CONTEXTUAL BANDITS WITH CONCAVE REWARDS, AND AN APPLICATION TO FAIR RANKING

Virginie Do
PSL University & Meta AI
virginiedo@meta.com

Elvis Dohmatob, Matteo Pirotta, Alessandro Lazaric, Nicolas Usunier
Meta AI
{dohmatob,pirotta,lazaric,usunier}@meta.com

ABSTRACT

We consider Contextual Bandits with Concave Rewards (CBCR), a multi-objective bandit problem where the desired trade-off between the rewards is defined by a known concave objective function, and the reward vector depends on an observed stochastic context. We present the first algorithm with provably vanishing regret for CBCR without restrictions on the policy space, whereas prior works were restricted to finite policy spaces or tabular representations. Our solution is based on a geometric interpretation of CBCR algorithms as optimization algorithms over the convex set of expected rewards spanned by all stochastic policies. Building on Frank-Wolfe analyses in constrained convex optimization, we derive a novel reduction from the CBCR regret to the regret of a *scalar-reward* bandit problem. We illustrate how to apply the reduction off-the-shelf to obtain algorithms for CBCR with both linear and general reward functions, in the case of non-combinatorial actions. Motivated by fairness in recommendation, we describe a special case of CBCR with rankings and fairness-aware objectives, leading to the first algorithm with regret guarantees for contextual combinatorial bandits with fairness of exposure.

1 INTRODUCTION

Contextual bandits are a popular paradigm for online recommender systems that learn to generate personalized recommendations from user feedback. These algorithms have been mostly developed to maximize a single scalar reward which measures recommendation performance for users. Recent fairness concerns have shifted the focus towards item producers whom are also impacted by the exposure they receive (Biega et al., 2018; Geyik et al., 2019), leading to optimize trade-offs between recommendation performance for users and fairness of exposure for items (Singh & Joachims, 2019; Zehlike & Castillo, 2020). More generally, there is an increasing pressure to insist on the multi-objective nature of recommender systems (Vamplew et al., 2018; Stray et al., 2021), which need to optimize for several engagement metrics and account for multiple stakeholders' interests (Mehrotra et al., 2020; Abdollahpouri et al., 2019). In this paper, we focus on the problem of contextual bandits with multiple rewards, where the desired trade-off between the rewards is defined by a known concave objective function, which we refer to as *Contextual Bandits with Concave Rewards* (CBCR). Concave rewards are particularly relevant to fair recommendation, where several objectives can be expressed as (known) concave functions of the (unknown) utilities of users and items (Do et al., 2021).

Our CBCR problem is an extension of Bandits with Concave Rewards (BCR) (Agrawal & Devanur, 2014) where the vector of multiple rewards depends on an observed stochastic context. We address this extension because contexts are necessary to model the user/item features required for personalized recommendation. Compared to BCR, the main challenge of CBCR is that optimal policies depend on the entire distribution of contexts and rewards. In BCR, optimal policies are distributions over actions, and are found by direct optimization in policy space (Agrawal & Devanur, 2014; Berthet & Perchet, 2017). In CBCR, stationary policies are mappings from a continuous context space to distributions over actions. This makes existing BCR approaches inapplicable to CBCR because the policy space is not amenable to tractable optimization without further assumptions or restrictions. As a matter of fact, the only prior theoretical work on CBCR is restricted to a finite policy set (Agrawal et al., 2016).

We present *the first algorithms with provably vanishing regret* for CBCR without restriction on the policy space. Our main theoretical result is a reduction where the CBCR regret of an algorithm is

bounded by its regret on a proxy bandit task with *single* (scalar) reward. This reduction shows that it is straightforward to turn *any* contextual (scalar reward) bandits into algorithms for CBCR. We prove this reduction by first re-parameterizing CBCR as an optimization problem in the space of feasible rewards, and then revealing connections between Frank-Wolfe (FW) optimization in reward space and a decision problem in action space. This bypasses the challenges of optimization in policy space.

To illustrate how to apply the reduction, we provide two example algorithms for CBCR with non-combinatorial actions, one for linear rewards based on LinUCB (Abbasi-Yadkori et al., 2011), and one for *general reward functions* based on the SquareCB algorithm (Foster & Rakhlin, 2020) which uses online regression oracles. In particular, we highlight that our reduction can be used together with any exploration/exploitation principle, while previous FW approaches to BCR relied exclusively on upper confidence bounds (Agrawal & Devanur, 2014; Berthet & Perchet, 2017; Cheung, 2019).

Since fairness of exposure is our main motivation for CBCR, we show how our reduction also applies to the *combinatorial* task of fair ranking with contextual bandits, leading to the *first algorithm with regret guarantees* for this problem, and we show it is *computationally efficient*. We compare the empirical performance of our algorithm to relevant baselines on a music recommendation task.

Related work. Agrawal et al. (2016) address a restriction of CBCR to a finite set of policies, where explicit search is possible. Cheung (2019) use FW for reinforcement learning with concave rewards, a similar problem to CBCR. However, they rely on a tabular setting where there are few enough policies to compute them explicitly. Our approach is the only one to apply to CBCR without restriction on the policy space, by removing the need for explicit representation and search of optimal policies.

Our work is also related to fairness of exposure in bandits. Most previous works on this topic either do not consider rankings (Celis et al., 2018; Wang et al., 2021; Patil et al., 2020; Chen et al., 2020), or apply to combinatorial bandits without contexts (Xu et al., 2021). Both these restrictions are impractical for recommender systems. Mansoury et al. (2021); Jeunen & Goethals (2021) propose heuristics with experimental support that apply to both ranking and contexts in this space, but they lack theoretical guarantees. We present the first algorithm with regret guarantees for fair ranking with contextual bandits. We provide a more detailed discussion of the related work in Appendix A.

2 Maximization of concave rewards in contextual bandits

Notation. For any $n \in \mathbb{N}$, we denote by $[\![n]\!] = \{1, \dots, n\}$. The dot product of two vectors x and y in $\mathbb{R}^n$ is either denoted $x^\intercal y$ or using braket notation $\langle x \,|\, y \rangle$, depending on which one is more readable.

Setting. We define a stochastic contextual bandit (Langford & Zhang, 2007) problem with D rewards. At each time step t, the environment draws a context $x_t \sim P$, where $x \in \mathcal{X} \subseteq \mathbb{R}^q$ and P is a probability measure over $\mathcal{X}$. The learner chooses an action $a_t \in \mathcal{A}$ where $\mathcal{A} \subseteq \mathbb{R}^K$ is the action space, and receives a noisy multi-dimensional reward $r_t \in \mathbb{R}^D$, with expectation $\mathbb{E}[r_t | x_t, a_t] = \mu(x_t) a_t$, where $\mu : \mathcal{X} \to \mathbb{R}^{D \times K}$ is the matrix-value contextual expected reward function.[1] The trade-off between the D cumulative rewards is specified by a known concave function $f : \mathbb{R}^D \to \mathbb{R} \cup \{\pm\infty\}$. Let $\overline{\mathcal{A}}$ denote the convex hull of $\mathcal{A}$ and $\overline{\pi} : \mathcal{X} \to \overline{\mathcal{A}}$ be a stationary policy,[2] then the optimal value for the problem is defined as $f^* = \sup_{\overline{\pi} : \mathcal{X} \to \overline{\mathcal{A}}} f\Big(\mathbb{E}_{x \sim P}\big[\mu(x)\overline{\pi}(x)\big]\Big)$.

