

By the lower bound encoded in κ_μ, we have

$$\boldsymbol{Q}_t = \lambda_{\mathrm{op}}\boldsymbol{I} + \sum_{s=1}^{t-1}\sum_{i=1}^{N} \dot{\mu}\big(\boldsymbol{x}_s(i)^\top\boldsymbol{\theta}_{s+1}\big)\boldsymbol{x}_s(i)\boldsymbol{x}_s(i)^\top \succeq \kappa_\mu \boldsymbol{V}_t.$$

We first decompose the instantaneous regret and then bound the linear and quadratic terms from the Taylor expansion separately. We have

$$\begin{aligned}\mathcal{R}_T^{\alpha} &= \sum_{t=1}^{T}(\alpha f_t(\pi_t^*;\boldsymbol{\theta}^*) - f_t(\pi_t;\boldsymbol{\theta}^*))\\ &\leq \sum_{t=1}^{T}\Big(f_t(\pi_t;\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}}) - f_t(\pi_t;\boldsymbol{\theta}^*)\Big) \leq L_r\sum_{t=1}^{T}\sum_{i=1}^{N}\Big|\mu(\boldsymbol{x}_t(i)^\top\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}}) - \mu(\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^*)\Big|\\ &\leq L_r\sum_{t=1}^{T}\sum_{i=1}^{N}\Bigg|\dot{\mu}(\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^*)\boldsymbol{x}_t(i)^\top(\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}} - \boldsymbol{\theta}^*)\\ &\qquad + \Big(\boldsymbol{x}_t(i)^\top\Big(\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}} - \boldsymbol{\theta}^*\Big)\Big)^2\int_0^1(1-v)\ddot{\mu}\Big(\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^* + v\boldsymbol{x}_t(i)^\top\Big(\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}} - \boldsymbol{\theta}^*\Big)\Big)\,dv\Bigg|\\ &\leq L_r\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\dot{\mu}(\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^*)\Big|\boldsymbol{x}_t(i)^\top(\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}} - \boldsymbol{\theta}^*)\Big| + RL_\mu\Big|\Big(\boldsymbol{x}_t(i)^\top\Big(\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}} - \boldsymbol{\theta}^*\Big)\Big)^2\Big|\right).\end{aligned} \tag{19}$$

Here the first inequality follows from the α-approximate OFU rule, since $\boldsymbol{\theta}^* \in C_t(\delta)$ implies

$$f_t(\pi_t; \widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}}) \geq \alpha \max_{\pi \in \Pi} \max_{\boldsymbol{\theta} \in C_t(\delta)} f_t(\pi; \boldsymbol{\theta}) \geq \alpha f_t(\pi_t^*; \boldsymbol{\theta}^*),$$

the second inequality follows from the Lipschitz continuity of r, the third inequality follows from the second-order Taylor expansion of μ around $\boldsymbol{\theta}^*$ with the integral remainder form, and the fourth inequality follows from Assumption 4.3.

We next control the first-order term in (19). By Cauchy–Schwarz and Lemma E.2,

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{i=1}^{N} \dot{\mu}(\boldsymbol{x}_t(i)^\top \boldsymbol{\theta}^*) \left| \boldsymbol{x}_t(i)^\top (\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}} - \boldsymbol{\theta}^*) \right| \\
&\leq \sum_{t=1}^{T} \sum_{i=1}^{N} \dot{\mu}(\boldsymbol{x}_t(i)^\top \boldsymbol{\theta}^*) \|\boldsymbol{x}_t(i)\|_{\boldsymbol{Q}_t^{-1}} \|\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}} - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_t} \leq 2\beta_T(\delta) \sum_{t=1}^{T} \sum_{i=1}^{N} \dot{\mu}(\boldsymbol{x}_t(i)^\top \boldsymbol{\theta}^*) \|\boldsymbol{x}_t(i)\|_{\boldsymbol{Q}_t^{-1}}.
\end{aligned} \tag{20}$$

Using $\dot{\mu}(\boldsymbol{x}_t(i)^\top \boldsymbol{\theta}^*) \leq L_\mu$ and $\boldsymbol{Q}_t \succeq \kappa_\mu \boldsymbol{V}_t$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{i=1}^{N} \dot{\mu}(\boldsymbol{x}_t(i)^\top \boldsymbol{\theta}^*) |\boldsymbol{x}_t(i)^\top (\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}} - \boldsymbol{\theta}^*)| \\
&\leq 2L_\mu \beta_T(\delta) \sum_{t=1}^{T} \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{Q}_t^{-1}} \\
&\leq 2\kappa_\mu^{-1/2} L_\mu \beta_T(\delta) \sum_{t=1}^{T} \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|_{\boldsymbol{V}_t^{-1}} \\
&\leq 2\kappa_\mu^{-1/2} L_\mu \beta_T(\delta) \left(\sqrt{2dNT \log\left(1 + \frac{NT}{d\lambda_{\mathrm{op}}}\right)} + \frac{2dN}{\sqrt{\lambda_{\mathrm{op}}}} \log\left(1 + \frac{NT}{d\lambda_{\mathrm{op}}}\right) \right),
\end{aligned} \tag{21}$$

where the last inequality follows from (8). By applying Takemura et al. [43, Lemma 5], we obtain

$$\sum_{t=1}^{T} \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|^2_{\boldsymbol{V}_t^{-1}} \leq 4dN \log\left(1 + \frac{NT}{d\lambda_{\mathrm{op}}}\right). \tag{22}$$

It remains to control the second-order remainder term in (19). By the self-concordance-type bound and Lemma E.2,

$$\begin{aligned}
RL_\mu \sum_{t=1}^{T} \sum_{i=1}^{N} \left| \left(\boldsymbol{x}_t(i)^\top \left(\widetilde{\boldsymbol{\theta}}_t^{\mathrm{op}} - \boldsymbol{\theta}^* \right) \right)^2 \right| &\leq 4RL_\mu \beta_T^2(\delta) \sum_{t=1}^{T} \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|^2_{\boldsymbol{Q}_t^{-1}} \\
&\leq 4\kappa_\mu^{-1} RL_\mu \beta_T^2(\delta) \sum_{t=1}^{T} \sum_{i=1}^{N} \|\boldsymbol{x}_t(i)\|^2_{\boldsymbol{V}_t^{-1}} \\
&\leq 16\kappa_\mu^{-1} RL_\mu dN \beta_T^2(\delta) \log\left(1 + \frac{NT}{d\lambda_{\mathrm{op}}}\right),
\end{aligned} \tag{23}$$

where the last inequality follows from (22).

