Journal of Combinatorial Optimization (2023) 45:102
https://doi.org/10.1007/s10878-023-01035-4

Beyond submodularity: a unified framework of randomized set selection with group fairness constraints

Shaojie Tang[1] · Jing Yuan[2]

Accepted: 13 April 2023 / Published online: 28 April 2023

Abstract

Machine learning algorithms play an important role in a variety of important decision-making processes, including targeted advertisement displays, home loan approvals, and criminal behavior predictions. Given the far-reaching impact of these algorithms, it is crucial that they operate fairly, free from bias or prejudice towards certain groups in the population. Ensuring impartiality in these algorithms is essential for promoting equality and avoiding discrimination. To this end we introduce a unified framework for randomized subset selection that incorporates group fairness constraints. Our problem involves a global utility function and a set of group utility functions for each group, here a group refers to a group of individuals (e.g., people) sharing the same attributes (e.g., gender). Our aim is to generate a distribution across feasible subsets, specifying the selection probability of each feasible set, to maximize the global utility function while meeting a predetermined quota for each group utility function in expectation. Note that there may not necessarily be any direct connections between the global utility function and each group utility function. We demonstrate that this framework unifies and generalizes many significant applications in machine learning and operations research. Our algorithmic results either improves the best known result or provide the first approximation algorithms for new applications.

Keywords Fair subset selection · Approximation algorithms · Randomized algorithms · Submodular maximization

✉ Shaojie Tang
shaojie.tang@utdallas.edu

Jing Yuan
jing.yuan@unt.edu

[1] Naveen Jindal School of Management, University of Texas at Dallas, Richardson, USA

[2] Department of Computer Science and Engineering, University of North Texas, Denton, USA

1 Introduction

The increasing use of machine learning algorithms in decision-making has raised concerns about the possibility of biases and discrimination. However, various efforts are being made to develop fair algorithms that ensure equitable outcomes for individuals or groups, even in sensitive domains. These efforts involve creating techniques for classification (Zafar et al. 2017), ranking (Celis et al. 2017), clustering (Chierichetti et al. 2017), bandit learning (Joseph et al. 2016), voting (Celis et al. 2018), college admission (Abdulkadiroğlu 2005), matching (Chierichetti et al. 2019), influence maximization (Tsang et al. 2019), and diverse data summarization (El Halabi et al. 2020).

Various concepts of fairness have been proposed in the literature, including individual fairness, group fairness, and subgroup fairness. These concepts intend to tackle discrimination and bias in algorithmic decision-making, particularly in sensitive domains like employment, housing, and criminal justice. However, there is no universal measure of fairness since it often depends on the context and can be affected by various factors such as the decision-making process's objectives and the sensitive attribute in question. In this paper, we propose a general group fairness notation that unifies many notations from previous works. We assume there are m groups defined by a set of shared attributes like race, gender, or age. To assess the appropriateness of a particular solution set S for a specific group $t \in [m]$, we introduce m group utility functions $g_1, g_2, \ldots, g_m : 2^V \rightarrow \mathbb{R} \geq 0$. Each group utility function $g_t(S)$ evaluates the utility that group t derives from the solution set S. For instance, in committee selection (Celis et al. 2018), $g_t(S)$ corresponds to the number of candidates chosen from group t in the solution set S. Given a set of feasible subsets $\mathcal{F}$, let $x \in [0, 1]^{\mathcal{F}}$ represent a distribution over solution sets in $\mathcal{F}$, where x_S is the probability of choosing $S \in \mathcal{F}$. A distribution x is deemed fair if $\sum S \in \mathcal{F} x_S g_t(S) \geq \alpha_t, \forall t \in [m]$, meaning the expected utility from each group $t \in [m]$ is lower bounded by α_t. This enables the consideration of each group's representation in the final outcome, promoting diversity and preventing under-representation of any particular group.

In addition, there is a global utility function $f : 2^V \rightarrow \mathbb{R}_{\geq 0}$. The expected (global) utility of a distribution x can be computed as $\sum_{S \in \mathcal{F}} x_S f(S)$. Our goal is to find a distribution x that maximizes $\sum_{S \in \mathcal{F}} x_S f(S)$ while ensuring that $\sum_{S \in \mathcal{F}} x_S g_t(S) \geq \alpha_t, \forall t \in [m]$. A formal definition of this problem is listed in **P.0**. We show that this formulation brings together and extends numerous noteworthy applications in both machine learning and operations research. We next summarize the main contributions of this paper.

- We develop a polynomial-time algorithmic framework for **P.0** based on ellipsoid method.
- One main algorithmic finding (as formally presented in Theorem 1) is that suppose that for all $z \in \mathbb{R}^m_{\geq 0}$, there exists a polynomial-time algorithm that returns a set $A \in \mathcal{F}$ such that

$$\forall S \in \mathcal{F}, f(A) + \sum_{t \in [m]} g_t(A) \cdot z_t \geq \rho \cdot f(S) + \mu \cdot \sum_{t \in [m]} g_t(S) \cdot z_t,$$

for some $\rho, \mu \in [0, 1]$. Then there is a polynomial-time (ρ, μ)-approximation algorithm for **P.0**. Here, the ρ represents the approximation ratio, while the μ indicates that the solution may violate the fairness constraint by at most a factor of μ. Hence, solving **P.0** is reduced to finding an approximation algorithm for a combinatorial subset selection problem. We utilize this result to tackle various applications such as fairness-aware submodular maximization, sequential submodular maximization with group constraints, and assortment planning with market share constraints. Our approach outperforms existing methods for some of these applications, while for others, we introduce novel applications and develop the first approximation algorithms.