We rely on either of the following assumptions on f:

Assumption A *f is closed proper concave*[3] *on $\mathbb{R}^D$ and $\mathcal{A}$ is a compact subset of $\mathbb{R}^K$. Moreover, there is a compact convex set $\mathcal{K} \subseteq \mathbb{R}^D$ such that*

- (Bounded rewards) $\forall (x, a) \in \mathcal{X} \times \mathcal{A}, \mu(x)a \in \mathcal{K}$ *and for all* $t \in \mathbb{N}_*$, $r_t \in \mathcal{K}$ *with probability* 1.
- (Local Lipschitzness) *f is L-Lipschitz continuous with respect to $\|.\|_2$ on an open set containing $\mathcal{K}$.*

Assumption B *Assumption A holds and f has C-Lipschitz-continuous gradients w.r.t. $\|.\|_2$ on $\mathcal{K}$.*

[1] Notice that linear structure between $\mu(x_t)$ and a_t is standard in combinatorial bandits (Cesa-Bianchi & Lugosi, 2012) and it reduces to the usual multi-armed bandit setting when $\mathcal{A}$ is the canonical basis of $\mathbb{R}^K$.

[2] In the multi-armed setting, stationary policies return a distribution over arms given a context vector. In the combinatorial setup, $\overline{\pi}(x) \in \overline{\mathcal{A}}$ is the average feature vector of a stochastic policy over $\mathcal{A}$. For the benchmark, we are only interested in expected rewards so there is to need to specify the full distribution over $\mathcal{A}$.

[3] This means that f is concave and upper semi-continuous, is never equal to $+\infty$ and is finite somewhere.

The most general version of our algorithm, described in Appendix D, removes the need for the smoothness assumption using smoothing techniques. We describe an example in Section 3.3. In the rest of the paper, we denote by $D_{\mathcal{K}} = \sup_{z,z' \in \mathcal{K}} \|z - z'\|_2$ the diameter of $\mathcal{K}$, and use $\tilde{C} = \frac{C}{2} D_{\mathcal{K}}^2$.

We now give two examples of this problem setting, motivated by real-world applications in recommender systems, and which satisfy Assumption A.

Example 1 (Optimizing multiple metrics in recommender systems.) *Mehrotra et al. (2020) formalized the problem of optimizing D engagement metrics (e.g. clicks, streaming time) in a bandit-based recommender system. At each t, x_t represents the current user's features. The system chooses one arm among K, represented by a vector a_t in the canonical basis of $\mathbb{R}^K$ which is the action space $\mathcal{A}$. Each entry of the observed reward vector $(r_{t,i})_{i=1}^D$ corresponds to a metric's value. The trade-off between the metrics is defined by the Generalized Gini Function: $f(z) = \sum_{i=1}^D w_i z_i^{\uparrow}$, where $(z_i^{\uparrow})_{i=1}^D$ denotes the values of z sorted increasingly and $w \in \mathbb{R}^D$ is a vector of non-increasing weights.*

Example 2 (Fairness of exposure in rankings.) *The goal is to balance the traditional objective of maximizing user satisfaction in recommender systems and the inequality of exposure between item producers (Singh & Joachims, 2018; Zehlike & Castillo, 2020). For a recommendation task with m items to rank, this leads to a problem with $D = m + 1$ objectives, which correspond to the m items' exposures, plus the user satisfaction metric. The context $x_t \in \mathcal{X} \subset \mathbb{R}^{md}$ is a matrix where each $x_{t,i} \in \mathbb{R}^d$ represents a feature vector of item i for the current user. The action space $\mathcal{A}$ is combinatorial, i.e. it is the space of rankings represented by permutation matrices:*

$$\mathcal{A} = \left\{ a \in \{0,1\}^{m \times m} : \forall i \in [\![m]\!], \sum_{k=1}^m a_{i,k} = 1 \text{ and } \forall k \in [\![m]\!], \sum_{i=1}^m a_{i,k} = 1 \right\} \tag{1}$$

For $a \in \mathcal{A}$, $a_{i,k} = 1$ if item i is at rank k. Even though we use a double-index notation and call a a permutation matrix*, we flatten a as a vector of dimension $K = m^2$ for consistency of notation.*

We now give a concrete example for f, which is concave as usual for objective functions in fairness of exposure (Do et al., 2021). It is inspired by Morik et al. (2020), who study trade-offs between average user utility and inequality[4] of item exposure:

$$f(z) = \underbrace{z_{m+1}}_{\text{user utility}} - \beta \underbrace{\frac{1}{2m} \sum_{i=1}^m \sum_{j=1}^m |z_i - z_j|}_{\text{inequality of item exposure}} \qquad \text{where } \beta > 0 \text{ is a trade-off parameter.} \tag{2}$$

The learning problem. In the bandit setting, P and μ are unknown and the learner can only interact online with the environment.Let $h_T = \big(x_t, a_t, r_t\big)_{t \in [\![T-1]\!]}$ be the history of contexts, actions, and reward observed up to time $T-1$ and $\delta' > 0$ be a confidence level, then at step t a bandit algorithm $\mathfrak{A}$ receives in input the history h_t, the current context x_t, and it returns a distribution over actions $\mathcal{A}$ and selects an action $a_t \sim \mathfrak{A}(h_t, x_t, \delta')$. The objective of the algorithm is to minimize the regret

$$R_T = f^* - f(\hat{s}_T) \qquad \text{where } \hat{s}_T = \frac{1}{T} \sum_{t=1}^T r_t. \tag{3}$$

Note that our setting subsumes classical stochastic contextual bandits: when $D = 1$ and $f(z) = z$, maximizing $f(\hat{s}_T)$ amounts to maximizing a cumulative scalar reward $\sum_{t=1}^T r_t$. In Lem. 9 (App. C.3), we show that alternative definitions of regret, with different choices of comparator or performance measure, would yield a difference of order $O(1/\sqrt{T})$, and hence not substantially change our results.

3 A GENERAL REDUCTION-BASED APPROACH FOR CBCR

In this section we describe our general approach for CBCR. We first derive our key reduction from CBCR to a specific scalar-reward bandit problem. We then instantiate our algorithm to the case of linear and general reward functions for smooth objectives f. Finally, we extend to the case of non-smooth objective functions using Moreau-Yosida regularization (Rockafellar & Wets, 2009).

