Mixed-Effects Contextual Bandits

Kyungbok Lee[1], **Myunghee Cho Paik**[1, 2], **Min-hwan Oh**[3], **Gi-Soo Kim** [4 *]

[1] Department of Statistics, Seoul National University
[2] Shepherd23 Inc.
[3] Graduate School of Data Science, Seoul National University
[4] Department of Industrial Engineering, Ulsan National Institute of Science and Technology
turtle107@snu.ac.kr, myungheechopaik@snu.ac.kr, minoh@snu.ac.kr, gisookim@unist.ac.kr

Abstract

We study a novel variant of a contextual bandit problem with multi-dimensional reward feedback formulated as a mixed-effects model, where the correlations between multiple feedback are induced by sharing stochastic coefficients called random effects. We propose a novel algorithm, Mixed-Effects Contextual UCB (ME−CUCB), achieving $\tilde{O}(d\sqrt{mT})$ regret bound after T rounds where d is the dimension of contexts and m is the dimension of outcomes, with either known or unknown covariance structure. This is a tighter regret bound than that of the naive canonical linear bandit algorithm ignoring the correlations among rewards. We prove a lower bound of $\Omega(d\sqrt{mT})$ matching the upper bound up to logarithmic factors. To our knowledge, this is the first work providing a regret analysis for mixed-effects models and algorithms involving weighted least-squares estimators. Our theoretical analysis faces a significant technical challenge in that the error terms do not constitute martingales since the weights depend on the rewards. We overcome this challenge by using covering numbers, of theoretical interest in its own right. We provide numerical experiments demonstrating the advantage of our proposed algorithm, supporting the theoretical claims.

Introduction

Many real-world decision-making problems involve multiple outcomes for each decision. As an example, clinical trials in medicine often yield multivariate outcomes in response to a treatment. Measurements obtained from different parts of the brain (Lennihan et al. 2000), as well as assessments involving the eyes, ears, or psychological distress comprising depression and stress scales, can exhibit correlations with one another. This correlation arises as all these measurements are derived from the same individual. For behavioral intervention trials, treatments such as standard of care in a hospital and educational intervention are implemented collectively, and the outcomes of the subjects who receive intervention in the same session are correlated. Another example of correlated responses arises in user-rating systems, where each user evaluates multiple items. The ratings provided by the same user can exhibit correlation (Karumur, Nguyen, and Konstan 2016). In scenarios where we recommend a bundle of items to a single user and observe the ratings for each item, these ratings are stochastically correlated. The correlation does not stem from the items being related to each other but from the inclusion of unobservable variables specific to that user, such as the user's preferences or mood at the time, which act as shared random effects across all ratings. While this problem setting of multiple correlated responses per decision is prevalent in various applications, the study of such aspects in online decision-making has been limited.

In this paper, we propose a novel contextual bandit model where the reward feedback from a single action is given as vector and rewards can be correlated with each other in a vector. We formulate this model using a *mixed-effects model*. In statistical literature (Laird and Ware 1982; Zhang et al. 2016), mixed-effects model is a multivariate regression model with both fixed effects and random effects to handle the correlated structure among longitudinal/clustered outcomes. Fixed effects refer to usual non-stochastic regression coefficients common to all subjects, while random effects refer to stochastic coefficients specific for each subject. In mixed-effects model, outcomes from the same subject are correlated by sharing the random effects common to all measurements from the same subject. Typically, weighted estimators taking the correlations into account are utilized to efficiently estimate the regression parameters. To the best of our knowledge, our work is the first to adapt the mixed-effects model to contextual bandits.

In other variants of multi-armed bandit for a vector of rewards, such as combinatorial bandits (Chen, Wang, and Yuan 2013; Qin, Chen, and Zhu 2014; Li et al. 2016), potential correlations among rewards are commonly ignored. It is well known that when the outcomes are correlated, the weighted least-squares estimator (WLSE) with the weight incorporating correlation has a smaller variance than the ordinary least-squares estimator. Since the regret is commonly related to the estimation error, utilizing an efficient estimator has the potential for a tighter bound for the regret. However, incorporating correlation between rewards in bandit algorithms poses technical challenges in analysis. Using the weighted estimator addressing correlation requires a novel yet more involved analysis since the self-normalized martingale theorem (de la Pena, Klass, and Lai 2004) does not apply as in the regret analysis of canonical linear bandit algorithms (Abbasi-Yadkori, Pál, and Szepesvári 2011;

*Corresponding author

Agrawal and Goyal 2013), or their combinatorial variants (Qin, Chen, and Zhu 2014; Li et al. 2016). Elaborating on the technical challenge in more detail, the issue arises when the weighted least-squares estimator is used with the inverse of the estimated covariance matrix as weights. An empirical covariance matrix, defined as $\widehat{V}_{\tau,t}$ in (4), involves data observed later than time τ and is not independent of the error terms. This lack of independence impedes invoking the self-normalizing theorems (de la Pena, Klass, and Lai 2004) highly utilized in many of the contextual bandit literature. These challenges have not been addressed in previous works since in the statistical literature, asymptotic results on the mixed-effects models are well established (Jiang 2017) but not the non-asymptotic results. We present a novel non-asymptotic error bound on the weighted least-squares estimator (WLSE) for the mixed-effects models.

To this end, we provide both algorithmic and theoretical contributions in this paper. First, we propose a novel algorithm for the mixed-effects bandit model, Mixed-Effects Contextual UCB (`ME-CUCB`). As a theoretical contribution, we provide the regret analysis of `ME-CUCB` achieving provable efficiency, yielding a tighter regret bound than that of the naive canonical linear bandit algorithm ignoring the correlation structure. To the best of our knowledge, this is the first to present a regret analysis for mixed-effects models and for algorithms involving weighted least-squares estimators. Our main contributions are summarized as follows:

Contributions

- We propose a new contextual bandit problem that allows for multi-dimensional reward feedback and potential correlations among rewards, using the mixed-effects model. We propose UCB-type algorithms (`ME-CUCB`, Algorithms 1 and 2) that achieve superior regret performance compared to the naive contextual bandit algorithm, with provable guarantees (Proposition 2).
- We provide an estimation error bound for the weighted regression coefficients where the estimator is a sum of contributions of each time point, but the weights depend on the data obtained from future time points (Theorem 4). To the best of our knowledge, this is the first error bound of such an estimator without martingale property.
- We provide the regret analysis of our proposed algorithms. We establish $O(d\sqrt{mT}\log T)$ worst-case regret bound for the both cases where the covariance structure is known (Theorem 3) and unknown (Theorem 6), for our proposed algorithms.
- We prove a lower bound of $\Omega(d\sqrt{(m\lambda_{\max}(D)+\sigma^2)T})$ for cumulative regrets of our problem considering the covariance matrix D and noise variance σ^2 (Theorem 7). This lower bound matches the regret upper bound of `ME-CUCB` up to logarithmic factors, proving the near-optimality of our method.
- Our numerical experiments demonstrate that our proposed algorithms outperform benchmarks in terms of cumulative regret.