- Notably, when f is non-negative monotone and submodular, and g_t is a modular function, our approach gives a *feasible* and *optimal* $(1 - 1/e, 1)$-approximation solution.
- We explore extensions to other commonly used fairness metrics and propose effective algorithms for solving them. In the first extension, we introduce additional upper bounds β_t on the expected utility of each group $t \in [m]$. Specifically, we require that $\alpha_t \leq \sum_{S \in \mathcal{F}} x_S \cdot g_t(S) \leq \beta_t, \forall t \in [m]$. In the second extension, we explore another frequently used measure of fairness that aims to achieve parity in pairwise utility between groups. The degree of fairness is defined by a parameter γ. Specifically, we require that for any two groups $t, t' \in [m]$, the difference between their expected utilities does not exceed γ, i.e., $\sum_{S \in \mathcal{F}} x_S \cdot g_t(S) - \sum_{S \in \mathcal{F}} x_S \cdot g_{t'}(S) \leq \gamma, \forall t, t' \in [m]$.

Additional Related Work. Over the years, there has been a significant effort to develop fair algorithms across various fields to address the issue of biased decision-making. In the domain of influence maximization and classification, researchers have been actively developing fair algorithms (Tsang et al. 2019; Zafar et al. 2017). Similarly, in voting systems, the focus has been on ensuring that election outcomes are a fair representation of the preferences of voters, leading to the development of fair algorithms (Celis et al. 2018). The field of bandit learning, which involves making sequential decisions based on uncertain information, has also seen a growing interest in the development of fair algorithms to address the issue of bias (Joseph et al. 2016). Additionally, the field of data summarization has seen an increasing focus on the development of fair algorithms (Celis et al. 2018) to provide a balanced representation of data and avoid biased decision-making.

The choice of fairness metric in previous studies depends on the context and type of bias being addressed, resulting in a range of optimization problems and fair algorithms customized to the particular demands of each application. We adopt a general group utility function to assess the solution's utility from each group's perspective. Our framework is general enough to encompass numerous existing fairness notations, including the 80%-rule (Biddle 2017), statistical parity (Dwork et al. 2012), and proportional representation (Monroe 1995). While most previous research on fairness-aware algorithm design (such as Celis et al. 2018; Yuan and Tang 2023; Wang et al. 2021; Mehrotra and Celis 2021) aims to find a deterministic solution set, our goal is to compute a randomized solution that can meet the group fairness constraints on average. This approach offers more flexibility in attaining group fairness. Our framework is general enough to encompass various existing studies on achieving group fairness through randomization, such as those examined in Asadpour et al. (2022), Chen et al. (2022), Tang et al. (2023).

2 Preliminaries and problem statement

We consider a set of n items V and m groups. There is a *global* utility function $f : 2^V \to \mathbb{R}_{\geq 0}$ and m *group* utility functions $g_1, g_2, \ldots, g_m : 2^V \to \mathbb{R}_{\geq 0}$. Given a subset of items $S \subseteq V$, we use $g_t(S)$ to assess the utility of S from each group t's perspective. There is a required minimum expectation of utility from each group, represented by $\alpha \in \mathbb{R}^m_{\geq 0}$, which acts as fairness constraints. This formulation enables us to design fair algorithms that take into account the preferences of each group, ensuring that the decision-making process is unbiased and leads to fair outcomes for all groups. Note that there may not be any connections between f and g_t. A common way to define g_t, as elaborated on in Sect. 4.1, is by counting the number of items selected from group t.

Suppose $\mathcal{F}$ contains all feasible subsets. For example, if there exists a constraint that limits the selection of items to k, then $\mathcal{F}$ can be defined as $\{S \subseteq V \mid |S| \leq k\}$. The objective of our problem, denoted as **P.0**, is to find a distribution $x \in [0, 1]^{\mathcal{F}}$ over $\mathcal{F}$ that maximizes the expected global utility, while also ensuring that the minimum expected utility from each group is met to comply with the fairness constraints. Here the decision variable $x_S \in [0, 1]$ specifies the selection probability of each feasible subset $S \in \mathcal{F}$. A formal definition of **P.0** is listed in below.

P.0 $\max_{x \in [0,1]^{\mathcal{F}}} \sum_{S \in \mathcal{F}} x_S f(S)$
subject to:

$$\begin{cases} \sum_{S \in \mathcal{F}} x_S \cdot g_t(S) \geq \alpha_t, \forall t \in [m]. \\ \sum_{S \in \mathcal{F}} x_S \leq 1. \end{cases}$$

This LP has $m + 1$ constraints, excluding the standard constraints of $x_S \geq 0$ for all $S \in \mathcal{F}$. However, the number of variables in the LP problem is equivalent to the size of $\mathcal{F}$, which can become exponential in n. Due to this, conventional LP solvers are unable to solve this LP problem efficiently.

3 Approximation algorithm for P.0

Before presenting our algorithm, we first introduce a combinatorial optimization problem called FAIRMAX. A solution to this problem serves as a subroutine of our final algorithm.