[4] $\text{Gini}(z_1, \ldots, z_m) = \frac{1}{2m} \sum_{i=1}^m \sum_{j=1}^m |z_i - z_j|$ is an unnormalized Gini coefficient.

3.1 REDUCTION FROM CBCR TO SCALAR-REWARD CONTEXTUAL BANDITS

There are two challenges in the CBCR problem: 1) the computation of the optimal policy $\sup_{\overline{\pi}:\mathcal{X}\to\overline{\mathcal{A}}} f\Big(\mathbb{E}_{x\sim P}\big[\mu(x)\overline{\pi}(x)\big]\Big)$ even with known μ; 2) the learning problem when μ is unknown.

1: Reparameterization of the optimization problem. The first challenge is that optimizing directly in policy space for the benchmark problem $\sup_{\overline{\pi}:\mathcal{X}\to\overline{\mathcal{A}}} f\Big(\mathbb{E}_{x\sim P}\big[\mu(x)\overline{\pi}(x)\big]\Big)$ is intractable without any restriction, because the policy space includes all mappings from the continuous context space $\mathcal{X}$ to distributions over actions. Our solution is to rewrite the optimization problem as a standard convex constrained problem by introducing the convex set $\mathcal{S}$ of feasible rewards:

$$\mathcal{S} = \left\{\mathbb{E}_{x\sim P}\big[\mu(x)\overline{\pi}(x)\big] \,\middle|\, \overline{\pi}:\mathcal{X}\to\overline{\mathcal{A}}\right\} \text{ so that } f^* = \sup_{\overline{\pi}:\mathcal{X}\to\overline{\mathcal{A}}} f\Big(\mathbb{E}_{x\sim P}\big[\mu(x)\overline{\pi}(x)\big]\Big) = \max_{s\in\mathcal{S}} f(s).$$

Under Assumption A, $\mathcal{S}$ is a compact subset of $\mathcal{K}$ (see Lemma 7 in App. C) so f attains its maximum over $\mathcal{S}$. We have thus reduced the complex initial optimization problem to a concave optimization problem over a compact convex set.

2: Reducing the learning problem to scalar-reward bandits. Unfortunately, since P and μ are unknown, the set $\mathcal{S}$ is unknown. This precludes the possibility of directly using standard constrained optimization techniques, including gradient descent with projections onto $\mathcal{S}$. We consider Frank-Wolfe, a projection-free optimization method robust to approximate gradients (Lacoste-Julien et al., 2013; Kerdreux et al., 2018). At each iteration t of FW, the update direction is given by the linear subproblem: $\operatorname{argmax}_{s\in\mathcal{S}}\langle\nabla f(z_{t-1})\,|\,s\rangle$, where z_{t-1} is the current iterate. Our main technical tool, Lemma 1, allows to connect the FW subproblem in the unknown reward space $\mathcal{S}$ to a workable decision problem in the action space (see Lemma 13 in Appendix E for a proof):

Lemma 1 *Let $\mathbb{E}_t\big[.\big]$ be the expectation conditional on h_t. Let $z_t \in \mathcal{K}$ be a function of contexts, actions and rewards up to time t. Under Assumption A, we have:*

$$\forall t\in\mathbb{N}_*, \mathbb{E}_t\big[\max_{a\in\mathcal{A}}\langle\nabla f(z_{t-1})\,|\,\mu(x_t)a\rangle\big] = \max_{s\in\mathcal{S}}\langle\nabla f(z_{t-1})\,|\,s\rangle. \tag{4}$$

For all $\delta\in(0,1]$, with probability at least $1-\delta$, we have:

$$\sum_{t=1}^{T}\Big(\max_{s\in\mathcal{S}}\langle\nabla f(z_{t-1})\,|\,s\rangle - \max_{a\in\mathcal{A}}\langle\nabla f(z_{t-1})\,|\,\mu(x_t)a\rangle\Big) \le LD_{\mathcal{K}}\sqrt{2T\ln(\delta^{-1})}. \tag{5}$$

Lemma 1 shows that FW for CBCR operates closely to a sequence of decision problems of the form $(\max_{a\in\mathcal{A}}\langle\nabla f(z_{t-1})\,|\,\mu(x_t)a\rangle)_{t=1}^T$. However, we have yet to address the problem that P and μ are unknown. To solve this issue, we introduce a **reduction to scalar-reward contextual bandits**. We can notice that solving for the sequence of actions maximizing $\sum_{t=1}^T\langle\nabla f(z_{t-1})\,|\,\mu(x_t)a\rangle$ corresponds to solving a contextual bandit problem with adversarial contexts and stochastic rewards. Formally, using $z_t = \hat{s}_t$[5], we define the extended context $\tilde{x}_t = (\nabla f(\hat{s}_{t-1}), x_t)$, the average scalar reward $\tilde{\mu}(\tilde{x}_t) = \nabla f(\hat{s}_{t-1})^\intercal\mu(x_t)$ and the observed scalar reward $\tilde{r}_t = \langle\nabla f(\hat{s}_{t-1})\,|\,r_t\rangle$. This fully defines a contextual bandit problem with *scalar reward*. Then, the objective of the algorithm is to minimize the following *scalar regret*:

$$R_T^{\text{scal}} = \sum_{t=1}^T\max_{a\in\mathcal{A}}\tilde{\mu}(\tilde{x}_t)^\intercal a - \sum_{t=1}^T\tilde{r}_t = \sum_{t=1}^T\max_{a\in\mathcal{A}}\langle\nabla f(\hat{s}_{t-1})\,|\,\mu(x_t)a\rangle - \sum_{t=1}^T\langle\nabla f(\hat{s}_{t-1})\,|\,r_t\rangle. \tag{6}$$

In this framework, the only information observed by the learning algorithm is $\tilde{h}_t := \big(\tilde{x}_{t'}, a_{t'}, \tilde{r}_{t'}\big)_{t'\in[\![t-1]\!]}$. This regret minimization problem has been extensively studied (see e.g., Slivkins, 2019, Chap. 8 for an overview). The following key reduction result[6] relates R_T^{scal} to R_T, the regret of the original CBCR problem:

[5]For simplicity, we presented our reduction with $z_t = \hat{s}_t$ but other choices of z_t are possible (see Appendix D). The important point is that the reduction works without restricting z_t to $\mathcal{S}$.

[6]In practice, this result is used in conjunction with an upper bound $\overline{R}^{\text{scal}}(T,\delta')$ on R_T^{scal} that holds with probability $\ge 1-\delta'$, which gives $R_T \le \overline{R}^{\text{scal}}(T,\delta')/T + O(\sqrt{\ln(1/\delta)/T})$ with probability at least $1-\delta-\delta'$ using the union bound.

