Improved Regret Bounds for Online Fair Division with Bandit Learning

Benjamin Schiffer, Shirley Zhang

Harvard University

Abstract

We study online fair division when there are a finite number of item types and the player values for the items are drawn randomly from distributions with unknown means. In this setting, a sequence of indivisible items arrives according to a random online process, and each item must be allocated to a single player. The goal is to maximize expected social welfare while maintaining that the allocation satisfies proportionality in expectation. When player values are normalized, we show that it is possible to with high probability guarantee proportionality constraint satisfaction and achieve $\tilde{O}(\sqrt{T})$ regret. To achieve this result, we present an upper confidence bound (UCB) algorithm that uses two rounds of linear optimization. This algorithm highlights fundamental aspects of proportionality constraints that allow for a UCB algorithm despite the presence of many (potentially tight) constraints. This result improves upon the previous best regret rate of $\tilde{O}(T^{2/3})$.

1 Introduction

The fair division of indivisible goods is a classic problem in computational social choice. In this problem, a set of goods must be fairly allocated among n players, where each player may have a different value for each good. For example, consider a central food bank which is in charge of distributing food to multiple food pantries across the region. Each food pantry may have differing preferences over various types of food depending on the populations they serve. The goal is then to divide the goods in a way that is fair relative to each player's valuations.

Online fair division adds another degree of difficulty. Instead of all items being known upfront, in online fair division goods arrive one at a time and must be irrevocably allocated at the time of arrival. The first goods that arrive must be allocated without knowing what future goods will be, and cannot be reallocated after more goods arrive. In the food bank example, this setting is especially relevant for perishable goods, which the food bank must quickly allocate after arrival. For example, (Mertzanidis, Psomas, and Verma 2024) describes a partnership with an Indiana program which redistributes rejected food by redirecting truck drivers from landfills to food banks. Truck drivers arrive on the app in an online manner, necessitating an online fair division algorithm to match drivers to food banks. The food available depends on what was rejected on a given day, which may be unpredictable.

A common fairness requirement is proportionality, which insists that each player receive at least $\frac{1}{n}$ of their total value for all goods. In settings where the allocation of an item may be randomized, it is natural to instead consider proportionality in expectation, where the expectation is taken over both the random player values and the random item types. For a given instance, there may be many proportional allocations, in which case we can differentiate further by also considering efficiency. Past works have incorporated efficiency by, for example, requiring Pareto optimality in addition to proportionality (Benadè et al. 2024) or maximizing utilitarian social welfare subject to proportionality (Procaccia, Schiffer, and Zhang 2024).

We study the online fair division problem in the setting where player values are *unknown* at the time of item arrival, and the value of a player for an item is only revealed if the item is allocated to that player. Specifically, we consider the setting where there are m item types, and each player's value for an item of each type is drawn from an (possibly different) unknown distribution. We would like to distribute items fairly in expectation to players despite not knowing their true mean values for item types. In our food bank example, this setting is most relevant when a food bank does not yet understand the needs of its food pantries, but can easily collect information from each food pantry (e.g. in the form of surveys) regarding how well items are received by each food pantry's visitors. (Yamada et al. 2024) gives further motivation for this setting in the form of allocating difficult tasks to users with different strengths and distributing humanitarian aid, both of which are online fair division problems in which values are only unveiled after requesting feedback. In such settings, it is necessary to learn each player's expected value for each item.

We consider the problem statement proposed in (Procaccia, Schiffer, and Zhang 2024), which is as follows. Subject to the fairness constraints of proportionality in expectation, we strive for efficiency as measured by the *expected utilitarian social welfare* of our solution. Specifically, we aim to mini-

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

mize the regret incurred when comparing to an optimal solution which must adhere to the fairness constraint but knows the true means. As the algorithm does not know the players' true means, it will be impossible to achieve non-trivial regret and guarantee proportionality in expectation at every time step. Therefore, we will instead require that the algorithm with high probability satisfies the proportionality in expectation constraints at every time step.

In summary, we consider online fair division with unknown means as a reinforcement learning problem subject to fairness constraints. In this paper, we study how to balance exploration and exploitation while also maintaining proportionality.

1.1 Our Contributions

In this work, we study the problem of online fair division with unknown means subject to proportionality constraints. Our main result is that, when player values are normalized, an algorithm can achieve $\tilde{O}(\sqrt{T})$ regret while satisfying proportionality in expectation constraints at every time step with high probability (Theorem 1). This is an improvement on the $\tilde{\Omega}(T^{2/3})$ regret of the explore-then-commit algorithm in (Procaccia, Schiffer, and Zhang 2024). The algorithm that achieves our result (Algorithm 1) uses a variant of upper confidence bound (UCB) logic with two rounds of linear program optimization. The first round of optimization guarantees that with high probability the constraints are satisfied, but does not provide sufficient exploration for UCB. Therefore, Algorithm 1 performs a second round of optimization that exploits the underlying structure of the fairness constraints to sufficiently explore without losing significant social welfare.

We complement our positive results for proportionality with an impossibility result for envy-freeness, another commonly studied fairness notion. Specifically, we show that when values are normalized, the best regret rate for envy-freeness is $\tilde{O}(T^{2/3})$, which matches the lower bound for envy-freeness when values are not normalized. This highlights a fundamental difference in the difficulty of maintaining envy-freeness versus proportionality when learning unknown values.