Combining (21) and (23), we obtain the following regret bound:

$$\begin{aligned}
\mathcal{R}_T^\alpha \leq\ & 2\kappa_\mu^{-1/2} L_r L_\mu \beta_T(\delta) \left(\sqrt{2dNT \log\left(1 + \frac{NT}{d\lambda_{\mathrm{op}}}\right)} + \frac{2dN}{\sqrt{\lambda_{\mathrm{op}}}} \log\left(1 + \frac{NT}{d\lambda_{\mathrm{op}}}\right) \right) \\
&+ 16\kappa_\mu^{-1} RL_\mu L_r dN \beta_T^2(\delta) \log\left(1 + \frac{NT}{d\lambda_{\mathrm{op}}}\right)
\end{aligned} \tag{24}$$

□

E.3 Useful lemmas for the proof of Theorem E.1

In this section, we present the lemmas to prove Theorem E.1.

First, we prove that the confidence set defined in Lemma E.2 contains the true parameter $\boldsymbol{\theta}^*$ with high probability.

Lemma E.2. *Let $\delta \in (0,1)$ and*

$$C_t(\delta) = \{\boldsymbol{\theta} \in \Theta \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|_{\boldsymbol{Q}_t} \leq \beta_t(\delta)\},$$

where

$$\beta_t(\delta) = \sqrt{4\lambda_{\mathrm{op}} D^2 + 2\eta \log\left(\frac{1}{\delta}\right) + d(6\eta^2 + \eta)\log\left(1 + \frac{L_\mu N t}{\lambda_{\mathrm{op}}}\right)}.$$

Set $\eta = 1 + RD$ and $\lambda_{\mathrm{op}} \geq \max\{14 d\eta R^2, 6\eta RDL_\mu N\}$ for Algorithm 4. Then, with probability at least $1-\delta$, we have $\boldsymbol{\theta}^* \in C_t(\delta)$, equivalently $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_t} \leq \beta_t(\delta)$, for any $t \in [T]$.

Proof. By Lemma E.4, we have

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|^2_{\boldsymbol{Q}_{t+1}} \leq 2\eta \sum_{s=1}^{t}\sum_{i=1}^{N}\big(\ell_{s,i}(\boldsymbol{\theta}^*) - \ell_{s,i}(\boldsymbol{\theta}_{s+1})\big) + 4\lambda_{\mathrm{op}} D^2 \\ + 2\eta RDL_\mu N \sum_{s=1}^{t}\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_2^2 - \sum_{s=1}^{t}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}\|^2_{\boldsymbol{Q}_s}. \end{aligned} \tag{25}$$

First, to upper bound the linear term, we decompose as

$$\sum_{s=1}^{t}\sum_{i=1}^{N}(\ell_{s,i}(\boldsymbol{\theta}^*) - \ell_{s,i}(\boldsymbol{\theta}_{s+1})) = \sum_{s=1}^{t}\left(\sum_{i=1}^{N}\ell_{s,i}(\boldsymbol{\theta}^*) - m_s(P_s)\right) + \sum_{s=1}^{t}\left(m_s(P_s) - \sum_{i=1}^{N}\ell_{s,i}(\boldsymbol{\theta}_{s+1})\right),$$

where $P_s = \mathcal{N}(\boldsymbol{\theta}_s, \zeta \boldsymbol{Q}_s^{-1})$ with $\zeta = 3\eta/2$ is a d-dimensional multivariate normal distribution and the function $m_s \colon P \mapsto \mathbb{R}$ is defined as $m_s(P_s) = -\log(\mathbb{E}_{\boldsymbol{\theta}\sim P_s}[\exp(-\sum_{i=1}^{N}\ell_{s,i}(\boldsymbol{\theta}))])$.

The first term can be bounded by applying Lemma E.5 to the aggregated loss $\sum_{i=1}^{N}\ell_{s,i}(\boldsymbol{\theta}^*)$. Thus, it holds that with probability at least $1-\delta$,

$$\sum_{s=1}^{t}\left(\sum_{i=1}^{N}\ell_{s,i}(\boldsymbol{\theta}^*) - m_s(P_s)\right) \leq \log\left(\frac{1}{\delta}\right).$$

In addition, by Zhang et al. [49, Lemma 6], we can bound the second term as

$$\sum_{s=1}^{t}\left(m_s(P_s) - \sum_{i=1}^{N}\ell_{s,i}(\boldsymbol{\theta}_{s+1})\right) \leq \frac{1}{3\eta}\sum_{s=1}^{t}\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|^2_{\boldsymbol{Q}_s} + d\left(3\eta + \frac{1}{2}\right)\log\left(1 + \frac{L_\mu N t}{\lambda_{\mathrm{op}}}\right).$$

Although there is a difference between Zhang et al. [49, Lemma 6] and our setting in that the latter is combinatorial, we can derive the above upper bound by treating $\sum_{i=1}^{N}\ell_{s,i}(\boldsymbol{\theta}_{s+1})$ as a single unit. Substituting these bounds into the (25), it holds that