Related Works

Multi-armed bandit algorithms based on upper confidence bounds (UCB) select arms by the principle of optimism in the face of uncertainty. The UCB approach was introduced by Lai, Robbins et al. (1985) and proven to achieve logarithmic regret bound with a known gap between optimal/sub-optimal arms (Auer, Cesa-Bianchi, and Fischer 2002). Auer (2002) introduced a contextual bandit algorithm `LinRel` with linear reward function. Li et al. (2010) proposed `LinUCB` with $\tilde{O}(d\sqrt{T})$ regret bound and applied the algorithm to news recommendation. Chu et al. (2011) proposed `SupLinUCB` with a tighter $\tilde{O}(\sqrt{dT})$ regret bound and proved matching lower bound up to logarithmic factors.

There are variants of contextual bandit problem that allow for multi-dimensional vector reward feedback: combinatorial bandits (Chen, Wang, and Yuan 2013; Qin, Chen, and Zhu 2014; Li et al. 2016) and multi-objective bandits (Tekin and Turgay 2017; Lu et al. 2019). In either case, the correlation structure has not been addressed. In combinatorial bandits, the agent pulls a set of multiple base arms, called super-arm at each round. A vector of outcomes with multiple values corresponding to multiple *base* arms are observed, and a specified reward function aggregates multiple values into a scalar reward. Qin, Chen, and Zhu (2014) introduced `C`^{2}`UCB` achieving $O(d\sqrt{mT}\log T)$ regret rate where m is the size of a super-arm, which corresponds to the dimension of a reward vector in our case. Another example is a multi-objective bandit (Drugan and Nowe 2013; Busa-Fekete et al. 2017; Turgay, Oner, and Tekin 2018) where a vector of rewards are observed when pulling an arm. In this setting, element-wise comparison is of interest. A particular element in the outcome vector can be better in one dimension, but another element is worse in another dimension. Drugan and Nowe (2013) proposed Pareto regret summarizing element-wise performance of multi-objective outcomes.

In statistical literature, to handle clustered data, mixed-effects models (Laird and Ware 1982) or generalized estimating equation (Liang and Zeger 1986) are commonly used. Both methods employ weighted least-squares estimators utilizing the covariance structure of clustered data to reduce the variance of the regression coefficient estimators. In contextual bandit algorithms, such estimators can potentially reduce the regret bound through smaller estimation error bound, but to our knowledge, no existing algorithms take advantage of weighted least-squares estimators.

Zhu and Kveton (2022a) proposed random effects bandit for the non-contextual case assuming the mean reward of each arm arising from a distribution. The setting differs from ours as they consider a single reward for each arm, and the rewards from same arm are correlated across time. We consider multiple outcomes for each arm correlated at each time point but independent across time.

Aouali, Kveton, and Katariya (2023) introduced a mixed-effect Thompson sampling algorithm, using similar terminology to our proposed *mixed-effects* contextual bandits. In this work, mixed-effect implies that the coefficients for each arm are sampled from a prior distribution based on a mixture of multiple effect parameters, within a Bayesian hierarchical

framework, and is different from the statistical mixed-effects model of Laird and Ware (1982). The *mixed-effects* in our work differs, as we employ a statistical mixed-effects model that combines fixed-effects and random-effects from a frequentist perspective.

Distinction From Bayesian Hierarchical Bandit Models Several Bayesian hierarchical bandit models have been proposed to handle correlated rewards across arms (Kveton et al. 2021; Wan, Ge, and Song 2021; Hong et al. 2022a; Zhu and Kveton 2022b; Aouali, Kveton, and Katariya 2023) or multiple tasks (Hong et al. 2022b). Our proposed mixed-effects bandit model is distinct in many aspects.

Firstly, the Bayesian hierarchical models are designed for handling correlated rewards from different arms or multiple tasks, whereas our model deals with a single task that pulls a single arm and returns a correlated outcome vector. This means that the problem we handle is fundamentally different from those addressed in the hierarchical bandit models.

Secondly, the theoretical justifications for hierarchical bandit models formulated within a Bayesian framework require a distributional Gaussian assumption with a known variance-covariance matrix, which is a major concern of Bayesian theory (Demidenko 2013). In contrast, our mixed-effects bandit model does not specify any distribution and allows the variance-covariance matrix to be unknown. This flexibility is a major advantage as, in practice, knowledge of the covariance structure is often unavailable.

This distinction leads to different theoretical results, as previous models derive Bayes regret bounds based on the Bayesian prior, while our approach provides a frequentist's worst-case regret bound by estimating the unknown covariance. In this process, we addressed a theoretical challenge not required in Bayesian methods: bounding the self-normalizing norm of terms that do not form a martingale.