1.2 Related Work

Our problem is most closely related to that in (Procaccia, Schiffer, and Zhang 2024), which introduces the problem setting we study. (Procaccia, Schiffer, and Zhang 2024) studies both envy-freeness in expectation and proportionality in expectation constraints, and provide explore-then-commit algorithms which achieve $\tilde{O}(T^{2/3})$ regret while maintaining these fairness constraints with high probability. (Procaccia, Schiffer, and Zhang 2024) also proves that no algorithm can have lower regret while maintaining these fairness constraints. In contrast, we show that when players' values are normalized, there do exist algorithms which achieve $\tilde{O}(\sqrt{T})$ regret while maintaining proportionality in expectation at each time step with high probability. The algorithm we present is an upper confidence bound algorithm rather than an explore-then-commit algorithm as used in (Procaccia, Schiffer, and Zhang 2024). While our algorithm relies on fundamental properties of fairness constraints similar to those in (Procaccia, Schiffer, and Zhang 2024), the properties needed for $\tilde{O}(\sqrt{T})$ regret are stronger and are *not* satisfied by envy-freeness constraints.

We briefly mention two other related works in online fair division. (Yamada et al. 2024) studies a similar setting in which a player's value for an item type is unknown and is only observed (with noise) when an item is allocated to that player. Rather than encoding fairness as a constraint, however, (Yamada et al. 2024) instead maximizes Nash social welfare in the objective function by leveraging algorithms which use dual averaging. (Benadè et al. 2024) studies a somewhat different online fair division setting in which items arrive in an adversarial manner, but the values of all agents for each item are known when the item arrives (and, crucially, before item allocation).(Benadè et al. 2024) also considers efficiency, but in the form of guaranteeing Pareto optimality rather than maximizing utilitarian social welfare.

There are many notions of fairness that have been studied in the multi-armed bandits literature. One such notion is that similar individuals are treated similarly (Chen, Li, and Ye 2021; Liu et al. 2017). Another related notion of fairness is that 'worse' arms are never pulled with higher probability than 'better' arms (Joseph et al. 2016b,a). These two notions of fairness are incompatible with proportionality, as proportionality may require that worse arms are pulled with higher probability. A third notion of fairness is the requirement that every arm is pulled a minimum proportion of the time (Chen et al. 2020; Claure et al. 2020; Li, Liu, and Ji 2019; Patil et al. 2021). Once again, this notion of fairness is not compatible with proportionality, as there exist proportional allocations where every item type is not allocated to every individual. We also focus solely on the non-contextual setting, however many works also study fairness when there is context (Grazzi et al. 2022; Schumann et al. 2019; Wang et al. 2021; Wu, Zheng, and Zhu 2023; Wei, Ma, and Wang 2024).

Because our fairness constraints and objective function are linear, our problem formulation is also related to the problem of multi-armed bandits under general linear constraints. One area of work studies linear bandits under linear safety constraints (Amani, Alizadeh, and Thrampoulidis 2019; Carlsson et al. 2024; Moradipari, Alizadeh, and Thrampoulidis 2020). (Amani, Alizadeh, and Thrampoulidis 2019) shows that if there is a single linear constraint and the optimal solution has positive slack with respect to this constraint, then $\tilde{O}(\sqrt{T})$ regret is possible. When the slack of the optimal solution is 0, (Amani, Alizadeh, and Thrampoulidis 2019) presents an algorithm that has regret of $\tilde{O}(T^{2/3})$. Our setting differs from (Amani, Alizadeh, and Thrampoulidis 2019) in that we have many linear constraints (one for each player), and the optimal solution frequently has zero slack for some of the constraints. The existence of the $\tilde{O}(\sqrt{T})$ regret algorithm in our setting despite these added difficulties fundamentally relies on the structure of our constraints. Other

works studying linear bandits have constraints that differ from our setting either because they are not applied at every time step or because they require slack in the constraints (Liu et al. 2021; Pacchiano et al. 2021).

Another related area is bandits with knapsacks, which also studies bandits problems with constraints (Liu et al. 2022; Badanidiyuru, Kleinberg, and Slivkins 2018). However, the knapsack constraints depend on resource consumption vectors rather than the unknown mean values, and therefore these constraints are significantly different than proportionality constraints.

2 Model

In our setting, there are $N = [n]$ players and $M = [m]$ item types, with T items arriving over time. For any item of type k, the value of player i for that item is drawn from a sub-Gaussian distribution with mean μ^*_{ik}. For each round t from 0 to $T-1$, an item j_t of type k_t is drawn uniformly at random. As shown in (Procaccia, Schiffer, and Zhang 2024), the assumption of uniformly random item types is WLOG, and all of our results hold when item types are drawn from any arbitrary distribution. After observing k_t, an algorithm allocates the item j_t (potentially randomly) to a player i_t, and then observes i_t's value v_t for j_t. In order to decide how to allocate j_t, an algorithm may consult the history $H_t = \{k_{t'}, i_{t'}, v_{t'}\}_{t'<t}$. Note that the algorithm never observes the values of a player i for item j if j is not allocated to i.