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|^2_{\boldsymbol{Q}_{t+1}} &\leq 4\lambda_{\mathrm{op}} D^2 + 2\eta\log\left(\frac{1}{\delta}\right) + d(6\eta^2 + \eta)\log\left(1 + \frac{L_\mu N t}{\lambda_{\mathrm{op}}}\right) \\ &\quad + 2\eta RDL_\mu N\sum_{s=1}^{t}\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_2^2 - \frac{1}{3\eta}\sum_{s=1}^{t}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}\|^2_{\boldsymbol{Q}_s} \\ &\leq 4\lambda_{\mathrm{op}} D^2 + 2\eta\log\left(\frac{1}{\delta}\right) + d(6\eta^2 + \eta)\log\left(1 + \frac{L_\mu N t}{\lambda_{\mathrm{op}}}\right) \\ &\quad + 2\eta RDL_\mu N\sum_{s=1}^{t}\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_2^2 - \frac{1}{3}\sum_{s=1}^{t}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}\|^2_{\boldsymbol{Q}_s} \\ &\leq 4\lambda_{\mathrm{op}} D^2 + 2\eta\log\left(\frac{1}{\delta}\right) + d(6\eta^2 + \eta)\log\left(1 + \frac{L_\mu N t}{\lambda_{\mathrm{op}}}\right),\end{aligned}$$

where the last inequality follows from $\lambda_{\mathrm{op}} \geq 6\eta RDL_\mu N$. □

The following lemma is regarding the property of the online mirror descent update.

Lemma E.3 (Zhang et al. 49, Lemma 1)**.** *Let $f\colon \Theta \to \mathbb{R}$ be a convex function, let Θ be a convex set, and let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be a positive definite matrix. $\boldsymbol{\theta}_+ = \arg\min_{\boldsymbol{\theta}\in\Theta}(f(\boldsymbol{\theta}) + \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\boldsymbol{A}}^2)$ satisfies*

$$\|\boldsymbol{\theta}_+ - \boldsymbol{u}\|_{\boldsymbol{A}}^2 \le 2\eta\langle \nabla f(\boldsymbol{\theta}_+), \boldsymbol{u} - \boldsymbol{\theta}_+\rangle + \|\boldsymbol{\theta}_0 - \boldsymbol{u}\|_{\boldsymbol{A}}^2 - \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_+\|_{\boldsymbol{A}}^2$$

for all $\boldsymbol{u} \in \Theta$.

This lemma is used to control the distance between the updated parameter and the true parameter.

Lemma E.4. *Assume that $\|\boldsymbol{\theta}^*\|_2 \le D$, $\Theta \subseteq \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \le D\}$, $\|\boldsymbol{x}_t(i)\|_2 \le 1$ for all t and i, $\dot{\mu}(z) \le L_\mu$ for all z, and $\ddot{\mu}(z) \le R$ for all z. When we use Algorithm 4 with $\eta = 1 + DR$, for any $\lambda_{\mathrm{op}} > 0$,*

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_{t+1}}^2 \le 2\eta \sum_{s=1}^{t}\sum_{i=1}^{N}\big(\ell_{s,i}(\boldsymbol{\theta}^*) - \ell_{s,i}(\boldsymbol{\theta}_{s+1})\big) + 4\lambda_{\mathrm{op}}D^2 \\ + 2\eta RDL_\mu N\sum_{s=1}^{t}\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_2^2 - \sum_{s=1}^{t}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}\|_{\boldsymbol{Q}_s}^2.\end{aligned}$$

Proof. Fix $s \in [t]$. By the second-order Taylor expansion of $\ell_{s,i}$ around $\boldsymbol{\theta}_{s+1}$,

$$\ell_{s,i}(\boldsymbol{\theta}^*) = \ell_{s,i}(\boldsymbol{\theta}_{s+1}) + \langle \nabla\ell_{s,i}(\boldsymbol{\theta}_{s+1}), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{s+1}\rangle + \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^*\|_{\widetilde{\boldsymbol{H}}_{s,i}}^2,$$

where

$$\widetilde{\boldsymbol{H}}_{s,i} = \int_0^1 (1-v)\nabla^2\ell_{s,i}((1-v)\boldsymbol{\theta}_{s+1} + v\boldsymbol{\theta}^*)\,dv.$$

Rearranging gives

$$\ell_{s,i}(\boldsymbol{\theta}_{s+1}) - \ell_{s,i}(\boldsymbol{\theta}^*) \le \langle \nabla\ell_{s,i}(\boldsymbol{\theta}_{s+1}), \boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^*\rangle - \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^*\|_{\widetilde{\boldsymbol{H}}_{s,i}}^2.$$

Next, let $\widetilde{\boldsymbol{H}}_{s,i} = \widetilde{\alpha}_{s,i}\boldsymbol{x}_s(i)\boldsymbol{x}_s(i)^\top$, and $\widetilde{\alpha}_{s,i} = \int_0^1 (1-v)\dot{\mu}\big(\boldsymbol{x}_s(i)^\top\big((1-v)\boldsymbol{\theta}_{s+1} + v\boldsymbol{\theta}^*\big)\big)dv$. By Zhang et al. [49, Lemma 8], we have

$$\widetilde{\alpha}_{s,i} \ge \frac{\dot{\mu}\big(\boldsymbol{x}_s(i)^\top\boldsymbol{\theta}_{s+1}\big)}{2 + 2DR}.$$

Therefore,

$$\widetilde{\boldsymbol{H}}_{s,i} \succeq \frac{1}{2+2DR}\nabla^2\ell_{s,i}(\boldsymbol{\theta}_{s+1}).$$