Problem Setting and Assumptions

Mixed-Effects Bandit Problem

Consider a contextual bandit problem where pulling each arm returns a vector of outcomes. Let T be the number of rounds, K be the number of arms, d and k be the dimension of the context vector of fixed effects and random effects, respectively. We select an arm a_t at each time step t. We assume that the outcome vector Y_{ti} of m elements for arm i at round t has the form of linear mixed-effects model

$$Y_{ti} = X_{ti}\beta + Z_{ti}\gamma_{ti} + e_{ti} = \mu_{ti} + \eta_{ti},$$

where $X_{ti} \in \mathbb{R}^{m\times d}$ and $Z_{ti} \in \mathbb{R}^{m\times k}$ are the context matrix for the fixed effect $\beta \in \mathbb{R}^d$ and the centered random effect $\gamma_{ti} \in \mathbb{R}^k$, respectively. The random effect γ_{ti} is a random variable common to all elements in Y_{ti} and induces correlation among them. We remark that we do not require γ_{ti} to be independent across arms given $\mathcal{H}_t$, and include the case where $\gamma_{t1} = \gamma_{t2} = \cdots = \gamma_{tK}$. The vector $e_{ti} \in \mathbb{R}^m$ is the noise of arm i at round t, and $\mu_{ti} = X_{ti}\beta$ and $\eta_{ti} = Z_{ti}\gamma_{ti} + e_{ti}$. We assume that γ_{ti} is independent of e_{ti}.

Two popular mixed-effects models are (i) *random intercept model* and (ii) *random coefficient model*. In random intercept model, $Z_{ti} = \mathbf{1}_m$ and γ_{ti} is a scalar random variable called random intercept. In this case, $\mathrm{Cov}(Y_{til}, Y_{tig}) = \mathrm{Var}(\gamma_{ti})$ for $l \neq g$, where Y_{til} denotes the l^{th} element of Y_{ti}. In random coefficient models, columns of Z_{ti} consist of all or some columns of X_{ti}. Detailed examples of these two models are provided below.

The goal is to maximize the expected cumulative reward $\sum_{t=1}^{T} f(\mu_{ta_t})$ over T rounds. Here $f : \mathbb{R}^m \to \mathbb{R}$ is a user-selected function measuring the quality of the selected arm. This is equivalent to minimizing the cumulative regret $R(T) = \sum_{t=1}^{T} \texttt{regret}(t)$, where $\texttt{regret}(t) = f(\mu_{ta_t^*}) - f(\mu_{ta_t})$ for the optimal arm $a_t^* = \arg\max_{i\in[K]} f(\mu_{ti})$.

Assumptions

For the theoretical regret analysis, we provide the following assumptions.

Assumption 1 (Boundedness)**.** For all $i \in [K]$ and $j \in [m]$, $\|X_{tij}\|_2 \leq 1$, $\|Z_{tij}\|_2 \leq 1$ and $\|\beta\|_2 \leq 1$ where X_{tij} and Z_{tij} are the j^{th} row of X_{ti} and Z_{ti}, respectively.

Assumption 2 (Sub-Gaussianity)**.** For a symmetric positive definite matrix $D \in \mathbb{R}^{k\times k}$ and $\sigma^2 > 0$,

$$\mathbb{E}(\gamma_{ti}) = 0, \mathrm{Var}(\gamma_{ti}) = D, \mathbb{E}\exp(\lambda^\top \gamma_{ti}) \leq \exp(\frac{1}{2}\lambda^\top D\lambda)$$

for any $\lambda \in \mathbb{R}^k$ and $\mathbb{E}(e_{ti}) = 0_m$, $\mathrm{Var}(e_{ti}) = \sigma^2 I_m$, $\mathbb{E}\exp(\lambda^\top e_{ti}) \leq \exp(\frac{1}{2}\sigma^2\lambda^\top\lambda)$ for any $\lambda \in \mathbb{R}^m$, where all expectations and variances are conditioned to $\mathcal{H}_t$, the history σ-field containing all the information at the start of round t.

Assumption 3 (Assumptions on f)**.**
(3-1) (Monotonicity). If $Y_1 \leq Y_2$ elementwisely, then $f(Y_1) \leq f(Y_2)$.
(3-2) (Lipschitz continuity). There exists $L > 0$ such that $|f(Y_1) - f(Y_2)| \leq L\|Y_1 - Y_2\|_2$ for any $Y_1, Y_2 \in \mathbb{R}^m$.

Assumption 1 is widely used in the contextual bandit literature (Chu et al. 2011; Agrawal and Goyal 2013) and the boundedness of Z_{tij} is added. Assumption 2 is an extension of the sub-Gaussian assumption commonly used in the bandit literature. It is required because our random effects and noises can be vectors with certain covariance matrices. Assumptions 3-1 and 3-2 are necessary in bandit problems where the reward depends on an outcome vector with multiple elements, such as combinatorial bandit problems (Chen, Wang, and Yuan 2013; Qin, Chen, and Zhu 2014; Li et al. 2016; Zhang, Li, and Liu 2019).

Weighted Least-Squares Estimator

By Assumption 2, the covariance of Y_{ta_t} conditioned on $\mathcal{H}_t$ is $V_t = Z_{ta_t} D Z_{ta_t}^\top + \sigma^2 I_m$. Let $V_t^* \in \mathcal{V}$ be an arbitrary symmetric positive definite matrix, where $\mathcal{V}$ is the class of all m-dimensional positive definite matrices with its eigenvalues in $[\lambda_0, \Lambda_0]$ containing the true V_t. Write the upper bound of the eigenvalues of V_t by $\Lambda = m\lambda_{\max}(D) + \sigma^2$ where $\lambda_{\max}(D)$ denoting the maximum eigenvalue of D.

We consider the following class of weighted least-squares estimators of β when V_t is known:

$$\beta_t^* = B_t^{*-1} \sum_{\tau=1}^{t} X_{\tau a_\tau}^\top V_\tau^{*-1} Y_{\tau a_\tau}, \tag{1}$$

where $B_t^* = \sum_{\tau=1}^{t} X_{\tau a_\tau}^\top V_\tau^{*-1} X_{\tau a_\tau} + I_d$. This class of estimators includes the ridge ordinary least-squares estimator when $V_\tau^* = I_m$, and the ridge weighted least-squares estimator when $V_\tau^* = V_\tau$. The variability of β_t^* is captured by

$$C_t = B_t^{*-1}(\sum_{\tau=1}^{t} X_{\tau,a_\tau}^\top V_\tau^{*-1} V_\tau V_\tau^{*-1} X_{\tau,a_\tau} + I_d) B_t^{*-1}. \tag{2}$$

The following proposition shows that C_t determines the upper bound of the estimation error of β_t^*.