An algorithm allocates items to players via fractional allocations $X \in \mathbb{R}^{n\times m}$, where a fractional allocation is said to be *valid* if $\sum_i X_{ik} = 1$ for all $k \in [m]$. Intuitively, the kth column of a fractional allocation represents how the algorithm will randomly allocate the item if the item has type k. One valid fractional allocation is the *uniform at random* (UAR) allocation, where every element is equal to $\frac{1}{n}$. At time t, before observing k_t, an algorithm considers the history H_t and outputs a valid fractional allocation $X^t = \mathrm{ALG}(H_t)$. After observing k_t, the algorithm will then allocate j_t based on the probabilities in $((X^t)^\top)_{k_t}$, i.e. the k_tth column of X^t. The expected value of player i for a fractional allocation X can be written as $\frac{1}{m}X_i \cdot \mu^*_i$, and the sum over all players' expected values for X is then

$$\frac{1}{m}\sum_{i\in[n]} X_i \cdot \mu^*_i = \frac{1}{m}\langle X, \mu^*\rangle_F.$$

Therefore, we can write the total expected social welfare of an algorithm which allocates according to X^t at time t as

$$\mathbb{E}[\text{social welfare of ALG}] = \frac{1}{m}\sum_{t=0}^{T-1}\langle X^t, \mu^*\rangle_F.$$

In this paper, we make two additional assumptions on the unknown mean matrix μ^*. The first assumption is that there exist known $a, b \in \mathbb{R}$ such that $0 < a \le \mu^*_{ik} \le b < \infty$ for all i, k. As shown in (Procaccia, Schiffer, and Zhang 2024), this is a necessary assumption in order to achieve $o(T)$ regret. The second assumption made in this paper is that the values for each player are normalized. In other words, we assume that for all players i, $\sum_{k=1}^m \mu^*_{ik} = 1$. Informally, normalizing values ensures that each player has equal say in the total social welfare. Normalized values is a standard assumption in fair division literature (see, e.g., (Gkatzelis, Psomas, and Tan 2021; Bogomolnaia, Moulin, and Sandomirskiy 2022; Yamada et al. 2024)) and further justification can be found in (Aziz 2020). Note that assuming normalized values does not affect the proportionality constraints, which are invariant to scaling.

2.1 Regret and Problem Formulation

In this section, we give our formal definitions for fairness and regret.

The main fairness notion we study is *proportionality in expectation*. Proportionality in expectation requires that, for every t, every player's expected value for the allocation X^t is at least as much as that player's expected value for the uniform at random allocation.

Formally, player i's expected value for fractional allocation X^t is equal to $\frac{1}{m}X^t_i \cdot \mu^*_i$, where the $1/m$ comes from every item having probability $1/m$ of being each item type. Player i's expected value for the UAR allocation is $\frac{1}{nm}\|\mu^*_i\|_1 = \frac{1}{nm}$ due to the normalized values assumption. Therefore, a fractional allocation is proportional in expectation if and only if $\frac{1}{m}X^t_i \cdot \mu^*_i \ge \frac{1}{nm}$, which is equivalent to $X^t_i \cdot \mu^*_i \ge \frac{1}{n}$. We define proportionality in expectation for an algorithm ALG in Definition 1.

Definition 1. An algorithm ALG that uses fractional allocation X^t at time t satisfies *proportionality in expectation* for μ^* if

$$X^t_i \cdot \mu^*_i \ge \frac{1}{n} \quad \forall t < T, i \in [n].$$

In this paragraph we will briefly summarize from (Procaccia, Schiffer, and Zhang 2024) the justification for studying proportionality in expectation rather than realized proportionality. First, note that satisfying proportionality in expectation does not guarantee that every player prefers their final set of allocated items at time T to a $1/n$ proportion of all T items (which we refer to as realized proportionality). In fact, no algorithm can with high probability guarantee that every player prefers their final set of allocated items at time T to a $1/n$ proportion of all T items. Define the *dis-proportionality* of the final allocation as the maximum across all players of how much each player prefers a $1/n$ proportion of all T items to their allocated items at time T. An algorithm that satisfies proportionality in expectation has the asymptotically optimal rate of dis-proportionality among all possible algorithms. See (Procaccia, Schiffer, and Zhang 2024) for more discussion about the optimality of proportionality in expectation, including formal statements and proofs.

For any value matrix μ, let Y^μ be the expected social welfare maximizing fractional allocation that satisfies proportional-

ity in expectation for μ. Formally, we define

$$Y^{\mu} := \arg\max \ \langle X, \mu \rangle_F$$
$$\text{s.t. } X_i^t \cdot \mu_i^* \geq \frac{1}{n} \quad \forall i$$
$$\sum_i X_{ik} = 1 \quad \forall k \tag{1}$$

If the true mean values matrix μ^* is known, then the social welfare maximizing algorithm ALG that satisfies proportionality in expectation is simply the algorithm that chooses $X^t = Y^{\mu^*}$ for all $t \in [0 : T-1]$. Using this optimal algorithm as a baseline, we define the regret of an arbitrary algorithm ALG as follows.

Definition 2. Define Y^{μ^*} as the solution to LP (1) when $\mu = \mu^*$. Then the T-step regret for μ^* of an algorithm ALG that uses allocation X^t at time t can be written as

$$\text{Regret of ALG for } \mu^* = T \cdot \langle Y^{\mu^*}, \mu^* \rangle_F - \sum_{t=0}^{T-1} \langle X^t, \mu^* \rangle_F.$$

The formal problem we address in this paper is as follows.

Problem 1. *Design an algorithm* ALG *such that the following result holds for any known n, m, T, a, b. Suppose that the true mean values satisfy $a \leq \mu_{ik}^* \leq b$ for all $i \in [n], k \in [m]$. Then with probability $1 - 1/T$,* ALG *will both satisfy the proportionality in expectation constraints for μ^* and have regret for μ^* of $\tilde{O}(\sqrt{T})$.*

2.2 Notation

Throughout this paper, we will use $O(), \tilde{O}(), \Omega(), \tilde{\Omega}()$ notation to represent the limiting behavior of functions with respect to T. For two matrices A and B, we use $\langle A, B \rangle_F$ to represent the Frobenius product of A and B. We use A_i to represent the ith row of matrix A and $A_i \cdot B_i$ to represent the dot product between vectors A_i and B_i. For matrices $\mu, \epsilon \in \mathbb{R}^{n \times m}$, define

$$B(\mu, \epsilon) = \{\mu' \in \mathbb{R}^{n \times m} : \mu_{ik} - \epsilon_{ik} \leq \mu'_{ik} \leq \mu_{ik} + \epsilon_{ik} \forall i, k\}.$$

3 Main Results

3.1 Algorithm Overview

In this section, we present our main algorithm (Algorithm 1) and main theorem (Theorem 1).