Summing over i yields

$$\begin{aligned}\sum_{i=1}^{N}\big(\ell_{s,i}(\boldsymbol{\theta}_{s+1}) - \ell_{s,i}(\boldsymbol{\theta}^*)\big) \le \sum_{i=1}^{N}\langle \nabla\ell_{s,i}(\boldsymbol{\theta}_{s+1}), \boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^*\rangle \\ - \frac{1}{2+2DR}\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^*\|_{\sum_{i=1}^{N}\nabla^2\ell_{s,i}(\boldsymbol{\theta}_{s+1})}^2. \qquad (26)\end{aligned}$$

We now control the linear term. Applying Lemma E.3 with $f = \widetilde{\ell}_s$, $\boldsymbol{A} = \boldsymbol{Q}_s$, $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_s$, $\boldsymbol{\theta}_+ = \boldsymbol{\theta}_{s+1}$, and $\boldsymbol{u} = \boldsymbol{\theta}^*$, we obtain

$$\langle \nabla\widetilde{\ell}_s(\boldsymbol{\theta}_{s+1}), \boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^*\rangle \le \frac{1}{2\eta}\big(\|\boldsymbol{\theta}_s - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_s}^2 - \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_s}^2 - \|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}\|_{\boldsymbol{Q}_s}^2\big). \qquad (27)$$

Let $\Delta_s = \boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s$. Since $\nabla\widetilde{\ell}_s(\boldsymbol{\theta}_{s+1}) = \sum_{i=1}^{N}\big(\nabla\ell_{s,i}(\boldsymbol{\theta}_s) + \nabla^2\ell_{s,i}(\boldsymbol{\theta}_s)\Delta_s\big)$, the difference between the true and surrogate gradients is

$$\sum_{i=1}^{N}\nabla\ell_{s,i}(\boldsymbol{\theta}_{s+1}) - \nabla\widetilde{\ell}_s(\boldsymbol{\theta}_{s+1}) = \sum_{i=1}^{N}\big(\nabla\ell_{s,i}(\boldsymbol{\theta}_{s+1}) - \nabla\ell_{s,i}(\boldsymbol{\theta}_s) - \nabla^2\ell_{s,i}(\boldsymbol{\theta}_s)\Delta_s\big).$$

By Taylor's theorem, for each i there exists $\xi_{s,i}$ on the line segment between $\boldsymbol{x}_s(i)^\top \boldsymbol{\theta}_s$ and $\boldsymbol{x}_s(i)^\top \boldsymbol{\theta}_{s+1}$ such that

$$\nabla \ell_{s,i}(\boldsymbol{\theta}_{s+1}) - \nabla \ell_{s,i}(\boldsymbol{\theta}_s) - \nabla^2 \ell_{s,i}(\boldsymbol{\theta}_s)\Delta_s = \frac{\ddot{\mu}(\xi_{s,i})}{2}\big(\boldsymbol{x}_s(i)^\top \Delta_s\big)^2 \boldsymbol{x}_s(i).$$

Hence,

$$\begin{aligned}
&\left\langle \sum_{i=1}^N \nabla \ell_{s,i}(\boldsymbol{\theta}_{s+1}) - \nabla \widetilde{\ell}_s(\boldsymbol{\theta}_{s+1}), \boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^* \right\rangle \\
&= \sum_{i=1}^N \frac{\ddot{\mu}(\xi_{s,i})}{2}\big(\boldsymbol{x}_s(i)^\top \Delta_s\big)^2 \boldsymbol{x}_s(i)^\top (\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^*) \\
&\le \sum_{i=1}^N \frac{R\dot{\mu}(\xi_{s,i})}{2} |\boldsymbol{x}_s(i)^\top (\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^*)| \big(\boldsymbol{x}_s(i)^\top \Delta_s\big)^2 \\
&\le \sum_{i=1}^N \frac{R}{2} \cdot 2D \cdot L_\mu \|\Delta_s\|_2^2 = RDL_\mu N \|\Delta_s\|_2^2,
\end{aligned}$$

where we used $\|\boldsymbol{x}_s(i)\|_2 \le 1$, $\|\boldsymbol{\theta}_{s+1}\|_2 \le D$, $\|\boldsymbol{\theta}^*\|_2 \le D$, and $\dot{\mu} \le L_\mu$. Combining this bound with (27), we obtain

$$\begin{aligned}
\sum_{i=1}^N \langle \nabla \ell_{s,i}(\boldsymbol{\theta}_{s+1}), \boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^* \rangle \le \frac{1}{2\eta}\big(\|\boldsymbol{\theta}_s - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_s}^2 - \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_s}^2 - \|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}\|_{\boldsymbol{Q}_s}^2\big) \\
+ RDL_\mu N \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_2^2. \qquad (28)
\end{aligned}$$

Since $\eta = 1 + DR$, we have $1/(2\eta) = 1/(2 + 2DR)$. Substituting (28) into (26) and using the definition of $\boldsymbol{Q}_s$, we obtain

$$\begin{aligned}
\sum_{i=1}^N \big(\ell_{s,i}(\boldsymbol{\theta}_{s+1}) - \ell_{s,i}(\boldsymbol{\theta}^*)\big) \le \frac{1}{2 + 2DR}\Big(\|\boldsymbol{\theta}_s - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_s}^2 - \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_{s+1}}^2 - \|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}\|_{\boldsymbol{Q}_s}^2\Big) \\
+ RDL_\mu N \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_2^2.
\end{aligned}$$

Summing over $s = 1, \ldots, t$ gives

$$\begin{aligned}
&(2 + 2DR) \sum_{s=1}^t \sum_{i=1}^N \big(\ell_{s,i}(\boldsymbol{\theta}_{s+1}) - \ell_{s,i}(\boldsymbol{\theta}^*)\big) \\
&\le \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_1}^2 - \|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_{t+1}}^2 - \sum_{s=1}^t \|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s+1}\|_{\boldsymbol{Q}_s}^2 + (2 + 2DR)RDL_\mu N \sum_{s=1}^t \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_2^2.
\end{aligned}$$

Since $\boldsymbol{Q}_1 = \lambda_{\mathrm{op}} \boldsymbol{I}$ and $\Theta \subseteq \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \le D\}$, we have

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_1}^2 = \lambda_{\mathrm{op}} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|_2^2 \le 4\lambda_{\mathrm{op}} D^2.$$

Rearranging proves the first claim.