Theorem 2 *Under Assmpt. B, for every $T \in \mathbb{N}_*$ and $\delta > 0$, algorithm $\mathfrak{A}$ satisfies, with prob. $\geq 1-\delta$:*

$$R_T = f^* - f(\hat{s}_T) \leq \frac{R_T^{\text{scal}} + LD_{\mathcal{K}}\sqrt{2T\ln(1/\delta)} + \tilde{C}\ln(eT)}{T}. \tag{7}$$

The reduction shown in Thm. 2 hints us at how to use or adapt scalar bandit algorithms for CBCR. In particular, any algorithm with sublinear regret will lead to a vanishing regret for CBCR. Since the worst-case regret of contextual bandits is $\Omega(\sqrt{T})$ (Dani et al., 2008), we obtain near minimax optimal algorithms for CBCR. We illustrate this with two algorithms derived from our reduction in Sec. 3.2.

Proof sketch of Theorem 2: CBCR *and Frank-Wolfe algorithms (full proof in Appendix E).*
Although the set $\mathcal{S}$ is not known, the standard telescoping sum argument for the analysis of Frank-Wolfe algorithms (see Lemma 14 in Appendix E, and e.g., (Berthet & Perchet, 2017, Lemma 12) for similar derivations) gives that under Assumption B, denoting $g_t = \nabla f(\hat{s}_{t-1})$:

$$TR_T \leq \sum_{t=1}^{T} \max_{s\in\mathcal{S}} \langle g_t \,|\, s - r_t\rangle + \tilde{C}\ln(eT).$$

The result is true for every sequence $(r_t)_{t\in[\![T]\!]} \in \mathcal{K}^T$, and only tracks the trajectory of $\hat{s}_t$ in reward space. We introduce now the reference of the scalar regret:

$$TR_T = \sum_{t=1}^{T}\big(\max_{s\in\mathcal{S}}\langle g_t \,|\, s\rangle - \max_{a\in\mathcal{A}}\langle g_t \,|\, \mu(x_t)a\rangle\big) + \underbrace{\sum_{t=1}^{T}\max_{a\in\mathcal{A}}\langle g_t \,|\, \mu(x_t)a - r_t\rangle}_{=R_T^{\text{scal}}} + \tilde{C}\ln(eT) \tag{8}$$

Lemma 1 bounds the leftmost term, from which Theorem 2 immediately follows using (8). □

3.2 Practical application: Two algorithms for multi-armed CBCR

To illustrate the effectiveness of the reduction from CBCR to scalar-reward bandits, we focus on the case where the action space $\mathcal{A}$ is the canonical basis of $\mathbb{R}^K$ (as in Example 1). We first study the case of linear rewards. Then, for general reward functions, we introduce the FW-SquareCB algorithm, the first example of a FW-based approach combined with an exploration principle other than optimism. This shows our approach has a much broader applicability to solve (C)BCR than previous strategies.

From LinUCB to FW-LinUCB (details in Appendix G). We consider a CBCR with linear reward function, i.e., $\mu(x) = \theta x$ where $\theta \in \mathbb{R}^{D\times d}$ (recall we have D rewards) and $x \in \mathbb{R}^{d\times K}$, where d is the number of features. Let $\tilde{\theta} := \text{flatten}(\theta)$ and $g_t = \nabla f(\hat{s}_{t-1})$. Using $[.;.]$ to denote the vertical concatenation of matrices, the expected reward for action a in context x at time t can be written $\langle g_t \,|\, \mu(x)a\rangle = g_t^\intercal \theta x a = \langle \tilde{\theta} \,|\, \tilde{x}_t a\rangle$ where $\tilde{x}_t \in \mathbb{R}^{Dd\times K}$ is the *extended* context with entries $\tilde{x}_t = [g_{t,0}x_t; \dots; g_{t,D}x_t] \in \mathbb{R}^{Dd\times K}$. This is an instance of a linear bandit problem, where at each time t, action a is associated to the vector $\tilde{x}_t a$ and its expected reward is $\langle \tilde{\theta} \,|\, \tilde{x}_t a\rangle$. As a result, we can immediately derive a LinUCB-based algorithm for linear CBCR by leveraging the equivalence FW-LinUCB$(h_t, x_t, \delta') =$ LinUCB$(\tilde{h}_t, \tilde{x}_t, \delta')$. LinUCB's regret guarantees imply $R_T^{\text{scal}} = O(d\sqrt{T})$ with high probability, which, in turn give a $O(1/\sqrt{T})$ for R_T.

From SquareCB to FW-SquareCB (details in Appendix H). We now consider a CBCR with general reward function $\mu(x)$. The SquareCB algorithm (Foster & Rakhlin, 2020) is a randomized exploration strategy that delegates the learning of rewards to an arbitrary online regression algorithm. The scalar regret of SquareCB is bounded depending on the regret of the base regression algorithm.

For FW-SquareCB, we have access to an online regression oracle $\hat{\mu}_t$, an estimate of μ which is a function of h_t, which has regression regret bounded by $R_{\text{oracle}}(T)$. The exploration strategy of FW-SquareCB follows the same principles as SquareCB: let $g_t = \nabla f(\hat{s}_{t-1})$ and denote $\underline{\hat{\mu}}_t = g_t^\intercal \hat{\mu}_t(x_t)$, so that $\underline{\hat{\mu}}_t^\intercal a = \langle g_t \,|\, \hat{\mu}_t(x_t)a\rangle$. Let $\mathfrak{A}_t =$ FW-SquareCB(h_t, x_t, δ') defined as

$$\forall a \in \mathcal{A}, \mathfrak{A}_t(a) = \begin{cases} \frac{1}{K+\gamma_t\left(\underline{\hat{\mu}}_t^* - \underline{\hat{\mu}}_t^\intercal a\right)} & \text{if } a \neq \underline{a}_t \\ 1 - \sum_{\substack{a\in\mathcal{A}\\ a\neq\underline{a}_t}} \mathfrak{A}_t(a) & \text{if } a = \underline{a}_t \end{cases} \quad \text{where } \underline{a}_t \in \operatorname*{argmax}_{a\in\mathcal{A}} \underline{\hat{\mu}}_t^\intercal a \text{ and } \underline{\hat{\mu}}_t^* = \underline{\hat{\mu}}_t \underline{a}_t$$

Then FW-SquareCB has R_T in $O(\sqrt{R_{\text{oracle}}(T)}/\sqrt{T})$ with high probability.

Table 1: Regret bounds depending on assumptions and base algorithm $\mathfrak{A}$, for multi-armed bandits with K arms (in dimension d for LinUCB). See Appendix G and H for the full details.