Definition 1 (*FairMax*) Given functions f and $g_1, g_2, \ldots, g_m$, a vector $z \in \mathbb{R}^m_{\geq 0}$ and a set of feasible subsets $\mathcal{F}$, $\textsc{FairMax}(z, \mathcal{F})$ aims to

$$\max_{S \in \mathcal{F}} \left(f(S) + \sum_{t \in [m]} g_t(S) \cdot z_t \right). \tag{1}$$

I.e., $\textsc{FairMax}(z, \mathcal{F})$ seeks to find the feasible subset $S \in \mathcal{F}$ that maximizes $f(S) + \sum_{t \in [m]} g_t(S) \cdot z_t$.

We next present the main theorem of this paper. A solution $y \in [0, 1]^{\mathcal{F}}$ is said to achieve a (a, b)-approximation for **P.0** if it satisfies the following conditions: $\sum_{S \in \mathcal{F}} y_S \leq 1, \sum_{S \in \mathcal{F}} y_S f(S) \geq a \times OPT$, where OPT denotes the optimal solution of **P.0**, and $\sum_{S \in \mathcal{F}} y_S g_t(S) \geq b \times \alpha_t, \forall t \in [m]$. Here, the a represents the approximation ratio, while the b indicates that the solution may violate the fairness constraint by at most a factor of μ. The following theorem establishes a connection between $\textsc{FairMax}(z, \mathcal{F})$ and **P.0**.

Theorem 1 *Suppose that for all $z \in \mathbb{R}^m_{\geq 0}$, there exists a polynomial-time algorithm that returns a set $A \in \mathcal{F}$ such that*

$$\forall S \in \mathcal{F}, f(A) + \sum_{t \in [m]} g_t(A) \cdot z_t \geq \rho \cdot f(S) + \mu \cdot \sum_{t \in [m]} g_t(S) \cdot z_t,$$

*for some $\rho, \mu \in [0, 1]$. Then there exists a polynomial-time (ρ, μ)-approximation algorithm for **P.0**.*

To prove this theorem, it suffices to present a polynomial (ρ, μ)-approximation algorithm for **P.0**, using a polynomial-time approximation algorithm for $\textsc{FairMax}(z, \mathcal{F})$ as a subroutine. To this end, we will investigate a relaxed form of **P.0**, which we refer to as **RP.0**.

RP.0 $\max_{x \in [0,1]^{\mathcal{F}}} \sum_{S \in \mathcal{F}} x_S f(S)$
subject to:

$$\begin{cases} \sum_{S \in \mathcal{F}} x_S \cdot g_t(S) \geq \mu \alpha_t, \forall t \in [m]. \\ \sum_{S \in \mathcal{F}} x_S \leq 1. \end{cases}$$

RP.0 is obtained by loosening the fairness constraint α_t in **P.0** by a factor of $\mu \in [0, 1]$, where μ is defined in Theorem 1. By solving **RP.0**, we can obtain a solution that is approximately feasible for **P.0**. Given the assumptions made in Theorem 1, in the following, we will focus on finding a solution for **RP.0** and show that this solution constitutes a bicriteria (ρ, μ)-approximation solution for the original problem **P.0**.

Note that the number of variables in **RP.0** is equal to the number of elements in $\mathcal{F}$, which can become very large when n is substantial. This results in standard LP solvers being unable to efficiently solve this LP problem. To address this challenge, we turn to the dual problem of **RP.0** and use the ellipsoid algorithm (Grötschel et al. 1981) to solve it.

The dual problem of **RP.0** (labeled as **Dual of RP.0**) involves assigning a weight $z_t \in \mathbb{R}_{\geq 0}$ to each group t and introducing an additional variable, $w \in \mathbb{R}_{\geq 0}$.

Dual of RP.0 $\min_{z \in \mathbb{R}^m_{\geq 0}, w \in \mathbb{R}_{\geq 0}} \sum_{t \in [m]} -\mu \alpha_t z_t + w$
subject to: $w \geq f(S) + \sum_{t \in [m]} g_t(S) \cdot z_t, \forall S \in \mathcal{F}.$

Observe that the number of constraints in **Dual of RP.0** might be exponential in n. At a high level, we aim to simplify this problem by reducing the number of constraints to a polynomial amount, without significantly altering the optimal solution.

We will now formally present our algorithm for **Dual of RP.0** which involves a series of iterations known as the ellipsoid algorithm. During each iteration, the ellipsoid algorithm is used to determine whether the current solution is feasible or not by approximately solving an instance of FAIRMAX. This problem acts as a test of feasibility and serves as a separation oracle, which helps to determine whether the current solution is located inside or outside the feasible region of the problem being solved. Let $C(L)$ denote the set of ($z \in \mathbb{R}^m_{\geq 0}$, $w \in \mathbb{R}_{\geq 0}$) satisfying that

$$\sum_{t \in [m]} -\mu \alpha_t z_t + w \leq L,$$
$$w \geq f(S) + \sum_{t \in [m]} g_t(S) \cdot z_t, \quad \forall S \in \mathcal{F}.$$

It is easy to verify that L is achievable with respect to **Dual of RP.0** if and only if $C(L)$ is non-empty. To find the minimum value of L such that $C(L)$ is non-empty, we use a binary search algorithm.