Theorem 1. *Suppose n, m, T, a, b are known and that the true mean values satisfy $0 < a \leq \mu_{ik}^* \leq b$ for all $i \in [n]$, $k \in [m]$. With probability $1 - 1/T$, Algorithm 1 will both satisfy the proportionality in expectation constraints for μ^* and have regret for μ^* of $\tilde{O}(n^5 m^3 \sqrt{T})$.*

The initial exploration phase of Algorithm 1 uses the fact that the uniform at random allocation is guaranteed to satisfy proportionality in expectation constraints for any mean value matrix μ. After the exploration phase, at each step we calculate an estimated mean value matrix ($\hat{\mu}^t$) and an uncertainty matrix (ϵ^t). Algorithm 1 then performs two rounds of optimization. The first optimization of Algorithm 1 calculates an optimistic estimate of expected social welfare and guarantees that the solution $\hat{X}^t$ will satisfy the proportionality in expectation constraints for any $\mu \in B(\hat{\mu}^t, \epsilon^t)$. Therefore, if the true mean value matrix μ^* is in $B(\hat{\mu}^t, \epsilon^t)$, then $\hat{X}^t$ will satisfy the proportionality in expectation constraints for μ^*. The algorithm unfortunately cannot directly use this $\hat{X}^t$ as the allocation in round t because $\hat{X}^t$ does not necessarily provide sufficient exploration of all (item, player) pairs. See Section 3.2 below for more details.

To avoid this issue, the algorithm includes a second round of optimization in LP (3) that calculates an allocation $\hat{Z}^{ik}$ for each (item, player) pair (i, k). $\hat{Z}^{ik}$ is guaranteed to sufficiently explore the (i, k) pair and have social welfare that is not significantly less than the social welfare of $\hat{X}^t$. By using the fractional allocation X^t that averages over all $\hat{Z}^{ik}$, the algorithm is able to guarantee that X^t sufficiently explores every (player, item) pair. We note that due to the maximization in the second round of optimization, the runtime of Algorithm 1 is exponential in n and m. In the algorithm notation below, subscripts represent matrix indexing while superscripts represent matrix names, i.e. Z^{ik} is a matrix, and X_{ik}^t is the (i, k) entry of X^t.

3.2 Algorithm Intuition

In this section, we discuss the intuition behind the $\tilde{O}(\sqrt{T})$ regret guarantee of Algorithm 1. First, we describe how Algorithm 1 relates to the standard multi-armed bandits upper confidence bound algorithm. We then describe the importance of the second round of optimization in Algorithm 1 and why the algorithm does not simply use the fractional allocation $\hat{X}^t$ at time t.

Consider the standard multi-armed bandits (MAB) setting, where there are n arms that each have an unknown mean reward. At each time step, the algorithm chooses one arm, and the goal is to maximize the T-step reward. In this setting, a standard UCB algorithm computes estimates of the mean rewards $\hat{\mu}^t$ (where $\hat{\mu}_j^t$ corresponds to arm j) and an uncertainty vector ϵ^t (where ϵ_j^t is the uncertainty in $\hat{\mu}_j^t$). The standard UCB algorithm at time t chooses arm $j_t = \arg\max_j \hat{\mu}_j^t + \epsilon_j^t$. The key idea underlying the low regret of standard UCB is that

$$[\text{Regret at time } t] = O(\epsilon_{j_t}). \tag{4}$$

In online fair division, instead of choosing a single arm j_t, the algorithm chooses a fractional allocation X^t. The generalization of Equation (4) to this setting is

$$[\text{Regret at time } t] = O\left(\langle X^t, \epsilon^t \rangle_F\right). \tag{5}$$

The standard UCB approach for finding a fractional allocation X^t that satisfies Equation (5) is to solve the optimization problem

$$\arg\max_X \langle X, \mu_U^t \rangle_F$$
$$\sum_i X_{ik} = 1 \quad \forall k \tag{6}$$

Algorithm 1: UCB Online Fair Division

Require: n, m, T, a, b

for $t \leftarrow 0$ to $\log^2(T)\sqrt{T} - 1$ **do**

Use $X^t = \text{UAR}$.

end for

for $t \leftarrow \log^2(T)\sqrt{T}$ to T **do**

$N_{ik}^t \leftarrow \sum_{\tau=0}^{t-1} 1_{k_\tau = k, i_\tau = i}$

$\hat{\mu}_{ik}^t \leftarrow \frac{1}{N_{ik}^t} \sum_{\tau=0}^{t-1} 1_{k_\tau = k, i_\tau = i} v_\tau$

$\epsilon_{ik}^t = \sqrt{\log^2(6nmT)/(N_{ik}^t)}$

$(\mu_U^t)_{ik} = \hat{\mu}_{ik}^t + \epsilon_{ik}^t$

$G^t = \left\{ \mu \in B(\hat{\mu}^t, \epsilon^t) : \sqrt{T}\mu_{ik} \in \mathbb{Z} \quad \forall i, k \right\}$