Algorithm (FW-<bandit>)	Assumptions (informal)	Bound on R_T (simplified, using $\delta' = \delta$)
FW-LinUCB	$\mu(x)a = \theta x a$ for $\theta \in \mathbb{R}^{D\times d}, x \in \mathbb{R}^{d\times K}$	$\frac{LD_{\mathcal{K}} dD \ln\left((1+\frac{TLD_{\mathcal{K}}}{dD})/\delta\right)}{\sqrt{T}}$
FW-SquareCB	$\sum_{t=1}^{T} \lVert \hat{\mu}_t(x_t)a_t - \mu(x_t)a_t \rVert_2^2 \leq R_{\text{oracle}}(T)$	$\frac{L\sqrt{K\left(R_{\text{oracle}}(T) + D_{\mathcal{K}}^2 \ln(T/\delta)\right)}}{\sqrt{T}}$

3.3 The case of nonsmooth f

When f is nonsmooth, we use a smoothing technique where the scalar regret is not measured using $\nabla f(\hat{s}_{t-1})$, but rather using gradients of a sequence $(f_t)_{t\in\mathbb{N}}$ of smooth approximations of f, whose smoothness decrease over time (see e.g., Lan, 2013, for applications of smoothing to FW). We provide a comprehensive treatment of smoothing in our general approach described in Appendix D, while specific smoothing techniques are discussed in Appendix F.

We now describe the use of Moreau-Yosida regularization (Rockafellar & Wets, 2009, Def. 1.22): $f_t(z) = \max_{y\in\mathbb{R}^D}\left(f(y) - \frac{\sqrt{t+1}}{2\beta_0}\|y-z\|_2^2\right)$. It is well-known that f_t is concave and L-Lipschitz whenever f is, and f_t is $\frac{\sqrt{t+1}}{\beta_0}$-smooth (see Lemma 15 in Appendix F). A related smoothing method was used by Agrawal & Devanur (2014) for (non-contextual) BCR. Our treatment of smoothing is more systematic than theirs, since we use a smoothing factor $\beta_0/\sqrt{t+1}$ that decreases over time rather than a fixed smoothing factor that depends on a pre-specified horizon. Our regret bound for CBCR is based on a scalar regret $R_T^{\text{scal,sm}}$ where $\nabla f_{t-1}(\hat{s}_{t-1})$ is used instead of $\nabla f(\hat{s}_{t-1})$:

$$R_T^{\text{scal,sm}} = \sum_{t=1}^{T} \max_{a\in\mathcal{A}} \langle \nabla f_{t-1}(\hat{s}_{t-1}) \,|\, \mu(x_t)a\rangle - \sum_{t=1}^{T} \langle \nabla f_{t-1}(\hat{s}_{t-1}) \,|\, r_t\rangle. \tag{9}$$

Theorem 3 *Under Assumptions A, for every $z_0 \in \mathcal{K}$, every $T \geq 1$ and every $\delta > 0, \delta' > 0$, Algorithm $\mathfrak{A}$ satisfies, with probability at least $1-\delta-\delta'$:*

$$R_T \leq \frac{R_T^{\text{scal,sm}}}{T} + \frac{LD_{\mathcal{K}}}{\sqrt{T}}\left(\frac{D_{\mathcal{K}}}{L\beta_0} + 3\frac{L\beta_0}{D_{\mathcal{K}}} + \sqrt{2\ln\frac{1}{\delta}}\right). \tag{10}$$

The proof is given in Appendix F. Taking $\beta_0 = \frac{D_{\mathcal{K}}}{L}$ leads to a simpler bound where $\frac{D_{\mathcal{K}}}{L\beta_0} + 3\frac{L\beta_0}{D_{\mathcal{K}}} = 4$.

4 Contextual ranking bandits with fairness of exposure

In this section, we apply our reduction to the combinatorial bandit task of fair ranking, and obtain the first algorithm with regret guarantees in the contextual setting. This task is described in Example 2 (Sec. 2). We remind that there is a fixed set of m items to rank at each timestep t, and that actions are flattened permutation matrices ($\mathcal{A}$ is defined in Ex. 2, Eq. (1)). The context $x_t \sim P$ is a matrix $x_t = (x_{t,i})_{i\in[\![m]\!]}$ where each $x_{t,i} \in \mathbb{R}^d$ represents a feature vector of item i for the current user.

Observation model. The *user utility* $u(x_t)$ is given by a position-based model with position weights $b(x_t) \in [0,1]^m$ and expected value for each item $v(x_t) \in [0,1]^m$. Denoting $u(x_t)$ the flattened version of $v(x_t)b(x_t)^\intercal \in \mathbb{R}^{m\times m}$, the user utility is (Lagrée et al., 2016; Singh & Joachims, 2018):

$$\langle u(x_t) \,|\, a\rangle = \sum_{i=1}^{m} v_i(x_t) \sum_{k=1}^{m} a_{i,k} b_k(x_t).$$

In this model, $b_k(x_t) \in [0,1]$ is the probability that the user observes the item at rank k. The quantity $\sum_{k=1}^{m} a_{i,k} b_k(x_t)$ is thus the probability that the user observes item i given ranking a. We denote

Algorithm 1: FW-LinUCBRank: linear contextual bandits for fair ranking.

input : $\delta' > 0, \lambda > 0, \hat{s}_0 \in \mathcal{K}$ $V_0 = \lambda \boldsymbol{I}_d, y_0 = \mathbf{0}_d, \hat{\theta}_0 = \mathbf{0}_d$

for $t = 1, \dots$ **do**

  Observe context $x_t \sim P$

  $\forall i, \hat{v}_{t,i} \leftarrow \hat{\theta}_{t-1}^\intercal x_{t,i} + \alpha_t\big(\frac{\delta'}{3}\big)\|x_{t,i}\|_{V_{t-1}^{-1}}$ // UCB on $v_i(x_t)$ (def. of α_t in Lem. 26, App. I)

  $a_t \leftarrow$ top-$\overline{k}\{\frac{\partial f}{\partial z_{m+1}}(\hat{s}_{t-1})\hat{v}_{t,i} + \frac{\partial f}{\partial z_i}(\hat{s}_{t-1})\}_{i=1}^m$ // FW linear optimization step

  Observe exposed items $e_t \in \{0,1\}^m$ and user feedback $c_t \in \{0,1\}^m$

  Update $\hat{s}_t \leftarrow \hat{s}_{t-1} + \frac{1}{t}(r_t - \hat{s}_{t-1})$

  $V_t \leftarrow V_{t-1} + \sum_{i=1}^m e_{t,i} x_{t,i} x_{t,i}^\intercal, \quad y_t \leftarrow y_{t-1} + \sum_{i=1}^m c_{t,i} x_{t,i}$ and $\hat{\theta}_t \leftarrow V_t^{-1} y_t$ // regression

end

$\overline{k} = \max_{x \in \mathcal{X}} \|b(x)\|_0 \leq m$ the maximum rank that can be exposed to any user. In most practical applications, $\overline{k} \ll m$. As formalized in Assumption D below, the position weights $b_k(x)$ are always non-increasing with k since the user browses the recommended items in order of their rank. We use a linear assumption for item values, where $D_{\mathcal{X}}$ and D_θ are known constants:

Assumption C $\sup_{x \in \mathcal{X}} \|x\|_2 \leq D_{\mathcal{X}}$ *and* $\exists \theta \in \mathbb{R}^d, \|\theta\|_2 \leq D_\theta$ *s.t.* $\forall x \in \mathcal{X}, \forall i \in [\![m]\!], v_i(x) = \theta^\intercal x_i$.