Proposition 1. *For any $\{V_\tau\}_{\tau=1}^t \subset \mathcal{V}$, $x \in \mathbb{R}^d$ with probability $1-\delta$,*

$$|x^\top(\beta_t^* - \beta)| \le \alpha_t^* \left\|x\right\|_{C_t}, \tag{3}$$

$$\alpha_t^* = \sqrt{d \log(\frac{m\Lambda}{\lambda_0^2 d} t + 1)\delta^{-\frac{2}{d}}} = O(\sqrt{d(\log t + \log m)}).$$

When $V_t^* = V_t$, that is, the true covariance matrices are used, the value of C_t is given by $C_t = B_t^{-1}$, where $B_t = \sum_{\tau=1}^{t} X_{\tau a_\tau}^\top V_\tau^{-1} X_{\tau a_\tau}$. Later in Proposition 2, we prove that the use of true covariance values results in the smallest instantaneous regret bound.

For the case where the D and σ^2 are unknown, we consider the following WLS estimator:

$$\widehat{\beta}_t = \widehat{B}_t^{-1} \sum_{\tau=1}^{t} X_{\tau a_\tau}^\top \widehat{V}_{\tau,t}^{-1} Y_{\tau a_\tau} \tag{4}$$

for matrix $\widehat{B}_t = \sum_{\tau=1}^{t} X_{\tau a_\tau}^\top \widehat{V}_{\tau,t}^{-1} X_{\tau a_\tau} + I_d$ and $\hat{V}_{\tau,t} = Z_{\tau a_\tau} \widehat{D}_t Z_{\tau a_\tau}^\top + \widehat{\sigma}_t^2 I_m$.

Consistent closed-form estimators $\widehat{D}_t, \widehat{\sigma}_t^2 \in \mathcal{H}_t$ for D and σ^2 are available for the random intercept models. One approach involves constructing an estimating equation using the sample covariance matrix of Y_t and its expected value. Another method is to maximize the marginal likelihood function of multivariate normal Y_t using Expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) with respect to D and σ^2. Score functions for D and σ^2 are unbiased estimating equations even when Y_t is not distributed as Gaussian, as long as the variance structure is correctly specified. Unbiasedness ensures that the solutions of the equations converge to the true values of D and σ^2 in probability.

Algorithm

We present novel algorithms, mixed-effects contextual UCB for mixed-effects contextual bandit problem for two different cases. `ME-CUCB1` (algorithm 1) is for the case where the true D and σ^2 are known, so that one needs not to estimate the covariance matrices. `ME-CUCB2` (algorithm 2) is for the general case where we estimate unknown D and σ^2.

Algorithm 1:
Mixed-Effects Contextual UCB1 (`ME-CUCB1`)

1: **INPUT:** Covariance parameters D, σ^2 and exploration parameter α.
2: **for** $t \in [T]$ **do**
3: observe the contexts $\{X_t\}$, $\{Z_t\}$ and compute $\widehat{Y}_{ti}$ for each arm i using equation (5).
4: Play arm $a_t = \arg\max_{i\in[m]} f(\widehat{Y}_{ti})$ and observe Y_{ta_t}.
5: Select $V_t^* \in \mathcal{V}$ and update $B_t^* = B_{t-1}^* + X_{ta_t}^\top V_t^{*-1} X_{ta_t}$.
6: Update $\beta_t^* = B_t^{*-1} \sum_{\tau=1}^{t} X_{\tau a_\tau}^\top V_\tau^{*-1} Y_{\tau a_\tau}$ and
$C_t := B_t^{*-1}(\sum_{\tau=1}^{t} X_{\tau a_\tau}^\top V_\tau^{*-1} V_\tau V_\tau^{*-1} X_{\tau a_\tau} + I_d) B_t^{*-1}$.
7: **end for**

`ME-CUCB` for Known D and σ^2

For the case where D and σ^2 are known, we present `ME-CUCB1` in algorithm 1, an UCB-type algorithm for mixed-effects bandit problem. At time step t we choose an arm $a_t = \arg\max_{i\in[K]} f(\widehat{Y}_{ti})$ and observe Y_{ta_t}, where the j^{th} element of the upper confidence bound vector $\widehat{Y}_{ti}$ is

$$\widehat{Y}_{tij} = X_{tij}^\top \beta_{t-1}^* + \alpha_t^* \left\|X_{tij}\right\|_{C_{t-1}}. \tag{5}$$

The following proposition presents the instantaneous regret bound of `ME-CUCB1` and demonstrates that the bound is minimized by utilizing the true covariance matrices.

Proposition 2. *Under the event where* (3) *holds, the regret of `ME-CUCB1` at time step t is bounded by $\texttt{regret}(t) \le 2L\alpha_t^* \sqrt{\sum_{j=1}^{m} \left\|X_{ta_t j}\right\|_{C_{t-1}}^2}$. The upper bound is minimized using the true $V_t^* = V_t$.*

Proposition 2 shows that among the estimators considered in (1), the estimator using true covariance yields the tightest bound by achieving the smallest estimation variance. Taking the correlation into account lead to a better regret bound than the standard UCB, ignoring the correlation.

`ME-CUCB` for Unknown D and σ^2

For the general case where the D and σ^2 are unknown, we present `ME-CUCB2` in algorithm 2.

The key difference is that we estimate $\widehat{D}_t$ and $\widehat{\sigma}_t^2$ at each round and we replace β_{t-1}^* and C_{t-1} with their plug-in estimators $\widehat{\beta}_{t-1}$ from (4) and $\widehat{B}_{t-1}^{-1}$ respectively to construct the UCB in (6). Although $\widehat{D}_t$ and $\widehat{\sigma}_t^2$ are plugged in instead of the true values D and σ^2, we show later that under mild assumptions, the estimation error of $\widehat{\beta}_t$ has the same order as the error of β_t^*. Consequently, the formula in (6) is a valid UCB as well. After c random exploration rounds, at the start of round t we obtain $\widehat{D}_t, \widehat{\sigma}_t^2 \in \mathcal{H}_t$. In practice, we do not need to update $\widehat{D}_t$ and $\widehat{\sigma}_t^2$ at every time step but only update occasionally. Given $\widehat{\beta}_{t-1}$, the upper bound of the j^{th} element of Y_{ti} is computed by

$$\widehat{Y}_{tij} = X_{tij}^\top \widehat{\beta}_{t-1} + \alpha_t \left\|X_{tij}\right\|_{\widehat{B}_{t-1}^{-1}} \tag{6}$$