$\hat{X}^t \leftarrow$ Solution to the following LP:

$$\max_X \langle X, \mu_U^t \rangle_F$$
$$\text{s.t. } X_{i'} \cdot \mu_{i'} \geq \frac{1}{n} \quad \forall i' \in [n],\ \forall \mu \in B(\hat{\mu}^t, \epsilon^t)$$
$$\sum_i X_{ik} = 1 \quad \forall k \tag{2}$$

$\forall i \in [n], \forall k \in [m]$, $\hat{Z}^{ik} \leftarrow$ Solution to the following LP:

$$\max X_{ik}$$
$$\text{s.t. } X_{i'} \cdot \mu_{i'} \geq \frac{1}{n} \quad \forall i' \in [n],\ \forall \mu \in B(\hat{\mu}^t, \epsilon^t)$$
$$\langle X, \mu_U^t \rangle_F \geq \langle \hat{X}^t, \mu_U^t \rangle_F - \frac{4bn}{a} \max_{\mu \in G^t} \langle Y^\mu, \epsilon^t \rangle_F - 2\langle \hat{X}^t, \epsilon_t \rangle_F$$
$$\sum_{i'} X_{i'k'} = 1 \quad \forall k' \tag{3}$$

Use $X^t = \frac{1}{nm}\left(\sum_{i,k} \hat{Z}^{ik}\right)$

end for

As in standard MAB, the solution to LP (6) will satisfy Equation (5). In online fair division, however, the algorithm must choose an X^t satisfying the proportionality constraints, and the solution to LP (6) may not satisfy these constraints. Instead, Algorithm 1 uses LP (2) (which has the same objective function as LP (6)) to find an allocation $\hat{X}^t$ that satisfies the proportionality constraints with high probability. However, $\hat{X}^t$ may no longer satisfy Equation (5). This means the algorithm cannot use the allocation $\hat{X}^t$ and bound the regret with a UCB argument.

Because we cannot directly use $\hat{X}^t$, Algorithm 1 instead leverages $\hat{X}^t$ to find an allocation X^t that will satisfy Equation (5). This is done using LP (3). LP (3) is designed to find Z^{ik} such that the (i,k) entry of Z^{ik} is relatively large compared to the (i,k) entry of Y^{μ^*}. Algorithm 1 chooses X^t to be a linear combination of the Z^{ik}. Therefore, every entry in X^t will be relatively large compared to the corresponding entry in Y^{μ^*}. Furthermore, the second constraint in LP (3) guarantees that X^t will not have significantly less social welfare than $\hat{X}^t$. The bulk of the theoretical work in proving Theorem 1 is showing that the previous two sentences together imply that X^t will satisfy a (slightly more complicated) version of Equation (5). Once we show that Equation (5) holds for X^t, a UCB argument bounds the regret of the algorithm to be $\tilde{O}(\sqrt{T})$.

4 Properties of Proportionality Constraints

The proof of Theorem 1 relies on three key results about proportionality in expectation constraints, which we outline in this section.

The first result, Lemma 1, is the reason that Algorithm 1 can explore and satisfy the proportionality in expectation constraints when the mean values are unknown.

Lemma 1. *The uniform at random allocation satisfies the proportionality in expectation constraints for any* $\mu^* \in [a,b]^{n \times m}$.

proof. If every player is given exactly $1/n$ proportion of every item type as in the UAR allocation, then every player has value $1/n$ for their allocation. This implies that the proportionality in expectation constraints will be satisfied. □

The second lemma shows that the total social welfare of the optimal allocation is continuous in the mean value matrix.

Lemma 2. *Define* Y^μ *as the solution to LP* (1)*. Then there exists a* γ_0 *such that if* $\|\mu - \mu'\|_1 \leq \gamma_0$ *for* $\mu, \mu' \in [a,b]^{n \times m}$*, then* $\langle Y^\mu, \mu \rangle_F - \langle Y^{\mu'}, \mu' \rangle_F \leq \frac{bn}{a}\|\mu - \mu'\|_1$.

proof. We provide a brief proof sketch and defer the formal proof to Appendix C. Define $\epsilon = \|\mu - \mu'\|_1$. Let Z^μ be the solution to the following modification of LP 1:

$$\max \langle X, \mu \rangle_F$$
$$\text{s.t. } X_i \cdot \mu_i - \frac{1}{n} \geq -\epsilon \quad \forall i \in [n]$$
$$\sum_i X_{ik} = 1 \quad \forall k \tag{7}$$

We next construct a fractional allocation W^μ such that W^μ is a solution to LP 1 and such that $\langle W^\mu, \mu \rangle_F - \langle Z^\mu, \mu \rangle_F \geq -O(\epsilon)$. That is to say, the social welfare of W^μ is not much worse than that of Z^μ. Because W^μ is a solution to LP 1, we have

$$\langle Y^\mu, \mu \rangle_F \geq \langle W^\mu, \mu \rangle_F \geq \langle Z^\mu, \mu \rangle_F - n\epsilon.$$

Next, we note that $Y^{\mu'}$ is also a solution to LP 7 by construction. Therefore, it must be the case that

$$\langle Z^\mu, \mu \rangle_F \geq \langle Y^{\mu'}, \mu \rangle_F.$$

Combining the previous two equations gives that $\langle Y^\mu, \mu \rangle_F \geq \langle Y^{\mu'}, \mu' \rangle_F - n\epsilon$. By symmetry, we also have that $\langle Y^{\mu'}, \mu' \rangle_F \geq \langle Y^\mu, \mu \rangle_F - n\epsilon$, and together with the previous equation this proves the desired result. □

Lemma 3 is the key reason that Algorithm 1 is able to find an allocation that satisfies the proportionality constraints without losing significant social welfare. Informally, this lemma states that for a known mean matrix μ, there exists an allocation X' such that X' has only slightly less expected social welfare than Y^μ and such that either $X' = \text{UAR}$, or X' is close to Y^μ *and* every proportionality constraint has non-negligible slack under X'.