For the second claim, note that $\boldsymbol{Q}_s \succeq \lambda_{\mathrm{op}} \boldsymbol{I}$, and thus

$$\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_{\boldsymbol{Q}_s}^2 \ge \lambda_{\mathrm{op}} \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_2^2.$$

If $\lambda_{\mathrm{op}} \ge 6RDL_\mu N(1 + DR)$, then

$$(2 + 2DR)RDL_\mu N \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_2^2 \le \frac{1}{3}\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_{\boldsymbol{Q}_s}^2.$$

Substituting this into the first inequality yields

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|_{\boldsymbol{Q}_{t+1}}^2 \le (2 + 2DR) \sum_{s=1}^t \sum_{i=1}^N \big(\ell_{s,i}(\boldsymbol{\theta}^*) - \ell_{s,i}(\boldsymbol{\theta}_{s+1})\big) - \frac{2}{3}\sum_{s=1}^t \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s\|_{\boldsymbol{Q}_s}^2 + 4\lambda_{\mathrm{op}} D^2,$$

as claimed. □

The following lemma extends Zhang et al. [49, Lemma 5] to the CAB setting, where multiple feedback observations can be obtained at the same time.

Lemma E.5. *Let $\mathcal{G}_t$ be the filtration defined by $\mathcal{G}_t = \sigma\big(\{\{\boldsymbol{x}_s(i), y_s(i)\}_{i=1}^N\}_{s=1}^{t-1}\big)$ and $\{P_t\}_{t=1}^\infty$ be a stochastic sequence of distributions over $\mathbb{R}^d$. P_t is $\mathcal{G}_t$-measurable for each $t \geq 1$. Define $\ell_{s,i}(\boldsymbol{\theta}) = -y_s(i)\boldsymbol{x}_s(i)^\top\boldsymbol{\theta} + m\big(\boldsymbol{x}_s(i)^\top\boldsymbol{\theta}\big)$, $L_t(\boldsymbol{\theta}^*) = \sum_{s=1}^t\sum_{i=1}^N \ell_{s,i}(\boldsymbol{\theta}^*)$, and $F_t = -\sum_{s=1}^t \log\Big(\mathbb{E}_{\boldsymbol{\theta}\sim P_s}[\exp\Big(-\sum_{i=1}^N(\ell_{s,i}(\boldsymbol{\theta}))\Big)]\Big)$ for any $t \geq 1$.*

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that

$$\mathbb{P}\left[\forall t \in [T], L_t(\boldsymbol{\theta}^*) - F_t \leq \log\left(\frac{1}{\delta}\right)\right] \geq 1-\delta.$$

Proof. Since it is necessary to properly handle the joint distribution, we extend the proof of Zhang et al. [49, Lemma 5] under the assumption of conditional independence. First, define $M_t = \exp(L_t(\boldsymbol{\theta}^*) - F_t)$. For each natural parameter z, let $p(y;z)$ denote the normalized density or mass function of the GLM. We use only the likelihood-ratio identity

$$\frac{p(y;z)}{p(y;z')} = \exp(y(z-z') - m(z) + m(z')).$$

For notational convenience, write $p(y(1),\dots,y(N);\{z_i\}_{i=1}^N) = \prod_{i=1}^N p(y(i);z_i)$. By the definition of M_t, we have

$$M_t = \frac{\prod_{s=1}^t \mathbb{E}_{\boldsymbol{\theta}\sim P_s}\left[\prod_{i=1}^N \exp(-\ell_{s,i}(\boldsymbol{\theta}))\right]}{\prod_{s=1}^t\prod_{i=1}^N \exp(-\ell_{s,i}(\boldsymbol{\theta}^*))}.$$

Since $\{y_t(i)\}_{i=1}^N$ are conditionally independent given $\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^*\}_{i=1}^N$, we have

$$p(y_t(1),\dots,y_t(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^*\}_{i=1}^N) = \prod_{i=1}^N p(y_t(i);\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^*).$$

By the likelihood-ratio identity, for any $\boldsymbol{\theta}$,

$$\frac{\prod_{i=1}^N \exp(-\ell_{t,i}(\boldsymbol{\theta}))}{\prod_{i=1}^N \exp(-\ell_{t,i}(\boldsymbol{\theta}^*))} = \frac{p(y_t(1),\dots,y_t(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}\}_{i=1}^N)}{p(y_t(1),\dots,y_t(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^*\}_{i=1}^N)}.$$

Together with the definition of M_t, this implies that

$$M_t = M_{t-1}\frac{\mathbb{E}_{\boldsymbol{\theta}\sim P_t}\left[p(y_t(1),\dots,y_t(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}\}_{i=1}^N)\right]}{p(y_t(1),\dots,y_t(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^*\}_{i=1}^N)}. \tag{29}$$

By using the above discussion, we can show that $\{M_t\}_{t=1}^\infty$ is a martingale with respect to $\{\mathcal{G}_t\}_{t=1}^\infty$. Here, $\mathbb{E}[\cdot]$ denotes expectation with respect to $y_t(1),\dots,y_t(N)$.