For a given L and (z, w), we first evaluate the inequality $\sum_{t \in [m]} -\mu \alpha_t z_t + w \leq L$. If the inequality holds, the algorithm runs a subroutine $\mathcal{A}$ to solve FAIRMAX$(z, \mathcal{F})$. Specifically, $\mathcal{A}$ aims to find the feasible subset $S \in \mathcal{F}$ that maximizes $f(S) + \sum_{t \in [m]} g_t(S) \cdot z_t$. Let A denote the set returned by $\mathcal{A}$.

- If the condition $f(A) + \sum_{t \in [m]} g_t(A) \cdot z_t \leq w$ holds, we mark $C(L)$ as a non-empty set (note that even in this case, $C(L)$ might still be empty because A is only an approximate solution of FAIRMAX$(z, \mathcal{F})$. However, as we will demonstrate later, this will not significantly impact our final solution). In such a scenario, we conclude that L is achievable and proceed to try a smaller value of L.
- If $f(A) + \sum_{t \in [m]} g_t(A) \cdot z_t > w$, this means that $(z, w) \notin C(L)$, and hence, A is a separating hyperplane. To continue the optimization process, we search for a smaller ellipsoid with a center that satisfies this constraint. We repeat this process until we either find a feasible solution in $C(L)$, in which case we attempt a smaller L, or until the volume of the bounding ellipsoid becomes so small that it is considered empty with respect to $C(L)$. In the latter case, we conclude that the current objective is unattainable and will therefore try a larger L.

For a detailed understanding of the individual steps required to run ellipsoid with separation oracles and attain (multiplicative and additive) approximate guarantees, we recommend referring to Chapter 2 of Bubeck (2015). After obtaining the results from the above ellipsoid methods, the subsequent procedures will encompass two primary steps. Firstly, an upper bound for the optimal solution of **P.0** will be calculated. Following that, a (ρ, μ)-approximation solution for **P.0** will be computed.

Establishing an upper bound on the optimal solution of **P.0**. Define L^* to be the smallest value of L for which $C(L)$ is marked as non-empty by our algorithm. We next show that the optimal solution of **P.0** is at most L^*/ρ. To avoid trivial cases, let us assume that $\rho > 0$.

Because $C(L^*)$ is marked as non-empty, there exists a (z^*, w^*) such that

$$\sum_{t\in[m]} -\mu\alpha_t z_t^* + w^* \leq L^* \tag{2}$$

and

$$f(A) + \sum_{t\in[m]} g_t(A) \cdot z_t^* \leq w^*. \tag{3}$$

Given the assumption made regarding A as stated in Theorem 1, we have

$$\forall S \in \mathcal{F},\ f(A) + \sum_{t\in[m]} g_t(A) \cdot z_t^* \geq \rho \cdot f(S) + \mu \cdot \sum_{t\in[m]} g_t(S) \cdot z_t^*. \tag{4}$$

Because $\rho > 0$, it follows that

$$\forall S \in \mathcal{F},\ f(S) + \frac{\mu}{\rho} \cdot \sum_{t\in[m]} g_t(S) \cdot z_t^* \leq (f(A) + \sum_{t\in[m]} g_t(A) \cdot z_t^*)/\rho \leq w^*/\rho \tag{5}$$

where the first inequality follows from (4) and the second inequality is by inequality (3).

Consider the dual of **P.0** (labeled as **Dual of P.0**), inequality (5) implies that $(\frac{\mu}{\rho} \cdot z^*, \frac{1}{\rho} \cdot w^*)$ is a feasible solution of **Dual of P.0**.

Dual of P.0 $\min_{z\in\mathbb{R}^m_{\geq 0}, w\in\mathbb{R}_{\geq 0}} -\alpha_t z_t + w$
subject to: $w \geq f(S) + \sum_{t\in[m]} g_t(S) \cdot z_t, \forall S \in \mathcal{F}$.

Plugging $(\frac{\mu}{\rho} \cdot z^*, \frac{1}{\rho} \cdot w^*)$ into the objective function of **Dual of P.0**, we can infer that the value of **Dual of P.0** is at most $\sum_{t\in[m]} -\mu\alpha_t z_t^*/\rho + w^*/\rho \leq L^*/\rho$ where the inequality is by (2). By strong duality, the value of **P.0** is at most L^*/ρ. Hence, by finding a solution to **RP.0** with a value of L^*, we can achieve an (ρ, μ)-approximation for the original problem **P.0**.

Finding a solution to **RP.0** *with a value of* $L^* - \epsilon$. Suppose $L^* - \epsilon$ is the largest value of L for which the algorithm identifies that $C(L)$ is empty. Here, ϵ denotes the precision of the binary search. We next focus on finding a solution to **RP.0** with a value of $L^* - \epsilon$. To this end we can utilize only the feasible subsets from $\mathcal{F}$ that correspond to the separating hyperplanes obtained by the separation oracle. To achieve this, we define a subset of $\mathcal{F}$, denoted as $\mathcal{F}'$, which contains all the feasible subsets for which the dual constraint is violated during the implementation of the ellipsoid algorithm on $C(L^* - \epsilon)$. The size of $\mathcal{F}'$ is polynomial since the dual constraints are violated for only a polynomial number of feasible subsets. We can use the feasible subsets in $\mathcal{F}'$ to construct a polynomial sized dual linear program of **RP.0** (labeled as **Poly-sized Dual of P.0**).