We propose an observation model where values $v_i(x)$ *and* position weights $b(x)$ are *unknown*. However, we assume that at each time step t, after computing the ranking a_t, we have two types of feedback: first, $e_{t,i} \in \{0,1\}$ is 1 if item i has been exposed to the user, and 0 otherwise. Second $c_{t,i} \in \{0,1\}$ which represents a binary `like`/`dislike` feedback from the user. We have

$$\mathbb{E}[e_{t,i} | x_t, a_t] = \sum_{k=1}^m a_{t,i,k} b_k(x_t) \qquad \mathbb{E}[c_{t,i} | x_t, e_{t,i}] = \begin{cases} v_i(x_t) & \text{if } e_{t,i} = 1 \\ 0 & \text{if } e_{t,i} = 0 \end{cases} \tag{11}$$

This observation model captures well applications such as newsfeed ranking on mobile devices or dating applications where only one post/profile is shown at a time. What we gain with this model is that $b(x)$ can *depend arbitrarily on the context* x, while previous work on bandits in the position-based model assumes b known and context-independent (Lagrée et al., 2016).[7]

Fairness of exposure. There are $D = m + 1$ rewards, i.e., $\mu(x) \in \mathbb{R}^{(m+1) \times m^2}$. Denoting $\mu_i(x)$ the ith-row of $\mu(x)$, seen as a column vector, each of the m first rewards is the exposure of a specific item, while the $m+1$-th reward is the user utility:

$$\forall i \in [\![m]\!], \langle \mu_i(x) \,|\, a \rangle = \sum_{k=1}^m a_{i,k} b_k(x) \qquad \text{and} \qquad \mu_{m+1}(x) = u(x) \tag{12}$$

The observed reward vector $r_t \in \mathbb{R}^D$ is defined by $\forall i \in [\![m]\!], r_{t,i} = e_{t,i}$ and $r_{t,m+1} = \sum_{i=1}^m c_{t,i}$. Notice that $\mathbb{E}[r_{t,m+1} | x_t] = u(x_t)$. Let $\mathcal{K}$ be the convex hull of $\{z \in \{0,1\}^{m+1} : \sum_{i=1}^m z_i \leq \overline{k}$ and $z_{m+1} \leq \sum_{i=1}^m z_i\}$, we have $D_{\mathcal{K}} \leq \sqrt{\overline{k}}\sqrt{\overline{k}+2} \leq \overline{k} + 1$ and $r_t \in \mathcal{K}$ with probability 1. The objective function $f : \mathbb{R}^D \to \mathbb{R}$ makes a trade-off between average user utility and inequalities in item exposure (we gave an example in Eq. (2)). The remaining assumptions of our framework are that the objective function is non-decreasing with respect to average user utility. This is not required but it is natural (see Example 2) and slightly simplifies the algorithm.

Assumption D *The assumptions of the framework described above hold, as well as Assumption B. Moreover,* $\forall z \in \mathcal{K}$ $\frac{\partial f}{\partial z_{m+1}}(z) > 0$, *and* $\forall x \in \mathcal{X}, 1 \geq b_1(x) \geq \dots \geq b_{\overline{k}}(x) = \dots = b_m(x) = 0$.

Algorithm and results. We present the algorithm in the setting of linear contextual bandits, using LinUCB (Abbasi-Yadkori et al., 2011; Li et al., 2010) as scalar exploration/exploitation algorithm in

[7] When b is unknown, depends on the context x, and we do not observe e_t, several approaches have been proposed to estimate the position weights (see e.g., Fang et al., 2019). Incorporating these approaches in contextual bandits for ranking is likely feasible but out of the scope of this work.

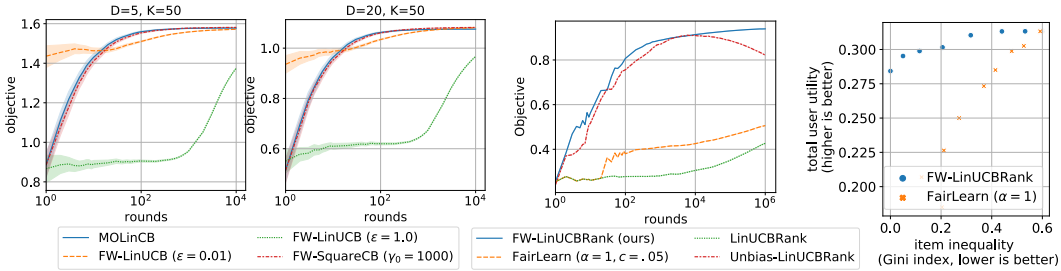

Figure 1: **(left)** Multi-armed CBCR: Objective values on environments from (Mehrotra et al., 2020). **(middle)** Ranking CBCR: Fairness objective value over timesteps on Last.fm data. **(right)** Ranking CBCR: Trade-off between user utility and item inequality after 5×10^6 iterations on Last.fm data.

Algorithm 1. It builds reward estimates based on Ridge regression with regularization parameter λ. As in the previous section, we focus on the case where f is smooth but the extension to nonsmooth f is straightforward, as described in Section 3. Appendix I provides the analysis for the general case.

As noted by Do et al. (2021), Frank-Wolfe algorithms are particularly suited for fair ranking in the position-based model. This is illustrated by line 4 of Alg. 1, where for $\tilde{u} \in \mathbb{R}^m$, top-$\overline{k}(\tilde{u})$ outputs a permutation (matrix) of $[\![m]\!]$ that sorts the top-$\overline{k}$ elements of $\tilde{u}$. Alg. 1 is thus *computationally fast*, with a cost dominated by the top-$\overline{k}$ sort. It also has an intuitive interpretation as giving items an adaptive bonus depending on ∇f (e.g., boosting the scores of items which received low exposure in previous steps). The following result is a consequence of (Do et al., 2021, Theorem 1):

Proposition 4 *Let $t \in \mathbb{N}_*$ and $\hat{\mu}_t$ such that $\forall i \in [\![m]\!], \hat{\mu}_{t,i} = \mu_i(x_t)$ and $\hat{\mu}_{t,m+1} = \hat{v}_t b(x_t)^\intercal$ viewed as a column vector, with $\hat{v}$ defined in line 3 of Algorithm 1. Then, under Assumption D, a_t defined on line 4 of Algorithm 1 satisfies: $\langle \nabla f(\hat{s}_{t-1}) \,|\, \hat{\mu}_t a_t \rangle = \operatorname*{argmax}_{a \in \mathcal{A}} \langle \nabla f(\hat{s}_{t-1}) \,|\, \hat{\mu}_t a \rangle$.*

The proposition says that even though computing a_t as in line 4 of Alg. 1 does not require the knowledge of $b(x_t)$, we still obtain the optimal update direction according to $\hat{\mu}_t$. Together with the usage of the observed reward r_t in FW iterates (instead of e.g., $\hat{\mu}_t a_t$ as would be done by Agrawal & Devanur (2014)), this removes the need for explicit estimates of $\mu(x_t)$. This is how our algorithm works *without knowing the position weights* $b(x_t)$, which are then allowed to depend on the context.