Lemma 3. *Define Y^μ as the solution to LP* (1)*. Then for any $\gamma < \frac{a}{bn}$ and any $\mu \in [a,b]^{n\times m}$, there exists an allocation X' such that $\langle X', \mu\rangle_F \geq \langle Y^\mu, \mu\rangle_F - \frac{bn\gamma}{a}$ and either $X' = \text{UAR}$ or for each $i \in [n]$,*

1. $X'_i \cdot \mu_i \geq \frac{1}{n} + \gamma$ *and*
2. $\forall i \in [n], k \in [m]$, $|X'_{ik} - Y^\mu_{ik}| \leq \frac{n\gamma}{a}$.

The proof of Lemma 3 uses the same construction as in Lemma 2 of (Procaccia, Schiffer, and Zhang 2024), and we defer the proof to Appendix D. Informally, the construction either sets $X' = \text{UAR}$ or constructs X' by redistributing allocation away from players with large proportionality surplus (i.e. players who strictly prefer their allocation to UAR). Unlike in (Procaccia, Schiffer, and Zhang 2024), the proof uses that this redistribution process always redistributes at most $n\gamma/a$ from any (player, item) pair, thereby satisfying the second condition of Lemma 3.

5 Proof Sketch of Theorem 1

We are now ready to present the proof sketch of Theorem 1. In Appendix A, we prove a more general result than Theorem 1 that applies to any set of fairness constraints satisfying general versions of Properties 1, 2, and 3. For this proof sketch, we will outline the proof for proportionality constraints. See Appendix A for more details on the general version of this result.

Proof sketch. By Lemma 1, the UAR allocations used for the first $\sqrt{T}\log^2(T)$ steps will satisfy the proportionality constraints. The regret of the first $\sqrt{T}\log^2(T)$ steps is $\tilde{O}(\sqrt{T})$ because the regret of any one step is upper bounded by $b - a$.

Now we study what happens in the algorithm for $t \geq \sqrt{T}\log^2(T)$. Define event E as the event that $\mu^* \in B(\hat{\mu}^t, \epsilon^t)$ and $\|\epsilon^t\|_1 = \tilde{O}(T^{-1/4})$ for every t. By two applications of Azuma–Hoeffding's inequality, $\Pr(E) = 1 - \frac{2}{3T}$. Under event E, for every round t and for all i, k, the allocation $\hat{Z}^{ik}$ will satisfy the proportionality in expectation constraints for μ^* due to the first constraint in LP (3). Because X^t is a linear combination of the $\hat{Z}^{ik}$ and the proportionality constraints are linear, this implies that under event E, X^t will also satisfy the proportionality in expectation constraints for μ^*.

Now we will bound the regret of Algorithm 1 for $t \geq \log^2(T)\sqrt{T}$. The key step in bounding this regret is showing that the regret at time $t \geq \log^2(T)\sqrt{T}$ is

$$\langle Y^{\mu^*}, \mu^*\rangle_F - \langle X^t, \mu^*\rangle_F = \tilde{O}\left(\langle X^t, \epsilon^t\rangle_F + \frac{1}{\sqrt{T}}\right). \quad (8)$$

In order to show Equation (8), we first bound the regret at time t by an expression which does not contain any terms involving μ^*.

Under event E, $\mu^* \in B(\hat{\mu}^t, \epsilon^t)$. G^t forms a grid on $B(\hat{\mu}^t, \epsilon^t)$, and therefore there exists some element $\mu_g \in G^t$ such that $\|\mu^* - \mu_g\|_\infty \leq \frac{1}{\sqrt{T}}$. By Lemma 2, this implies that

$$|\langle Y^{\mu^*}, \mu^*\rangle_F - \langle Y^{\mu_g}, \mu_g\rangle_F| = O\left(\frac{1}{\sqrt{T}}\right). \quad (9)$$

Using the first constraint of LP (3), we can bound $\langle \hat{X}^t, \mu^U_t\rangle_F - \langle \hat{Z}^{ik}, \mu^U_t\rangle_F$. By construction of X^t and μ^g, this implies that

$$\begin{aligned}&\langle \hat{X}^t, \mu^g\rangle_F - \langle X^t, \mu^*\rangle_F \\ &= \tilde{O}_T\left(\max_{\mu\in G^t}\langle Y^\mu, \epsilon^t\rangle_F + \langle \hat{X}^t, \epsilon_t\rangle_F + \langle \hat{Z}^{ik}, \epsilon^t\rangle_F\right).\end{aligned} \quad (10)$$

Furthermore, a series of algebraic steps with a UCB argument shows that

$$\begin{aligned}&\langle Y^{\mu^g}, \mu^g\rangle_F - \langle \hat{X}^t, \mu^g\rangle_F \\ &= \tilde{O}\left(\max_{\mu\in G^t}\langle Y^\mu, \epsilon^t\rangle_F + \langle \hat{X}^t, \epsilon^t\rangle_F\right).\end{aligned} \quad (11)$$

Combining Equations (9), (10), and (11) gives the following key result.

$$\begin{aligned}&\langle Y^{\mu^*}, \mu^*\rangle_F - \langle X^t, \mu^*\rangle_F \\ &= \tilde{O}\left(\langle \hat{X}^t, \epsilon^t\rangle_F + \max_{\mu\in G^t}\langle Y^\mu, \epsilon^t\rangle_F + \langle X^t, \epsilon^t\rangle_F + \frac{1}{\sqrt{T}}\right).\end{aligned} \quad (12)$$