Algorithm 2:
Mixed-Effects Contextual UCB2 (ME-CUCB2)

INPUT: number of random exploration rounds c, exploration parameter α.
for $t \leq c$ **do**
  Sample an arm $a_t \in [K]$ randomly and observe Y_{ta_t}.
end for
Compute $\widehat{D}_c, \widehat{\sigma}_c^2, \widehat{V}_{\tau,c}, \widehat{B}_c$ and $\widehat{\beta}_c$ by equation (4).
for $t > c$ **do**
  Observe the contexts $\{X_t\}$, $\{Z_t\}$ and compute $\widehat{D}_t, \widehat{\sigma}_t^2 \in \mathcal{H}_t$.
  Compute $\widehat{Y}_{ti}$ for each arm i using equation (6).
  Play arm $a_t = \arg\max_{i\in[m]} f(\widehat{Y}_{ti})$ and observe Y_{ta_t}.
  Compute $\widehat{V}_{\tau,t} = Z_{\tau,a_\tau}\widehat{D}_t Z_{\tau,a_\tau}^\top + \widehat{\sigma}_t^2 I_m$ for $\tau \in [t]$ and $\widehat{B}_t = \sum_{\tau=1}^{t} X_{\tau a_\tau}^\top \widehat{V}_{\tau,t}^{-1} X_{\tau a_\tau} + I_d$.
  Update $\widehat{\beta}_t = \widehat{B}_t^{-1} \sum_{\tau=1}^{t} X_{\tau a_\tau}^\top \widehat{V}_{\tau,t}^{-1} Y_{\tau a_\tau}$.
end for

for some $\alpha_t > 0$. Then we choose an arm with maximum value of $f(\widehat{Y}_{ti})$ by $a_t = \arg\max_{i\in[K]} f(\widehat{Y}_{ti})$ to observe Y_{ta_t}. In the following section, we conduct regret analysis for the cases when D and σ^2 are known and unknown.

Regret Analysis

Why the Usual Self-Normalizing Martingale Norm Technique Cannot Be Applied?

In the regret analysis of ME-CUCB2, an additional term emerges compared to that of ME-CUCB1 due to the estimation error in $\widehat{D}_t$ and $\widehat{\sigma}_t^2$. Using the approach in Lattimore, Crammer, and Szepesvári (2015), it is possible to construct a martingale based only on data up to round $\tau < t$. However, this approach fails to fully exploit the available data up to round t, leading to an inability to fully leverage the consistency of the estimators. This could result in additional regret bound terms that cannot be neglected.

To overcome this limitation, our proposed algorithm uses $\widehat{V}_{\tau,t}$ as the weight matrix. The computation of $\widehat{V}_{\tau,t}$ uses all the available data up to round t. As a result, a dependency between the weight and error term emerges, causing the terms to not satisfy the martingale property. This non-martingale property prevents conventional techniques for martingale norms from being applicable. As an alternative, we introduce a new approach that utilizes the covering number to derive the regret bound.

Regret Bound of ME-CUCB1 (Algorithm 1) for Known D and σ^2

Theorem 3 (Regret bound of ME-CUCB1)**.** *With probability $1-\delta$, the cumulative regret of ME-CUCB1 by time T is bounded by*

$$R(T) \leq 2\sqrt{2}Ld\sqrt{\Lambda T}\sqrt{\log\frac{\frac{mT}{\sigma^2 d}+1}{\delta^{2/d}}}\sqrt{\log(\frac{mT}{\Lambda d}+1)}.$$

Since $\Lambda = m\lambda_{\max}(D) + \sigma^2$, the regret bound rate is $R(T) = O(d\sqrt{mT}\log T)$. The regret bound of ME-CUCB1 does not depend on k, the dimension of random effects. Intuitively, we do not estimate the random effects that determine the covariance structure; the impact of random effects on the regret bound is reflected in the value of Λ.

Regret Bound of ME-CUCB2 (Algorithm 2) Using Consistent Estimator of D and σ^2

A natural question that arises is whether plugging in a 'good' estimator for D and σ^2 provides a comparable regret bound. Here, we provide an affirmative answer to this question. For any $\delta > 0$, consider a monotonically decreasing ϵ_t such that, with probability $1-\delta$,

$$\left\|\hat{D}_t - D\right\|_{\mathrm{F}} \leq \epsilon_t, \;\; |\hat{\sigma_t}^2 - \sigma^2| \leq \epsilon_t \tag{7}$$

for all $t \in [T]$. That is, the estimators of D and σ^2 are assumed to be consistent with a rate of ϵ_t. An example of such an estimator is the maximum likelihood type estimator of D and σ^2, which satisfies the condition (7) with $\epsilon_t = O(t^{-1/2})$, under mild regularity conditions. This ensures the existence of the estimators $\widehat{D}_t$ and $\widehat{\sigma}_t^2$ assumed in this section. In the following section, we use $\epsilon_t = t^{-1/2}$ up to a constant coefficient, without loss of generality.

Let E_1 be the event where (7) holds so that $\mathbb{P}(E_1) \geq 1-\delta$, and define $\Lambda_1 = m\lambda_{\max}(D) + \sigma^2 + m + 1$, the upper bound of the eigenvalues of $\widehat{V}_{\tau,t}$ under E_1. For ME-CUCB2, one need invertibility of $\widehat{V}_{\tau,t}$ to bound the regret. The exploration rounds with $c = \lceil 4/\sigma^4 \rceil$ guarantee the minimum eigenvalue of $\widehat{V}_{\tau,t}$ to be bounded below by $\sigma^2/2$ under E_1. The regret incurred by c exploration rounds is bounded by $\sum_{\tau=1}^{c}(f(\mu_{\tau a_\tau^*}) - f(\mu_{\tau a_\tau})) \leq 2L\sqrt{mc}$, independent of T.