The usage of $\hat{v}_t$ to compute a_t follows the usual confidence-based approach to explore/exploitation principles for linear bandits, which leads to the following result (proven in Appendix I):

Theorem 5 *Under Assumptions B, C and D, for every $\delta' > 0$, every $T \in \mathbb{N}_*$, every $\lambda \geq D_{\mathcal{X}}^2 \overline{k}$, with probability at least $1 - \delta'$, Algorithm 1 has scalar regret bounded by*

$$R_T^{\text{scal}} = O\left(L\sqrt{T\overline{k}}\sqrt{d \ln(T/\delta')}\left(\sqrt{d \ln(T/\delta')} + D_\theta \sqrt{\lambda} + \sqrt{\overline{k}/d}\right)\right). \tag{13}$$

Thus, considering only $d, T, \overline{k}$ and $\delta = \delta'$ Alg. 1 has regret $R_T \leq O\left(\frac{d\overline{k}\ln(T/\delta)}{\sqrt{T}}\right)$ w.p. at least $1 - \delta$.

5 Experiments

We present two experimental evaluations of our approach, which are fully detailed in App. B.

5.1 Multi-armed CBCR: Application to multi-objective bandits

We first focus on the multi-objective recommendation task of Example 1 where $f(z) = \sum_{i=1}^D w_i z_i^{\uparrow}$.

Algorithms. We evaluate our two instantiations presented in Sec. 3.2 with the Moreau-Yosida smoothing technique of Sec. 3.3: (i) FW-SquareCB with Ridge regression and (ii) FW-LinUCB, where exploration is controlled by a scaling variable ϵ on the exploration bonus of each arm. We compare them to MOLinCB from (Mehrotra et al., 2020).

Environments. We reproduce the synthetic environments of Mehrotra et al. (2020), where the context and reward parameters are generated randomly, and $w_i = \frac{1}{2^{i-1}}$. We set $K = 50$ and $D \in \{5, 20\}$ (we also vary K in App. B). Each simulation is repeated with 100 random seeds.

Results. Following (Mehrotra et al., 2020), we evaluate the algorithms' performance by measuring the value of $f(\frac{1}{T}\sum_{t=1}^{T}\mu(x_t)a_t)$ over time. Our results are shown in Figure 1 (left). We observe that our algorithm FW-SquareCB obtains comparable performance with the baseline MOLinCB. These algorithms converge after ≈ 100 rounds. In this environment from (Mehrotra et al., 2020), only little exploration is needed, hence FW-LinUCB obtains better performance when ϵ is smaller ($\epsilon = 0.01$). The advantage of using an FW instantiation for the multi-objective bandit optimization task is that unlike MOLinCB, its convergence is also supported by our theoretical regret guarantees.

5.2 Ranking CBCR: Application to fairness of exposure in rankings

We now tackle the ranking problem of Section 4. We show how FW-LinUCBRank allows to fairly distribute exposure among items on a music recommendation task with bandit user feedback.

Environment. Following (Patro et al., 2020), we use the Last.fm music dataset from (Cantador et al., 2011), from which we extract the top 50 users and items with the most listening counts. We use a protocol similar to Li et al. (2016) to generate context and rewards from those. We use $\bar{k} = 10$ ranking slots, and exposure weights $b_k(x) = \frac{\log(2)}{1+\log(k)}$. Simulations are repeated with 10 seeds.

Algorithms. Our algorithm is FW-LinUCBRank with the nonsmooth objective f of Eq. (2), which trades off between user utility and item inequality. We study other fairness objectives in App. B. Our first baseline is *LinUCBRank* (Ermis et al., 2020), designed for ranking without fairness. Then, we study two baselines with amortized fairness of exposure criteria. Mansoury et al. (2021) proposed a fairness module for UCB-based ranking algorithms, which we plug into *LinUCBRank*. We refer to this baseline as *Unbiased-LinUCBRank*. Finally, the *FairLearn*(c, α) algorithm (Patil et al., 2020) enforces as fairness constraint that the pulling frequency of each arm be $\geq c$, up to a tolerance α. We implement as third baseline a simple adaptation of *FairLearn* to contextual bandits and ranking.

Dynamics. Figure 1 (middle) represents the values of f over time achieved by the competing algorithms, for fixed $\beta = 1$. As expected, compared to the fairness-aware and -unaware baselines, our algorithm FW-LinUCBRank reaches the best values of f. Interestingly, *Unbiased-LinUCBRank* also obtains high values of f on the first 10^4 rounds, but its performance starts decreasing after more iterations. This is because *Unbiased-LinUCBRank* is not guaranteed to converge to an optimal trade-off between user fairness and item inequality.

At convergence. We analyse the trade-offs achieved after $5 \cdot 10^6$ rounds between user utility and item inequality measured by the Gini index. We vary β in the objective f of Eq. (2) for FW-LinUCBRank and the strength c in *FairLearn*(c, α), with tolerance $\alpha = 1$. In Fig. 1 (right), we observe that compared to *FairLearn*, FW-LinUCBRank converges to much higher user utility at all levels of inequality among items. In particular, it achieves zero-unfairness at little cost for user utility.

6 Conclusion

We presented the first general approach to contextual bandits with concave rewards. To illustrate the usefulness of the approach, we show that our results extend randomized exploration with generic online regression oracles to the concave rewards setting, and extend existing ranking bandit algorithms to fairness-aware objective functions. The strength of our reduction is that it can produce algorithms for CBCR from any contextual bandit algorithm, including recent extensions of SquareCB to infinite compact action spaces (Zhu & Mineiro, 2022; Zhu et al., 2022) and future ones.

In our main application to fair ranking, the designer sets a fairness trade-off f to optimize. An interesting open problem is the integration of recent elicitation methods for f (e.g., Lin et al., 2022) in the bandit setting. Another interesting issue is the generalization of our framework to include constraints (Agrawal & Devanur, 2016). Finally, we note that the deployment of our algorithms requires to carefully design the whole machine learning setup, including the specification of reward functions (Stray et al., 2021), the design of online experiments (Bird et al., 2016), while taking feedback loops into account (Bottou et al., 2013; Jiang et al., 2019; Dean & Morgenstern, 2022).

ACKNOWLEDGMENTS

The authors would like to thank Clément Vignac, Marc Jourdan, Yaron Lipman, Levent Sagun and the anonymous reviewers for their helpful comments.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Beyond personalization: Research directions in multistakeholder recommendation. *arXiv preprint arXiv:1905.01986*, 2019.

Alekh Agarwal, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. *Advances in Neural Information Processing Systems*, 24, 2011.

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006, 2014.

Shipra Agrawal, Nikhil R Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pp. 4–18. PMLR, 2016.