To show Equation (8) from Equation (12), we bound the first two terms in Equation (12) by $O(\langle X^t, \epsilon^t\rangle_F + \frac{1}{\sqrt{T}})$ conditional on event E. First, we show that for all i and k, $\hat{X}^t$ satisfies the constraints of LP (3). We can then conclude that $\hat{Z}^{ik}_{ik} \geq \hat{X}^t_{ik}$ because $\hat{Z}^{ik}$ is the solution to LP (3). X^t is a linear combination of the $\hat{Z}^{ik}$, and therefore the previous sentence implies that $X^t_{ik} \geq \frac{1}{nm}\hat{X}^t_{ik}$. This implies that

$$\langle \hat{X}^t, \epsilon^t\rangle_F = O(\langle X^t, \epsilon^t\rangle_F). \quad (13)$$

Under event E, Lemma 3 with a carefully chosen value of γ implies that for every $\mu \in G^t$, there exists an allocation X^μ such that $\langle X^\mu, \mu\rangle_F$ is similar to $\langle Y^\mu, \mu\rangle_F$ and such that X^μ is a solution to LP 3. The fact that X^μ is a solution to LP 3 is a result of the careful construction of constraints in LP 3 and choice of γ for Lemma 3. Because X^μ is a solution to LP (3), by the same logic as in the previous paragraph, we have that $\langle X^\mu, \epsilon^t\rangle_F = O(\langle X^t, \epsilon^t\rangle_F)$. X^μ was constructed in Lemma 3 such that the elements of X^μ are close to the elements

of Y^μ. Therefore, $\langle Y^\mu, \epsilon^t \rangle_F = O(\langle X^t, \epsilon^t \rangle_F + \frac{1}{\sqrt{T}})$ for all $\mu \in G^t$, or equivalently

$$\max_{\mu \in G^t} \langle Y^\mu, \epsilon^t \rangle_F = O(\langle X^t, \epsilon^t \rangle_F + \frac{1}{\sqrt{T}}). \tag{14}$$

Putting together Equations (12), (13), and (14) gives the desired result of Equation (8). Using Equation (8), we can conclude with an upper confidence bound argument to upper bound the total T-step regret by

$$\begin{aligned}
&\sum_{t=0}^{T-1} \langle Y^{\mu^*}, \mu^* \rangle_F - \langle X^t, \mu^* \rangle_F \\
&\leq \tilde{O}(\sqrt{T}) + \sum_{t=\log^2(T)\sqrt{T}}^{T-1} \tilde{O}\left(\langle X^t, \epsilon^t \rangle_F + \frac{1}{\sqrt{T}} \right) \\
&= \tilde{O}(\sqrt{T}).
\end{aligned}$$

□

5.1 Lower Bounds

The regret of $\tilde{O}(\sqrt{T})$ in Theorem 1 is tight for T up to log factors because $\tilde{\Omega}(\sqrt{T})$ is the standard lower bound for stochastic multi-armed bandit problems. A natural follow-up question to Theorem 1 is whether an equivalent result holds when the algorithm must satisfy other notions of fairness such as envy-freeness in expectation. For the online fair division problem, envy-freeness in expectation can be represented as constraints in the following way:

Definition 3. An algorithm ALG that uses fractional allocation X^t at time t satisfies *envy-freeness in expectation* if for all $t < T$ and all $i \in [n]$,

$$(X_t)_i \cdot \mu_i^* \geq \max_{i' \in [n]} (X_t)_{i'} \cdot \mu_i^* \quad \forall i \in [n].$$

Theorem 2 shows that an equivalent result to Theorem 1 does not hold for envy-freeness in expectation, and in fact the best regret possible while maintaining envy-freeness in expectation is $\tilde{\Omega}(T^{2/3})$.

Theorem 2. *There exists a, b, n, m such that no algorithm can, for all $\mu^* \in [a, b]^{n \times m}$ with rows that add to 1, both satisfy the envy-freeness in expectation constraints and achieve regret of less than $\frac{T^{2/3}}{\log(T)}$ with probability at least $1 - 1/T$.*

Proof sketch. The proof of this theorem extends the lower bound construction of (Procaccia, Schiffer, and Zhang 2024) to an example with normalized mean values. We prove the desired result by contradiction. First, we assume that an algorithm ALG does achieve regret of less than $\frac{T^{2/3}}{\log(T)}$ and satisfies the envy-freeness in expectation constraints with probability greater than $1 - 1/T$ for all mean value matrices with normalized values. We then construct two mean value matrices each with three players and three types of items that differ only in two entries by $T^{-1/3}$, and therefore are difficult to distinguish between. We then show that ALG cannot simultaneously have low regret and satisfy the envy-freeness in expectation constraints for both of these mean value matrices, which leads to a contradiction. See Appendix E for the formal proof. □

6 Discussion

6.1 Non-random Item Types

We studied the online fair division problem where T items arrive online and each item's type is drawn uniformly at random. In this section, we discuss how our results extend to a similar problem where all item types are deterministic. Suppose that instead of a single item arriving at each of T time steps, a basket containing exactly one of every item type arrives at each of T/m time steps. The algorithm then must allocate every item in the basket among the n players using a random fractional allocation. The goal in this alternative setting is to maximize expected social welfare while satisfying the proportionality in expectation constraints for each basket.

For the setting studied in this paper, there are three sources of randomness: random item types, random player values, and random fractional allocations. In this alternative setting the item types are not random, and therefore there are only two sources of randomness: random player values and random fractional allocations. Note that m items with uniform at random types are equal in expectation to a basket of m items with one of every item type. Furthermore, both the constraints and the reward function for both settings are expressed as expectations. Therefore, the difference in sources of randomness does not fundamentally change the problem. This implies that the results from this paper (such as Theorem 1) carry over to this setting with deterministic item types.