Estimator Error Bound We proceed our analysis under E_1 from now on. de la Pena, Klass, and Lai (2004) and Abbasi-Yadkori, Pál, and Szepesvári (2011) derived tight bounds for estimation errors when the error is expressed as a normalized martingale, where the normalization term has dependency on data from all time steps. The error of our estimator also decomposes as a normalized sum, $\widehat{\beta}_t - \beta = \widehat{B}_t^{-1}\sum_{\tau=1}^{t} X_\tau^\top \widehat{V}_{\tau,t}^{-1}\eta_\tau$. However, the term $\sum_{\tau=1}^{t} X_\tau^\top \widehat{V}_{\tau,t}^{-1}\eta_\tau$ without normalization is not a martingale since $\hat{V}_{\tau,t}$ involves the data observed after τ, so we cannot apply corollary 4.3 of de la Pena, Klass, and Lai (2004) or theorem 1 of Abbasi-Yadkori, Pál, and Szepesvári (2011). To address this new challenge, we decompose the error into three terms, and use the covering number arguments to break the dependency between $\widehat{V}_{\tau,t}^{-1}$ and η_τ.

Theorem 4 (estimation error bound of ME-CUCB2)**.** *For any $x \in \mathbb{R}^d$ and all $t \geq c$, with probability $1-3\delta$*

$$\left|x^\top(\widehat{\beta}_t - \beta)\right| \leq \alpha_t(1+R_t)\left\|x\right\|_{B_t^{-1}} \tag{8}$$

for $\alpha_t = \sqrt{d\log\frac{\frac{mt}{\sigma^2 d}+1}{\delta^{2/d}}} = O(\sqrt{d\log t})$ and $R_t = O(m\sqrt{d}\epsilon_t) = O(m\sqrt{d}t^{-1/2})$.

Theorem 4 shows that the estimation error bound from algorithm ME−CUCB2 is of the same main order as that of algorithm ME−CUCB1 using true D and σ^2. The extra term including R_t is of non-dominant order.

Regret Analysis Finally, we take care of the fact that the output at each round is a vector and the expected rewards depend on the f, and present the instantaneous regret bound.

Proposition 5. *With probability $1-3\delta$, for all $t > c$, the regret at time step t is bounded above by*

$$\texttt{regret}(t) \leq 2L\alpha_t(1+2S_t)\sqrt{\sum_{j=1}^{m}\|X_{ta_tj}\|^2_{B_{t-1}^{-1}}},$$

$$S_t = 2\sigma^{-2}\sqrt{m\epsilon_t} + R_{t-1} = O(\sqrt{m\epsilon_t}) = O(\sqrt{m}t^{-1/4}).$$

The first term of S_t emerges due to using $\hat{B}_t$ instead of B_t in the bonus term in (6) while the second term, R_t, is extra variation for estimating the variance in the WLSE as shown in Theorem 4. We are now ready to bound the cumulative regret of ME−CUCB2.

Theorem 6 (Regret bound of ME−CUCB2 with consistent estimators)**.** *With probability $1-3\delta$, with c random exploration rounds the cumulative regret $R(T)$ of the algorithm 2 is bounded by*

$$R(T) \leq 2\sqrt{2}Ld\sqrt{\Lambda T}\sqrt{\log\frac{\frac{mT}{\sigma^2 d}+1}{\delta^{2/d}}}\sqrt{\log(\frac{mT}{\Lambda d}+1)} + O(dmT^{1/4}\log T).$$

The regret bound rate for ME−CUCB2 is expressed as $O(d\sqrt{mT}\log T)+O(dmT^{1/4}\log T)$. The second term captures the additional variation that arises from using the estimated covariance instead of the true covariance. The the term $O(dmT^{1/4}\log T)$ represents the difference between the regret bounds of ME−CUCB2 and ME−CUCB1, which is negligible compared to the leading term $O(d\sqrt{mT}\log T)$. This implies that ME−CUCB2 employing the consistent estimator for D and σ^2, achieves the same regret rate as the optimal ME−CUCB1.

Matching Lower Bounds

We establish the regret lower bound for the multi-dimensional feedback bandit problem matching the regret upper bound outlined in Theorems 3 and 6 up to logarithmic factors. We present instances that achieve a lower bound of $\Omega(d\sqrt{\Lambda T}) = \Omega(d\sqrt{(m\Lambda_{\max}(D)+\sigma^2)T})$. To construct the context for random effects, we decompose the D as $D = P\bar{\Lambda}P^\top$, for an orthonormal matrix P and the diagonal matrix $\bar{\Lambda}$ with sorted eigenvalues of D as its elements, with $\lambda_{\max}(D) = \bar{\Lambda}_{11}$. Denote the first column of P as P_1. We can prove the following theorem.

Theorem 7. *Consider an instance with $\{X_{tij}\}_{i=1}^{2^d} = \{-\sqrt{\Lambda d/T}, \sqrt{\Lambda d/T}\}^d$ for all $j \in m$ and $\{Z_{tij}\} = P_1^\top$ for all $i \in [2^d]$ and $j \in m$. Then for any algorithm, there exists $\beta \in \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$ such that $R(T) = \Omega(d\sqrt{\Lambda T})$ for $T > (m\lambda_{\max}(D)+\sigma^2)d^2$.*

The lower bound in Theorem 7 matches our regret upper bound for ME−CUCB as stated in Theorems 3 and 6 up to a logarithmic factor. Therefore, our proposed algorithms are provably near-optimal.

Numerical Experiments

Simulation Data

We compare the cumulative regret of the following algorithms: (i) C^{2}UCB (Qin, Chen, and Zhu 2014) (ii) the proposed ME−CUCB1 with true D and σ^2 (iii) the proposed ME−CUCB2 with estimated D and σ^2. For C^{2}UCB we restrict the super-arms to K arms containing m context vectors. All three have α as a hyperparameter to control the exploration rate. We run the experiments with $\alpha \in \{10^{-3}, 2\cdot 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ and choose the value with smallest average cumulative regret. We present the results for the random intercept model and the random coefficient model.