Poly-zied Dual of RP.0 $\min_{z\in\mathbb{R}^m_{\geq 0}, w\in\mathbb{R}_{\geq 0}} \sum_{t\in[m]} -\mu\alpha_t z_t + w$
subject to: $w \geq f(S) + \sum_{t\in[m]} g_t(S)\cdot z_t, \forall S \in \mathcal{F}'.$

The objective of **Poly-sized Dual of RP.0** is to maximize the dual objective function subject to the constraints defined by the feasible subsets in $\mathcal{F}'$. Because $C(L^* - \epsilon)$ is empty, the value of **Poly-sized Dual of RP.0** at least $L^* - \epsilon$. Hence, the optimal solution to the dual of **Poly-sized Dual of RP.0** (labeled as **Poly-sized RP.0**) is at least $L^* - \epsilon$.

Poly-sized RP.0
$\max_{x\in[0,1]^{\mathcal{F}'}} \sum_{S\in\mathcal{F}'} x_S f(S)$
subject to:

$$\begin{cases} \sum_{S\in\mathcal{F}'} x_S \cdot g_t(S) \geq \mu\alpha_t, \forall t \in [m]. \\ \sum_{S\in\mathcal{F}'} x_S \leq 1. \end{cases}$$

It is important to note that the size of **Poly-sized RP.0** is polynomial, since $\mathcal{F}'$ contains only a polynomial number of feasible subsets. Thus, we can solve **Poly-sized RP.0** efficiently and obtain a solution with a value of $L^* - \epsilon$. This solution is a (ρ, μ)-approximation (with additive error ϵ) for **P.0**.

4 Applications

This section covers a range of applications for our framework, some of which yield better results than previously known methods. In other cases, we present new applications and provide the first approximation algorithms for them.

4.1 Submodular maximization with group fairness constraints

In this problem, we make the assumption that the global utility function f and m group utility functions $g_1, g_2, \ldots, g_m$ are submodular.[1]. This problem setting is general enough to cover a wide range of optimization problems that can be modeled using submodular global utility functions. Examples of such applications include data summarization (El Halabi et al. 2020), influence maximization (Kempe and Mahdian 2008), and information retrieval (Yue and Guestrin 2011). It is worth noting that this

[1] A function $h : 2^V \rightarrow \mathbb{R}$ is considered submodular if, for any sets X and Y that are subsets of V with $X \subseteq Y$ and any item $e \in V \setminus Y$, the following inequality holds: $h(X \cup \{e\}) - h(X) \geq h(Y \cup \{e\}) - h(Y)$. It is considered monotone if, for any set $X \subseteq V$ and any item $e \in V \setminus X$, it holds that $h(X \cup \{e\}) - h(X) \geq 0$

scenario encompasses *submodular maximization under submodular coverage* (Ohsaka and Matsuoka 2021) as a special case. Their objective is to maximize a monotone submodular function subject to a lower quota constraint on a single submodular group utility function. However, their focus is on finding a deterministic solution set, whereas our approach provides greater flexibility in achieving group fairness. Next, we discuss the rationale behind assuming that the group utility function is submodular. In many prior works (El Halabi et al. 2020; Celis et al. 2018), the concept of balance with respect to a sensitive attribute (such as race or gender) has been a widely used criterion for evaluating the solution obtained by fairness-aware optimization algorithms. This notion of balance typically involves ensuring that the solution does not significantly disadvantage any particular group with respect to the sensitive attribute, while also achieving good performance on the global objective f. We next provide a specific example in this context. Consider a set V of n items (such as people), where each item is associated with a sensitive attribute. Let $V_1, \dots, V_m$ denote the m groups of items with the same attribute. We define a solution $x \in [0, 1]^{\mathcal{F}}$, which encodes the selection probability of each set from $\mathcal{F}$, to be fair if the expected number of selected items from each group V_t is at least α_t, where α_t often set proportional to the fraction of items of V_t in the entire set V. In this case, we define $g_t(S) = |S \cap V_t|$ for each $t \in [m]$. It is easy to verify that g_t is a monotone and submodular function.

Recall that solving **P.0** is reduced to solving FairMax$(z, \mathcal{F})$ (Definition 1). Here the objective of FairMax$(z, \mathcal{F})$ is to $\max_{S \in \mathcal{F}}(f(S) + \sum_{t\in[m]} g_t(S) \cdot z_t)$. Fortunately, if f and g_t are submodular functions, $f(S) + \sum_{t\in[m]} g_t(S) \cdot z_t$ is also a submodular function by the fact that a linear combination of submodular functions is still submodular. If we assume that the family $\mathcal{F}$ is defined based on cardinality constraints, specifically as $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ for a positive integer k, then the problem of FairMax$(z, \mathcal{F})$ can be reduced to maximizing a submodular function while satisfying a cardinality constraint. According to Nemhauser et al. (1978), if the objective function is non-negative, monotone and submodular, then an optimal $(1 - 1/e)$-approximation algorithm exists for this problem, that is, $\rho = \mu = 1 - 1/e$. On the other hand, if the objective function is non-monotone (in addition to non-negative and submodular), then an approximation of 0.385 is possible (Buchbinder and Feldman 2019), that is, $\rho = \mu = 0.385$. This, together Theorem 1, implies the following proposition.