6.2 Limitations and Future Directions

In this section, we discuss the limitations of this paper and some potential future directions. One limitation of Algorithm 1 is that the runtime is not linear due to the second round of optimization. An open question is whether an algorithm with linear runtime can achieve the same rate of regret. Furthermore, the main result of this paper, Theorem 1, only applies to proportionality constraints. In the proof of Theorem 1, we do present a general algorithm that achieves $\tilde{O}(\sqrt{T})$ regret for any fairness constraints satisfying certain properties. An open question is whether there exist other types of fairness constraints (e.g. equitability) for which $\tilde{O}(\sqrt{T})$ is also possible. We leave this as an open question for future work.

Another interesting question is whether the ideas in this paper can be extended to other fair division problems. More broadly, the techniques used in this paper are not exclusive to fairness constraints. Therefore, another question is whether similar ideas can lead to algorithms that achieve $\tilde{O}(\sqrt{T})$ regret under even more general classes of constraints.

Acknowledgements

Schiffer was supported by an NSF Graduate Research Fellowship. Zhang was supported by an NSF Graduate Research Fellowship. The authors would also like to thank Ariel Procaccia for helpful discussions.

References

Amani, S.; Alizadeh, M.; and Thrampoulidis, C. 2019. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32.

Aziz, H. 2020. Justifications of welfare guarantees under normalized utilities. *ACM SIGecom Exchanges*, 17(2): 71–75.

Badanidiyuru, A.; Kleinberg, R.; and Slivkins, A. 2018. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3): 1–55.

Benadè, G.; Kazachkov, A. M.; Procaccia, A. D.; Psomas, A.; and Zeng, D. 2024. Fair and Efficient Online Allocations. *Operations Research*. Forthcoming.

Bogomolnaia, A.; Moulin, H.; and Sandomirskiy, F. 2022. On the fair division of a random object. *Management Science*, 68(2): 1174–1194.

Carlsson, E.; Basu, D.; Johansson, F.; and Dubhashi, D. 2024. Pure exploration in bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, 334–342. PMLR.

Chen, G.; Li, X.; and Ye, Y. 2021. Fairer LP-based online allocation via analytic center. *arXiv preprint arXiv:2110.14621*.

Chen, Y.; Cuellar, A.; Luo, H.; Modi, J.; Nemlekar, H.; and Nikolaidis, S. 2020. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, 181–190. PMLR.

Claure, H.; Chen, Y.; Modi, J.; Jung, M.; and Nikolaidis, S. 2020. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 299–308.

Gkatzelis, V.; Psomas, A.; and Tan, X. 2021. Fair and efficient online allocations with normalized valuations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 5440–5447.

Grazzi, R.; Akhavan, A.; Falk, J. I.; Cella, L.; and Pontil, M. 2022. Group meritocratic fairness in linear contextual bandits. *Advances in Neural Information Processing Systems*, 35: 24392–24404.

Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2016a. Fair algorithms for infinite and contextual bandits. *arXiv:1610.09559*.

Joseph, M.; Kearns, M.; Morgenstern, J. H.; and Roth, A. 2016b. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29.

Li, F.; Liu, J.; and Ji, B. 2019. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3): 1799–1813.

Liu, Q.; Xu, W.; Wang, S.; and Fang, Z. 2022. Combinatorial bandits with linear constraints: Beyond knapsacks and fairness. *Advances in Neural Information Processing Systems*, 35: 2997–3010.

Liu, X.; Li, B.; Shi, P.; and Ying, L. 2021. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34: 24075–24086.

Liu, Y.; Radanovic, G.; Dimitrakakis, C.; Mandal, D.; and Parkes, D. C. 2017. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*.

Mertzanidis, M.; Psomas, A.; and Verma, P. 2024. Automating Food Drop: The Power of Two Choices for Dynamic and Fair Food Allocation. *arXiv preprint arXiv:2406.06363*.

Moradipari, A.; Alizadeh, M.; and Thrampoulidis, C. 2020. Linear thompson sampling under unknown linear constraints. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3392–3396. IEEE.

Pacchiano, A.; Ghavamzadeh, M.; Bartlett, P.; and Jiang, H. 2021. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, 2827–2835. PMLR.

Patil, V.; Ghalme, G.; Nair, V.; and Narahari, Y. 2021. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 22(174): 1–31.

Procaccia, A. D.; Schiffer, B.; and Zhang, S. 2024. Honor Among Bandits: No-Regret Learning for Online Fair Division. *arXiv preprint arXiv:2407.01795*.

Schumann, C.; Lang, Z.; Mattei, N.; and Dickerson, J. P. 2019. Group fairness in bandit arm selection. *arXiv preprint arXiv:1912.03802*.

Wang, L.; Bai, Y.; Sun, W.; and Joachims, T. 2021. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*, 10686–10696. PMLR.

Wei, W.; Ma, X.; and Wang, J. 2024. Fair adaptive experiments. *Advances in Neural Information Processing Systems*, 36.

Wu, Y.; Zheng, Z.; and Zhu, T. 2023. Best arm identification with fairness constraints on subpopulations. In *2023 Winter Simulation Conference (WSC)*, 540–551. IEEE.

Yamada, H.; Komiyama, J.; Abe, K.; and Iwasaki, A. 2024. Learning Fair Division from Bandit Feedback. In *International Conference on Artificial Intelligence and Statistics*, 3106–3114. PMLR.