Random Intercept Model We consider the random intercept model $Y_{ti} = X_{ti}\beta + b_{ti}\mathbf{1}_m + e_{ti}$, and $f(\mu) = m^{-1}\mathbf{1}_m^\top\mu$ to maximize the expected average outcome. We generate each element of $X_{ti} \in \mathbb{R}^{m\times d}$ from $\mathcal{N}(0,1)$ and truncate them to satisfy $\|X_{tij}\|_2 \leq 1$. Each element of the fixed effect β is sampled from a uniform distribution $U(\pm 1/\sqrt{d})$. To generate the stochastic output we sample the random intercept $b_{ti} \sim \mathcal{N}(0, D)$ and the noise vector $e_{ti} \sim \mathcal{N}(0, I_m)$. We fix (d, K, m) to $(10, 100, 10)$ and choose the value of D from $\{0.0, 1.0, 5.0\}$ to observe how the algorithms perform as the correlation changes. The case with $D = 0$ represents when there is no correlation. We run $c = 10$ random exploration rounds.

The upper row of the Figure 1 shows the mean cumulative regret $R(T)$ for $T = 1000$ over 100 runs. When $D > 0$, the cumulative regrets of ME−CUCB1 with the true variance (blue) is significantly smaller than that of C^{2}UCB (orange) ignoring the correlation. The cumulative regrets of ME−CUCB2 (green) gives similar values as ME−CUCB1. For the case $D = 0$ all three algorithms have almost the same cumulative regret and estimating D and σ^2 does not harm the performance even if there is no correlation.

Random Coefficient Model We generate outcomes from the random coefficient model $Y_{ti} = X_{ti}(\beta+\gamma_{ti})+e_{ti}$, where $\gamma_{ti} \sim \mathcal{N}(0, D)$ for a covariance matrix D. We use the same reward function, (d, K, m), contexts and fixed effect β as in random intercept model. The value of D is chosen from $\{I_d, 2I_d, 5I_d\}$. We run $c = 20$ exploration rounds and use EM algorithm for ME−CUCB2 to estimate $\widehat{D}_t$ and $\widehat{\sigma}_t^2$. The lower row of the Figure 1 presents the mean cumulative regret for $T = 1000$ over 100 experiments. For all three values of D, the ME−CUCB1 and ME−CUCB2 are compatible and significantly outperform C^{2}UCB.

Real-World Data: MovieLens Dataset

The MovieLens 10M dataset (Harper and Konstan 2015) contains 10 million triplets of users, movies, and the ratings from 0 to 5 across 10,681 movies. We split the dataset into train/test sets by 8:2. The data construction process for

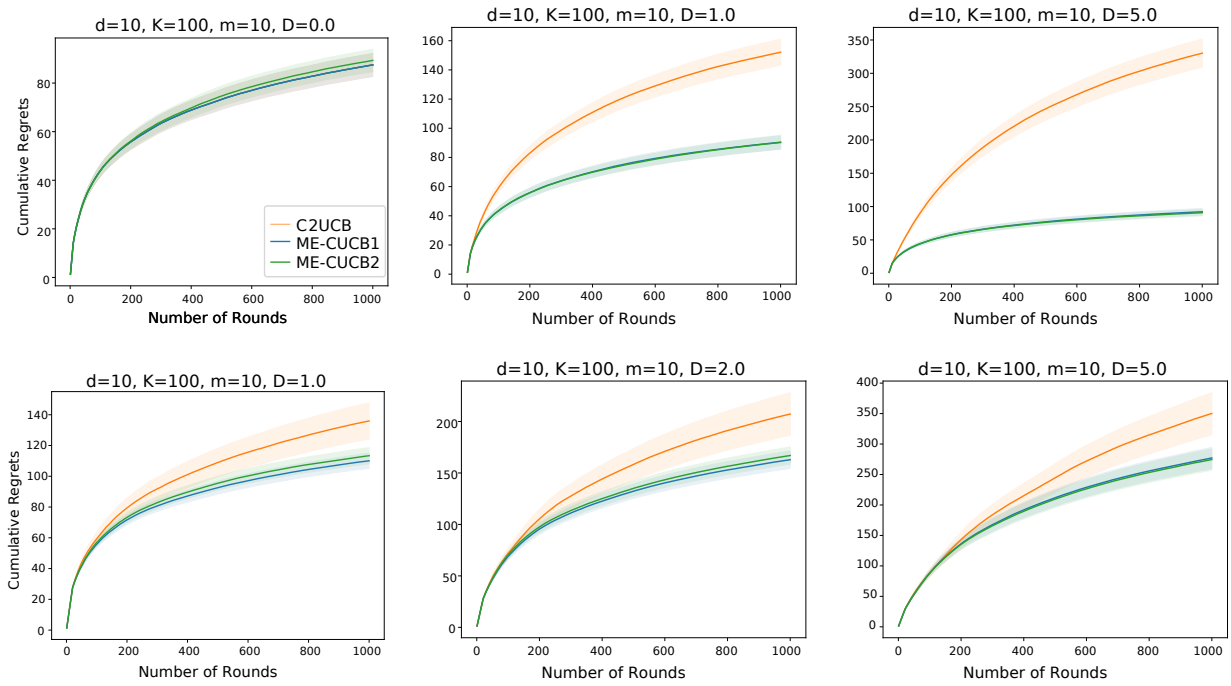

Figure 1: Comparison of the average cumulative regrets on synthetic dataset from random intercept model (upper) and random coefficient model (lower) over 100 repeated runs with $T = 1000$. The shaded areas show the 95% confidence interval.

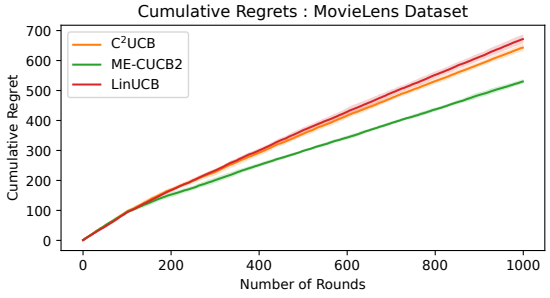

Figure 2: Comparison of the average cumulative regret on MovieLens 10M dataset over ten repeated runs with $T = 1000$. The shaded areas show the 95% confidence interval.