$$\begin{aligned}
\mathbb{E}[M_t|\mathcal{G}_t] &= M_{t-1}\mathbb{E}\left[\frac{\mathbb{E}_{\boldsymbol{\theta}\sim P_t}\left[p(y_t(1),\dots,y_t(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}\}_{i=1}^N)\right]}{p(y_t(1),\dots,y_t(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^*\}_{i=1}^N)}\middle|\mathcal{G}_t\right] \\
&= M_{t-1}\int \frac{\mathbb{E}_{\boldsymbol{\theta}\sim P_t}\left[p(y(1),\dots,y(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}\}_{i=1}^N)\right]}{p(y(1),\dots,y(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^*\}_{i=1}^N)} \\
&\qquad \cdot p(y(1),\dots,y(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}^*\}_{i=1}^N)\,dy(1)\dots dy(N) \\
&= M_{t-1}\int \mathbb{E}_{\boldsymbol{\theta}\sim P_t}\left[p(y(1),\dots,y(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}\}_{i=1}^N)\right]dy(1)\dots dy(N) \\
&= M_{t-1}\mathbb{E}_{\boldsymbol{\theta}\sim P_t}\left[\int p(y(1),\dots,y(N);\{\boldsymbol{x}_t(i)^\top\boldsymbol{\theta}\}_{i=1}^N)dy(1)\dots dy(N)\right] \\
&= M_{t-1},
\end{aligned}$$

where we used the fact that M_{t-1} is $\mathcal{G}_t$-measurable and (29) in the first equality. Therefore, since $\{M_t\}_{t=1}^\infty$ is a martingale, $M_0 = 1$, and $M_t \geq 0$, we can use Lattimore and Szepesvári [26, Theorem 3.9] to obtain

$$\mathbb{P}\left[\sup_t M_t \geq \varepsilon\right] \leq \frac{1}{\varepsilon},$$

for any $\varepsilon > 0$. By setting $\varepsilon = 1/\delta$, we can prove the argument. □

Algorithm 5 Max match

Input: The total rounds T, the number of users N, and tuning parameter λ_0 and α.

1: $\mathcal{D}_0 = \varnothing$.
2: $\boldsymbol{V}_1 \leftarrow \lambda_0 \boldsymbol{I}$.
3: **for** $t = 1, \ldots, T$ **do**
4: $\quad \overline{\boldsymbol{\theta}}_t \leftarrow \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \widetilde{\mathcal{L}}(\mathcal{D}_t; \boldsymbol{\theta}, \kappa_\mu \lambda_0)$.
5: $\quad$ Choose $\pi_t = \arg\max_\pi \left(\sum_{a \in [K]} \sum_{i \in \pi^{-1}(a)} \mu(\boldsymbol{\phi}_t(i,a)^\top \overline{\boldsymbol{\theta}}_t) + g_t(\pi) \right)$, where g_t is defined in (3).
6: $\quad$ Observe $y_t(i)$ for any $i \in [N]$.
7: $\quad \boldsymbol{x}_t(i) \leftarrow \boldsymbol{\phi}_t(i, \pi_t(i))$ for any $i \in [N]$ and $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(\boldsymbol{x}_t(i), y_t(i))\}_{i \in [N]}$.
8: $\quad V_{t+1} \leftarrow \lambda_0 \boldsymbol{I} + \sum_{s=1}^t \sum_{i=1}^N \boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^\top$.

F Details of Section 5

This section describes the detailed experiment settings and reports additional results.

F.1 Baseline algorithms

We first describe detailed algorithms of "Max match" and "FairX". First, we show the detailed algorithm of Max match in Algorithm 5, which is a UCB-based algorithm maximizing the sum of matches. We can see the important difference between Max match and CAB-UCB in line 6. Max match chooses arms with the highest sum of expected matches instead of satisfaction.

Algorithm 6 shows FairX's algorithm in detail. FairX is a UCB-based fairness algorithm proposed by [46]. This method ensures that each arm receives a share of exposure that is proportional to its expected match, aiming to mitigate the over-selection of specific arms. [46] proposes UCB-based and TS-based algorithms in the stochastic linear bandit setting. For fair comparisons, we use the UCB-based algorithm and extend it to CAB. To approximate the argmax over CR_t, we draw 50 candidate parameters from the confidence region as follows: if $\boldsymbol{V}_t = \boldsymbol{L}_t \boldsymbol{L}_t^\top$, we first sample $\boldsymbol{x}$ uniformly from the unit Euclidean ball and then set $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}_t + \boldsymbol{L}_t^{-1}(\sqrt{\gamma}\boldsymbol{x})$, so that $\boldsymbol{\theta} \in \mathrm{CR}_t$. For each sampled candidate, we evaluate $\sum_{i \in [N]} \sum_{a \in [K]} \frac{\mu(\boldsymbol{\phi}_t(i,a)^\top \boldsymbol{\theta})}{\sum_{a' \in [K]} \mu(\boldsymbol{\phi}_t(i,a')^\top \boldsymbol{\theta})} \cdot \mu(\boldsymbol{\phi}_t(i,a)^\top \boldsymbol{\theta})$ and keep the best one.

We then describe the implementation of "CAB-TS (ε)" in Algorithm 2. The optimization problem in Algorithm 2 does not satisfy monotonicity since $h_t(\pi; \widetilde{\boldsymbol{\varepsilon}}_t)$ might be a negative value. For convenience, we set $w_a(S \cup j) - w_a(S)$ to 0 to retain monotonicity if $w_a(S \cup j) - w_a(S) < 0$. This implementation is a practical proxy for the perturbed allocation step and is not an exact implementation of the oracle assumed in the CAB-TS regret analysis.

Finally, we describe the hyperparameters of these algorithms. In the main text, we use $\lambda_0 = d$, $c_1 = \sqrt{d}$ for CAB-UCB and Max match, $a = \sqrt{dN}$ for CAB-TS, and $\gamma = 0.1$ for FairX.