Proposition 1 *If f and g_t are non-negative monotone submodular functions, and $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ for a positive integer k, then there exists an optimal $(1 - 1/e, 1 - 1/e)$-approximation algorithm for **P.0**. If f and g_t are non-negative non-monotone submodular functions, there exists a $(0.385, 0.385)$-approximation algorithm for **P.0**.*

4.1.1 Improved results for monotone submodular *f* and modular g_t

We next investigate an important special case of this application where we make the assumption that the global utility function f is non-negative monotone and submodular; m group utility functions $g_1, g_2, \dots, g_m$ are modular functions. It is easy to verify that the example previously discussed, wherein $g_t(S) = |S \cap V_t|$, satisfies the properties of a modular group utility function. We show that there exists a *feasible* optimal $(1 - 1/e)$-approximation algorithm for this special case.

Observe that if f is non-negative monotone and submodular, g_t is a modular function (hence $\sum_{t\in[m]} g_t(\cdot)\cdot z_t$ is also a modular function), and $\mathcal{F} = \{S \subseteq V \mid |S| \le k\}$ for a positive integer k, then Sviridenko et al. (2017) presented a randomized polynomial-time algorithm that produces a set $A \in \mathcal{F}$ such that for every $S \in \mathcal{F}$, it holds that

$$f(A) + \sum_{t\in[m]} g_t(A) \cdot z_t \ge (1 - 1/e) f(S) + \sum_{t\in[m]} g_t(S) \cdot z_t. \tag{6}$$

That is, $\rho = 1 - 1/e$ and $\mu = 1$. Substituting these values into Theorem 1, we have the following proposition. Note that $\mu = 1$ indicates that our solution strictly satisfies all fairness constraints.

Proposition 2 *If the global utility function f is a non-negative monotone submodular function; m group utility functions $g_1, g_2, \dots, g_m$ are modular functions; $\mathcal{F} = \{S \subseteq V \mid |S| \le k\}$ for a positive integer k, then there exists a* feasible $(1 - 1/e, 1)$-*approximation algorithm for **P.0**.*

Note that the above result may be subject to a small additive error due to the omission of a similar error present in the original result (Inequality (6)) presented in Sviridenko et al. (2017), which has been left out for simplicity.

4.2 Sequential submodular maximization

This problem was first studied in Asadpour et al. (2022) where the objective is to determine the optimal *ordering* of a set of items to maximize a linear combination of various submodular functions. This variant of submodular maximization arises from the scenario where a platform displays a list of products to a user. The user examines the first l items in the list, where l is randomly selected from a given distribution and the user's decision to purchase an item from the set depends on a choice model, resulting in the platform's goal of maximizing the engagement of the shopper, which is defined as the probability of purchase. Formally, we are given monotone submodular functions $h_1, \dots, h_n : 2^V \to \mathbb{R}_{\ge 0}$ and $h_1^t, \dots, h_n^t : 2^V \to \mathbb{R}_{\ge 0}$ for each group $t \in [m]$, nonnegative coefficients $\lambda_1, \dots, \lambda_n$ and $\lambda_1^t, \dots, \lambda_n^t$ for each group $t \in [m]$. By abuse of notation, let S be a permutation over items in V and let $S_{[l]}$ represent the first l items in S. Define the global utility function as $f(S) = \sum_{l\in[n]} \lambda_l h_l(S_{[l]})$. In the context of product ranking, λ_l represents the fraction of users with patience level l, while h_l corresponds to the aggregate purchase probability function of users with patience level l. Similarly, the group utility function is defined as $g_t(S) = \sum_{l\in[n]} \lambda_l^t h_l^t(S_{[l]})$ for each group of users $t \in [m]$, where λ_l^t should be interpreted as the fraction of users with patience level l in group t. Despite f and g_t being defined over permutations instead of sets, it is easy to verify that Theorem 1 remains valid. In particular, suppose for all $z \in \mathbb{R}^m_{\ge 0}$, there exists a polynomial-time algorithm that returns a permutation $A \in \mathcal{F}$ such that

$$\forall S \in \mathcal{F},\ f(A) + \sum_{t\in[m]} g_t(A) \cdot z_t \ge \rho \cdot f(S) + \mu \cdot \sum_{t\in[m]} g_t(S) \cdot z_t,$$

for some $\rho, \mu \in [0, 1]$. Here $\mathcal{F}$ is the set of all possible permutations over items in V. Then there exists a polynomial-time (ρ, μ)-approximation algorithm for the sequential submodular maximization problem.