the experiment is as follows. We apply probabilistic matrix factorization (PMF, Mnih and Salakhutdinov (2007)) to the training set to extract 10-dimensional contexts for each movie, and gather data that contains user, movie, movie context vectors, and corresponding ratings. To simulate the recommendation of bundles of movies to an individual in each round, we sample a single user and then select $K \times m = 250$ movie ratings from that user, to create $K = 50$ arms each containing $m = 5$ movies with context matrix in $\mathbb{R}^{m \times d}$. Although the individual movies are not inherently related to each other, the ratings are provided by a single user. This results in a shared unobserved random effect specific to that user, leading to correlations among the ratings. Consequently, the collection of movie ratings is regarded as correlated since they are generated from the same user in each round. The reward function f is given as the average of the ratings. We evaluate and compare the proposed `ME-CUCB2` with the benchmark models `C`^{2}`UCB` and `LinUCB`. For `LinUCB`, we approach the problem as maximizing the scalar reward given by the average rating, as in the original linear bandit problem, since we utilize the average function for f. The algorithm `ME-CUCB1` cannot be applied here since the true covariance structure is unknown. For `ME-CUCB2`, we assume a random intercept model. Figure 2 displays the cumulative regret of each algorithm, and `ME-CUCB2` estimating the covariance matrix achieves significantly smaller regrets compared to the baseline algorithms.

Conclusion

We address a framework called mixed-effects bandit that can effectively handle correlations between multiple outcomes. We developed an efficient algorithm `ME-CUCB`, solving the mixed-effects bandit problem, for both the cases where the covariance matrices are known and unknown. The proposed algorithm achieves a regret bound of $\tilde{O}(d\sqrt{mT})$, matching the lower bound under the problem setting. To bound the error terms that do not form a martingale, a novel covering number method has been employed. Empirical evaluation on synthetic and public MovieLens dataset supports the theoretical claims and demonstrates that the algorithm outperforms existing methods in practice. Overall, this work achieves both provable near-optimality and practicality for the mixed-effects bandit problem.

Acknowledgments

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT, No.2020R1A2C1A01011950) and by the Institute of Information & communications Technology Planning & evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2020-0-01336, Artificial Intelligence Graduate School Program (UNIST); No. 2022-0-00469, Development of Core Technologies for Task-oriented Reinforcement Learning for Commercialization of Autonomous Drones; No.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)), and by Creative-Pioneering Researchers Program through Seoul National University.

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Agrawal, S.; and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, 127–135. PMLR.

Aouali, I.; Kveton, B.; and Katariya, S. 2023. Mixed-effect thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, 2087–2115. PMLR.

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

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2): 235–256.

Busa-Fekete, R.; Szörényi, B.; Weng, P.; and Mannor, S. 2017. Multi-objective bandits: Optimizing the generalized gini index. In *International Conference on Machine Learning*, 625–634. PMLR.

Chen, W.; Wang, Y.; and Yuan, Y. 2013. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, 151–159. PMLR.

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

de la Pena, V. H.; Klass, M. J.; and Lai, T. L. 2004. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability*, 1902–1933.

Demidenko, E. 2013. *Mixed models: theory and applications with R*. John Wiley & Sons.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22.

Drugan, M. M.; and Nowe, A. 2013. Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Harper, F. M.; and Konstan, J. A. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4): 1–19.

Hong, J.; Kveton, B.; Katariya, S.; Zaheer, M.; and Ghavamzadeh, M. 2022a. Deep Hierarchy in Bandits. *arXiv preprint arXiv:2202.01454*.

Hong, J.; Kveton, B.; Zaheer, M.; and Ghavamzadeh, M. 2022b. Hierarchical bayesian bandits. In *International Conference on Artificial Intelligence and Statistics*, 7724–7741. PMLR.

Jiang, J. 2017. *Asymptotic analysis of mixed effects models: theory, applications, and open problems*. CRC press.

Karumur, R. P.; Nguyen, T. T.; and Konstan, J. A. 2016. Exploring the value of personality in predicting rating behaviors: a study of category preferences on movielens. In *Proceedings of the 10th ACM conference on recommender systems*, 139–142.

Kveton, B.; Konobeev, M.; Zaheer, M.; Hsu, C.-w.; Mladenov, M.; Boutilier, C.; and Szepesvari, C. 2021. Meta-thompson sampling. In *International Conference on Machine Learning*, 5884–5893. PMLR.

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

Laird, N. M.; and Ware, J. H. 1982. Random-effects models for longitudinal data. *Biometrics*, 963–974.

Lattimore, T.; Crammer, K.; and Szepesvári, C. 2015. Linear multi-resource allocation with semi-bandit feedback. *Advances in Neural Information Processing Systems*, 28.

Lennihan, L.; Mayer, S. A.; Fink, M. E.; Beckford, A.; Paik, M. C.; Zhang, H.; Wu, Y.-C.; Klebanoff, L. M.; Raps, E. C.; and Solomon, R. A. 2000. Effect of hypervolemic therapy on cerebral blood flow after subarachnoid hemorrhage: a randomized controlled trial. *Stroke*, 31(2): 383–391.

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

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

Liang, K.-Y.; and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1): 13–22.

Lu, S.; Wang, G.; Hu, Y.; and Zhang, L. 2019. Multi-objective generalized linear bandits. *arXiv preprint arXiv:1905.12879*.

Mnih, A.; and Salakhutdinov, R. R. 2007. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20.

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

Tekin, C.; and Turgay, E. 2017. Multi-objective contextual bandits with a dominant objective. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. IEEE.

Turgay, E.; Oner, D.; and Tekin, C. 2018. Multi-objective contextual bandit problem with similarity information. In *International Conference on Artificial Intelligence and Statistics*, 1673–1681. PMLR.

Wan, R.; Ge, L.; and Song, R. 2021. Metadata-based multi-task bandits with bayesian hierarchical models. *Advances in Neural Information Processing Systems*, 34: 29655–29668.

Zhang, X.; Li, S.; and Liu, W. 2019. Contextual combinatorial conservative bandits. *arXiv preprint arXiv:1911.11337*.

Zhang, X.; Zhou, Y.; Ma, Y.; Chen, B.-C.; Zhang, L.; and Agarwal, D. 2016. Glmix: Generalized linear mixed models for large-scale response prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 363–372.

Zhu, R.; and Kveton, B. 2022a. Random Effect Bandits. In *International Conference on Artificial Intelligence and Statistics*, 3091–3107. PMLR.

Zhu, R.; and Kveton, B. 2022b. Robust Contextual Linear Bandits. *arXiv preprint arXiv:2210.14483*.