F.2 Details of the setup

In this subsection, we provide the detailed setup of the synthetic experiments. We first define the 5-dimensional feature vector $\boldsymbol{\phi}(i,a)$ as follows.

$$\boldsymbol{\phi}(i,a) = \lambda \cdot \boldsymbol{\phi}_{pop}(i,a) + (1-\lambda) \cdot \boldsymbol{\phi}_{base}(i,a), \tag{30}$$

where $\boldsymbol{\phi}_{pop}$ and $\boldsymbol{\phi}_{base}$ are sampled from the standard normal distribution. Moreover, $\boldsymbol{\phi}_{pop}$ is componentwise strictly increasing with respect to the arm index, meaning that for every user i, component $m \in [5]$, and all $a \in [K-1]$, $[\boldsymbol{\phi}_{pop}(i,a)]_m < [\boldsymbol{\phi}_{pop}(i,a+1)]_m$. The parameter λ controls the strength of arm popularity. We use $\mu(x) = 1/(1+\exp(-x))$ and $r(x) = \min\{x, \beta\}$ as the feedback mean and satisfaction functions, respectively, so matches beyond β have no additional effect on satisfaction. The unknown true parameter $\boldsymbol{\theta}^*$ is sampled from the standard uniform distribution.

Algorithm 6 FairX

Input: The total rounds T, the number of users N, and tuning parameter λ_0 and γ.

$\mathcal{D}_0 = \varnothing$.
$\boldsymbol{V}_1 \leftarrow \lambda_0 \boldsymbol{I}$.
for $t = 1, \ldots, T$ **do**
  $\overline{\boldsymbol{\theta}}_t \leftarrow \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \widetilde{\mathcal{L}}(\mathcal{D}_t; \boldsymbol{\theta}, \kappa_\mu \lambda_0)$.
  Set $\mathrm{CR}_t = (\boldsymbol{\theta} : \|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_t\|_{\boldsymbol{V}_t} \le \sqrt{\gamma})$.
  Choose $\boldsymbol{\theta}_t = \arg\max_{\boldsymbol{\theta} \in \mathrm{CR}_t} \left(\sum_{i \in [N]} \sum_{a \in [K]} \frac{\mu(\boldsymbol{\phi}_t(i,a)^\top \boldsymbol{\theta})}{\sum_{a' \in [K]} \mu(\boldsymbol{\phi}_t(i,a')^\top \boldsymbol{\theta})} \cdot \mu(\boldsymbol{\phi}_t(i,a)^\top \boldsymbol{\theta}) \right)$.
  Construct policy $P_t(i,a) = \frac{\mu(\boldsymbol{\phi}_t(i,a)^\top \boldsymbol{\theta}_t)}{\sum_{a \in [K]} \mu(\boldsymbol{\phi}_t(i,a)^\top \boldsymbol{\theta}_t)}$.
  Sample $\pi_t \sim P_t$.
  Observe $y_t(i)$ for any $i \in [N]$.
  $\boldsymbol{x}_t(i) \leftarrow \boldsymbol{\phi}_t(i, \pi_t(i))$ for any $i \in [N]$ and $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(\boldsymbol{x}_t(i), y_t(i))\}_{i \in [N]}$.
  $V_{t+1} \leftarrow \lambda_0 \boldsymbol{I} + \sum_{s=1}^{t} \sum_{i=1}^{N} \boldsymbol{x}_s(i) \boldsymbol{x}_s(i)^\top$.

We compare CAB-UCB, CAB-TS (ε), the heuristic variant CAB-TS (θ), and the one-pass OMD against Random, Max match, and FairX. We fix $N = 50$ and $K = 10$ in all settings. The default comparisons in Figures 2a and 2b use $T = 10000$, $\lambda = 0.5$, and $\beta = 5.0$ over 10 runs. The λ- and β-sweeps in Figures 2c and 2d vary the corresponding parameter while fixing the others. The histogram analyses in Figures 2e and 2f use $T = 5000$, $\lambda = 1.0$, and $\beta = 5.0$ with 5 runs. For the appendix-only results in Figures 3a to 3c, we use the same data-generation procedure and set $T = 5000$, $\lambda = 0.5$, and $\beta = 5.0$ unless otherwise specified. Throughout the experiments, the allocation routines are practical proxies for the offline oracle steps used in the theory. Accordingly, CAB-TS (θ) should be interpreted only as an empirical heuristic variant, not as an algorithm covered by the theoretical guarantees, and the one-pass variant should likewise be interpreted as an empirical proxy for the theory-side oracle-based procedure.

Implementation of the one-pass variant For the one-pass OMD implementation, we set $\lambda_{\mathrm{op}} = 5.0$, $\eta = 1.0$, and $\delta = 0.05$ in all experiments. The projection set is $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \le \sqrt{d}\}$, which contains the true parameter generated from $[0,1]^d$. We initialize $\boldsymbol{\theta}_1 = \mathbf{0}$ and $\boldsymbol{Q}_1 = \lambda_{\mathrm{op}} \boldsymbol{I}$, and project each iterate onto the parameter set Θ. At round t, the implementation computes a confidence radius of the same form as in Appendix E using the current number of past observations, with $L_\mu = 1/4$ for the logistic link, and then forms the optimistic match estimate

$$\widetilde{\mu}_t(i,a) = \mu\Big(\boldsymbol{\phi}_t(i,a)^\top \boldsymbol{\theta}_t + \widehat{\beta}_t \|\boldsymbol{\phi}_t(i,a)\|_{\boldsymbol{Q}_t^{-1}} \Big).$$