Observe that in this case, the objective function of FairMax$(z, \mathcal{F})$ can be written as

$$f(S) + \sum_{t\in[m]} g_t(S)\cdot z_t = \sum_{l\in[n]} \lambda_l h_l(S_{[l]}) + \sum_{t\in[m]} \left(\sum_{l\in[n]} \lambda_l^t h_l^t(S_{[l]}) \cdot z_t \right) \tag{7}$$

$$= \sum_{l\in[n]} \lambda_l h_l(S_{[l]}) + \sum_{l\in[n]} \left(\sum_{t\in[m]} \lambda_l^t h_l^t(S_{[l]}) \cdot z_t \right) \tag{8}$$

$$= \sum_{l\in[n]} \left(\lambda_l h_l(S_{[l]}) + \sum_{t\in[m]} \lambda_l^t h_l^t(S_{[l]}) \cdot z_t \right). \tag{9}$$

If h_l and h_l^t are both monotone and submodular for all $l \in [n]$ and $t \in [m]$, then $\lambda_l h_l(\cdot) + \sum_{t\in[m]} z_t \lambda_l^t h_l^t(\cdot)$ is also monotone and submodular for all $l \in [n]$ by the fact that a linear combination of monotone submodular functions is still monotone and submodular. According to the analysis presented in Theorem 1 of Asadpour et al. (2022), if $\lambda_l h_l(\cdot) + \sum_{t\in[m]} z_t \lambda_l^t h_l^t(\cdot)$ is monotone and submodular for all $l \in [n]$, the problem of identifying a permutation S that maximizes the right-hand side of Eq. (7) can be transformed into a submodular maximization problem subject to a (laminar) matroid constraint. Hence, there exists a $(1-1/e)$-approximation algorithm (Calinescu et al. 2007) for FairMax$(z, \mathcal{F})$, that is, $\rho = \mu = 1-1/e$. Using Theorem 1, we obtain a $(1 - 1/e, 1 - 1/e)$-approximation algorithm for **P.0**. Note that the current state-of-the-art result (Asadpour et al. 2022) provides a bi-criteria $((1 - 1/e)^2, (1 - 1/e)^2)$-approximation for **P.0**. Our proposed framework offers significant improvements over their results in both approximation ratio and feasibility.

Proposition 3 *There exists an optimal* $(1 - 1/e, 1 - 1/e)$*-approximation algorithm for sequential submodular maximization.*

4.3 Random assortment planning with group market share constraints

The third application concerns assortment planning, which is a problem that is widely recognized within the operations research community. Assortment planning with group market share constraints aims to identify the best possible combination of products to present to customers while ensuring that a minimum market share requirement of each group is met. Existing studies on this problem focus on finding a *deterministic* solution that meets a minimum market share of a *single* group. We extend this study to consider a randomized setting with multiple groups. Formally, this problem takes a set V of n products as input, which is divided into m (possibly non-disjoint) groups denoted by $V_1, V_2, \ldots, V_m$. Under the well-known multinomial logit (MNL) model, each product $i \in V$ has a preference weight v_i and let v_0 denote the preference for no purchase. Let r_i denote the revenue of selling a product $i \in V$. Given an assortment $S \subseteq V$, the purchase probability of any product $i \in S$ is $\frac{v_i}{v_0+\sum_{i\in S} v_i}$. Hence the expected revenue of offering S is $f(S) = \frac{\sum_{i\in S} r_i v_i}{v_0+\sum_{i\in S} v_i}$ and the resulting market share of group $t \in [m]$ is $g_t(S) = \frac{\sum_{i\in S\cap V_t} v_i}{v_0+\sum_{i\in S} v_i}$. Assume $\mathcal{F}$ is comprised of all possible subsets of V, the goal of **P.0** is to compute the selection probability x_S of each assortment of products $S \subseteq V$ such that the expected revenue $\sum_{S\in\mathcal{F}} x_S f(S)$ is maximized while the expected market share is at least $\sum_{S\in\mathcal{F}} x_S g_t(S) \geq \alpha_t$ for each group $t \in [m]$. To solve this problem, we consider its corresponding FAIRMAX$(z, \mathcal{F})$, whose objective is to find a $S \subseteq V$ that maximizes $f(S) + \sum_{t\in[m]} g_t(S) \cdot z_t$ for a given vector $z \in \mathbb{R}^m_{\geq 0}$. By the definitions of f and g_t, the objective function of FAIRMAX$(z, \mathcal{F})$ can be written as

$$f(S) + \sum_{t\in[m]} g_t(S) \cdot z_t = \frac{\sum_{i\in S} r_i v_i}{v_0 + \sum_{i\in S} v_i} + \sum_{t\in[m]} \frac{\sum_{i\in S\cap V_t} v_i}{v_0 + \sum_{i\in S} v_i} \cdot z_t \tag{10}$$

$$= \frac{\sum_{i\in S}(r_i + \sum_{t\in[m]} z_t \cdot \mathbf{1}_{i\in V_t})v_i}{v_0 + \sum_{i\in S} v_i}, \tag{11}$$

where $\mathbf{1}_{i\in V_t} \in \{0, 1\}$ is an indicator variable such that $\mathbf{1}_{i\in V_t} = 1$ if and only if $i \in V_t$.

Hence, the goal of FAIRMAX$(z, \mathcal{F})$ is to

$$\max_{S\subseteq V} \frac{\sum_{i\in S}(r_i + \sum_{t\in[m]} z_t \cdot \mathbf{1}_{i\in V_t})v_i}{v_0 + \sum_{i\in S} v_i}. \tag{12}$$

This problem can be viewed as an unconstrained assortment planning problem, where each product $i \in V$ has a revenue of $(r_i + \sum_{t\in[m]} z_t \cdot \mathbf{1}_{i\in V_t})$ and a preference weight of v_i, and the preference weight of no purchase is v_0. According to Talluri and Van Ryzin (2004), the optimal solution for this problem is a revenue-ordered assortment. In other words, the assortment that maximizes revenue consists of the l products with the highest revenues $(r_i + \sum_{t\in[m]} z_t \cdot \mathbf{1}_{i\in V_t})$, where $l \in [n]$. As a result, FAIRMAX$(z, \mathcal{F})$ (problem (12)) can be solved optimally in polynomial time by examining at most n potential assortments. Thus, the combination of Theorem 1 and the ability to solve FAIRMAX$(z, \mathcal{F})$ optimally in polynomial time (i.e., $\rho = \mu = 1$) implies that an optimal solution for **P.0** exists.