Quentin Berthet and Vianney Perchet. Fast rates for bandit optimization with upper-confidence frank-wolfe. *Advances in Neural Information Processing Systems*, 30, 2017.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2212–2220, 2019.

Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, pp. 405–414, 2018.

Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14 (11), 2013.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, and Shie Mannor. Multi-objective bandits: Optimizing the generalized gini index. In *International Conference on Machine Learning*, pp. 625–634. PMLR, 2017.

Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011, New York, NY, USA, 2011. ACM.

L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth K Vishnoi. An algorithmic framework to control bias in bandit-based personalization. *arXiv preprint arXiv:1802.08674*, 2018.

Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, pp. 181–190. PMLR, 2020.

Wang Chi Cheung. Regret minimization for reinforcement learning with vectorial feedback and complex objectives. *Advances in Neural Information Processing Systems*, 32, 2019.

Kenneth L Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):1–30, 2010.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

Sarah Dean and Jamie Morgenstern. Preference dynamics under personalized recommendations. *arXiv preprint arXiv:2205.13026*, 2022.

Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 275–284, 2020.

Virginie Do and Nicolas Usunier. Optimizing generalized gini indices for fairness in rankings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 737–747, 2022.

Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Two-sided fairness in rankings via lorenz dominance. *Advances in Neural Information Processing Systems*, 34, 2021.

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

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Beyza Ermis, Patrick Ernst, Yannik Stein, and Giovanni Zappella. Learning to rank in the position based model with bandit feedback. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2405–2412, 2020.

Zhichong Fang, Aman Agarwal, and Thorsten Joachims. Intervention harvesting for context-dependent examination-bias estimation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 825–834, 2019.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta. Adversarial attacks on linear contextual bandits. *Advances in Neural Information Processing Systems*, 33:14362–14373, 2020.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: the mean-field game viewpoint. *arXiv preprint arXiv:2106.03787*, 2021.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pp. 2221–2231, 2019.

Thomas P Hayes. A large-deviation inequality for vector-valued martingales. *Combinatorics, Probability and Computing*, 2005.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pp. 427–435. PMLR, 2013.

Olivier Jeunen and Bart Goethals. Top-k contextual bandits with equity of exposure. In *Fifteenth ACM Conference on Recommender Systems*, pp. 310–320, 2021.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 383–390, 2019.

Thomas Kerdreux, Fabian Pedregosa, and Alexandre d'Aspremont. Frank-wolfe with subsampling oracle. In *International Conference on Machine Learning*, pp. 2591–2600. PMLR, 2018.

Till Kletti, Jean-Michel Renders, and Patrick Loiseau. Introducing the expohedron for efficient pareto-optimal fairness-utility amortizations in repeated rankings. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 498–507, 2022.

Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International conference on machine learning*, pp. 767–776. PMLR, 2015.

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate frank-wolfe optimization for structural svms. In *International Conference on Machine Learning*, 2013.

Paul Lagrée, Claire Vernade, and Olivier Cappe. Multiple-play bandits in the position-based model. *Advances in Neural Information Processing Systems*, 29, 2016.

Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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.

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*, pp. 661–670, 2010.

Shuai Li, Baoxiang Wang, Shengyu Zhang, and Wei Chen. Contextual combinatorial cascading bandits. In *International conference on machine learning*, pp. 1245–1253. PMLR, 2016.

Zhiyuan Jerry Lin, Raul Astudillo, Peter Frazier, and Eytan Bakshy. Preference exploration for efficient bayesian optimization with multiple outcomes. In *International Conference on Artificial Intelligence and Statistics*, pp. 4235–4258. PMLR, 2022.

Debmalya Mandal and Jiarui Gan. Socially fair reinforcement learning. *arXiv preprint arXiv:2208.12584*, 2022.

Masoud Mansoury, Himan Abdollahpouri, Bamshad Mobasher, Mykola Pechenizkiy, Robin Burke, and Milad Sabouri. Unbiased cascade bandits: Mitigating exposure bias in online learning to rank recommendation. *arXiv preprint arXiv:2108.03440*, 2021.

Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*, pp. 2243–2251, 2018.

Rishabh Mehrotra, Niannan Xue, and Mounia Lalmas. Bandit based optimization of multiple objectives on a music streaming platform. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3224–3233, 2020.

Kaisa Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 2012.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 429–438, 2020.

Hervé Moulin. *Fair division and collective welfare*. MIT press, 2003.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari. Achieving fairness in the stochastic multi-armed bandit problem. In *AAAI*, pp. 5379–5386, 2020.

Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of The Web Conference 2020*, pp. 1194–1204, 2020.

Gourab K Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. Fair ranking: a critical review, challenges, and future directions. *arXiv preprint arXiv:2201.12662*, 2022.

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*, pp. 461–469. SIAM, 2014.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, pp. 8905–8915. PMLR, 2020.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2219–2228. ACM, 2018.

Ashudeep Singh and Thorsten Joachims. Policy learning for fairness in ranking. *Advances in Neural Information Processing Systems*, 32, 2019.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

Jonathan Stray, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. What are you optimizing for? aligning recommender systems with human values. *arXiv preprint arXiv:2107.10939*, 2021.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Projection efficient subgradient method and optimal nonsmooth frank-wolfe method. *Advances in Neural Information Processing Systems*, 33:12211–12224, 2020.

Nicolas Usunier, Virginie Do, and Elvis Dohmatob. Fast online ranking with fairness of exposure. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2157–2167, 2022.

Peter Vamplew, Richard Dazeley, Cameron Foale, Sally Firmin, and Jane Mummery. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 20(1):27–40, 2018.

Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*, pp. 10686–10696. PMLR, 2021.

Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. Joint multisided exposure fairness for recommendation. *arXiv preprint arXiv:2205.00048*, 2022.

Huanle Xu, Yang Liu, Wing Cheong Lau, and Rui Li. Combinatorial multi-armed bandits with concave rewards and fairness constraints. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 2554–2560, 2021.

Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*, pp. 1–6, 2017.

Nicholas C. Yannelis. *Integration of Banach-Valued Correspondence*, pp. 2–35. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991. ISBN 978-3-662-07071-0.

Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

Alp Yurtsever, Olivier Fercoq, Francesco Locatello, and Volkan Cevher. A conditional gradient framework for composite convex minimization with applications to semidefinite programming. In *International Conference on Machine Learning*, pp. 5727–5736. PMLR, 2018.

Meike Zehlike and Carlos Castillo. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*, pp. 2849–2855, 2020.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000*, 2021.

Yinglun Zhu and Paul Mineiro. Contextual bandits with smooth regret: Efficient learning in continuous action spaces. In *International Conference on Machine Learning*, pp. 27574–27590. PMLR, 2022.

Yinglun Zhu, Dylan J Foster, John Langford, and Paul Mineiro. Contextual bandits with large action spaces: Made practical. In *International Conference on Machine Learning*, pp. 27428–27453. PMLR, 2022.