Instead of solving the nested maximization in (18) exactly, we use the same sequential randomized welfare routine as in the other experimental methods. This replacement is heuristic and lies outside the scope of the theory in Appendix E. If s_a denotes the current accumulated optimistic matches for arm a, then arm a receives weight $\big(r(s_a + \widetilde{\mu}_t(i,a)) - r(s_a)\big)^{K-1}$; these weights are normalized to probabilities, and one arm is sampled. After observing the Bernoulli matches $y_t(i)$, the update step computes

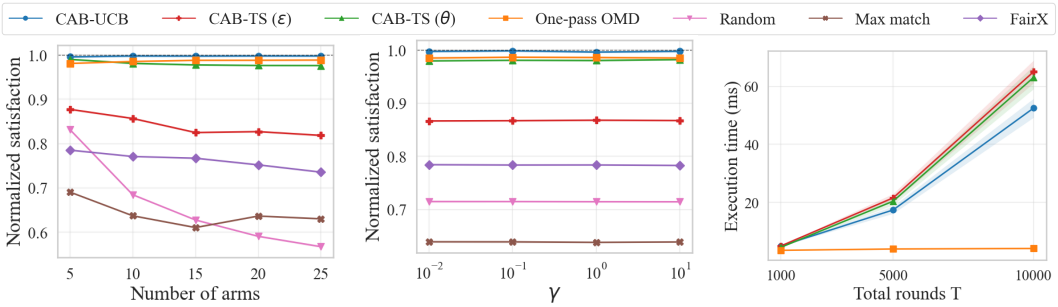
$$\boldsymbol{\Delta}_t = \left(\sum_{i=1}^{N} \nabla^2_{\boldsymbol{\theta}} \ell_{t,i}(\boldsymbol{\theta}_t) + \boldsymbol{Q}_t/\eta \right)^{-1} \left(\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \ell_{t,i}(\boldsymbol{\theta}_t) \right), \qquad \boldsymbol{\theta}_{t+1} = \Pi_\Theta(\boldsymbol{\theta}_t - \boldsymbol{\Delta}_t),$$

that is, it first takes the unconstrained minimizer of the quadratic surrogate and then projects it onto Θ. Finally, it updates

$$\boldsymbol{Q}_{t+1} = \boldsymbol{Q}_t + \sum_{i=1}^{N} \dot{\mu}\big(\boldsymbol{\phi}_t(i, \pi_t(i))^\top \boldsymbol{\theta}_{t+1}\big) \boldsymbol{\phi}_t(i, \pi_t(i)) \boldsymbol{\phi}_t(i, \pi_t(i))^\top.$$

We evaluate all methods in terms of cumulative arm satisfaction (our objective) and cumulative matches. These metrics are calculated as follows.

$$\text{Cumulative arm satisfaction} = \sum_{t=1}^{T} f_t(\pi_t; \boldsymbol{\theta}^*)$$



(a) Normalized cumulative satisfaction under varying K.

(b) Normalized cumulative satisfaction under varying γ.

(c) Average execution time per round under varying T.

Figure 3: Additional synthetic results. Panels (a) and (b) report normalized cumulative satisfaction, where each value is normalized by the cumulative satisfaction of a reference allocation computed with the true parameter and the same allocation oracle. Panel (a) varies the number of arms, panel (b) varies FairX's confidence parameter γ, and panel (c) compares the average execution time per round as the horizon length changes. Unless otherwise specified, we use $N = 50$, $T = 5000$, $\lambda = 0.5$, and $\beta = 5.0$, while fixing the remaining parameters to their default values.

$$\text{Cumulative matches} = \sum_{t=1}^{T} \sum_{a \in [K]} \sum_{i \in \pi_t^{-1}(a)} y_t(i),$$

For Figures 2c and 2d, we normalize cumulative satisfaction by the cumulative satisfaction of a reference allocation computed using the true parameter and the same allocation oracle. Solid lines denote means over runs, and shaded regions indicate 95% confidence intervals where applicable. All experiments were performed on a MacBook Air (13-inch, Apple M3, 16 GB RAM).

F.3 Additional results

We report three additional synthetic experiments in Figure 3, covering robustness to the number of arms, sensitivity to FairX's confidence parameter γ, and computational cost as the horizon length increases. In Figures 3a and 3b, each value is the cumulative satisfaction normalized by a reference allocation computed with the true parameter and the same allocation oracle. Unless otherwise specified, we fix $N = 50$, $K = 10$, $T = 5000$, $\lambda = 0.5$, and $\beta = 5.0$.

Varying the number of arms. Figure 3a shows that CAB-UCB remains essentially indistinguishable from the reference value across all tested values of K, while CAB-TS (ε), the heuristic variant CAB-TS (θ), and the one-pass variant also stay close to the reference value. In contrast, Random, Max match, and FairX remain substantially below the CAB methods, with the gap becoming particularly large for Random and Max match as K increases. These results indicate that the proposed CAB methods preserve arm-side satisfaction even when the action space becomes larger.

Varying FairX's parameter γ. Figure 3b shows that changing γ over several orders of magnitude has little effect on any method's normalized satisfaction. CAB-UCB consistently attains the highest value, and the other CAB variants remain close behind, whereas FairX stays well below the proposed methods for all tested values of γ. This suggests that FairX's limitation in this problem is not primarily a matter of tuning γ. Rather, it stems from optimizing a fairness criterion based on expected matches instead of the arm-satisfaction objective.

Runtime comparison. Figure 3c compares the average execution time per round while varying the horizon length T. The one-pass variant is the fastest method and remains nearly constant as T grows, whereas CAB-UCB, CAB-TS (ε), and the heuristic variant CAB-TS (θ) become noticeably slower for larger T. This pattern is consistent with the design of the one-pass variant, which avoids solving a regularized MLE from the full history at every round, and shows the expected computational trade-off between accuracy and efficiency.