Proposition 4 *There exists an optimal and feasible algorithm for assortment planning with group market share constraints.*

5 Discussion on other variants of fairness notations

In this section, we examine two additional notations of fairness that are frequently employed in the literature.

5.1 Incorporating fairness upper bound constraints β_t

One natural way to extend the fairness notation we introduced in **P.0** would be to impose further upper bounds β_t on the expected utility of every group. Formally,

P.A $\max_{x\in[0,1]^{\mathcal{F}}} \sum_{S\in\mathcal{F}} x_S f(S)$
subject to:

$$\begin{cases}\alpha_t \leq \sum_{S\in\mathcal{F}} x_S \cdot g_t(S) \leq \beta_t, \forall t \in [m].\\ \sum_{S\in\mathcal{F}} x_S \leq 1.\end{cases}$$

This general formulation can be seen in several previous studies on fairness-aware optimization, such as Celis et al. (2018), El Halabi et al. (2020). Fortunately, we can still use ellipsoid method to solve **P.A** to obtain a bicriteria algorithm. As we will see later, the following problem of FairMax serves as a separation oracle, which helps to determine whether the current solution is located inside or outside the feasible region of the problem being solved.

Definition 2 (*FairMax*) Given functions f and $g_1, g_2, \ldots, g_m$, two vectors $z \in \mathbb{R}^m_{\geq 0}$ and $u \in \mathbb{R}^m_{\geq 0}$, and a set of feasible subsets $\mathcal{F}$, FairMax$(z, u, \mathcal{F})$ aims to

$$\max_{S\in\mathcal{F}}(f(S) + \sum_{t\in[m]} g_t(S) \cdot (z_t - u_t)). \tag{13}$$

Unlike the objective function in (1), the coefficient of g_t in the above utility function might take on negative values. The next theorem builds a connection between FairMax$(z, u, \mathcal{F})$ and **P.A**. Here we extend the definition of (a, b)-approximation such that a solution $y \in \mathbb{R}^m_{\geq 0}$ is said to achieve a (a, b)-approximation for **P.A** if it satisfies the following conditions: $\sum_{S\in\mathcal{F}} y_S \leq 1$, $\sum_{S\in\mathcal{F}} y_S f(S) \geq a \times OPT$, where OPT denotes the optimal solution of **P.A**, and $\beta_t \geq \sum_{S\in\mathcal{F}} y_S g_t(S) \geq b \times \alpha_t, \forall t \in [m]$. The proof of the following theorem is moved to appendix.

Theorem 2 *Assuming for all $z \in \mathbb{R}^m_{\geq 0}$ and $u \in \mathbb{R}^m_{\geq 0}$, there exists a polynomial-time algorithm that returns a set $A \in \mathcal{F}$ such that*

$$\forall S \in \mathcal{F}, f(A) + \sum_{t\in[m]} g_t(A) \cdot z_t \geq \rho \cdot f(S) + \mu \cdot \sum_{t\in[m]} g_t(S) \cdot (z_t - u_t),$$

*for some $\rho, \mu \in [0, 1]$. Then there exists a polynomial-time (ρ, μ)-approximation algorithm for **P.A**.*

Observe that if f is non-negative monotone and submodular, g_t is a modular function (hence $\sum_{t\in[m]} g_t(\cdot) \cdot (z_t - u_t)$ is also a modular function), and $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ for a positive integer k, then Sviridenko et al. (2017) presented a randomized polynomial-time algorithm that produces a set $A \in \mathcal{F}$ such that for every $S \in \mathcal{F}$, it holds that

$$f(A) + \sum_{t\in[m]} g_t(A) \cdot (z_t - u_t) \geq (1 - 1/e) f(S) + \sum_{t\in[m]} g_t(S) \cdot (z_t - u_t). \tag{14}$$

The following proposition follows immediately from Theorem 2 and inequality (14). Note that this result recovers the findings from Tang et al. (2023).

Proposition 5 *If the global utility function f is a non-negative monotone submodular function; m group utility functions $g_1, g_2, \dots, g_m$ are modular functions; $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ for a positive integer k, then there exists a* feasible $(1 - 1/e, 1)$-*approximation algorithm for* ***P.A***.

For the case when f is non-negative *non-monotone* and submodular, g_t is a modular function, and $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ for a positive integer k, Qi (2022) presented an algorithm for FairMax that achieves $\rho = \frac{te^{-t}}{t+e^{-t}} - \epsilon$ and $\mu = \frac{t}{t+e^{-t}}$ for every constant $t \in [0, 1]$. This, together with Theorem 2, indicates a $(\frac{te^{-t}}{t+e^{-t}} - \epsilon, \frac{t}{t+e^{-t}})$-approximation algorithm for **P.A**.

5.2 Pairwise fairness

We will now explore another frequently utilized notation of fairness, which relies on achieving parity in pairwise utility between groups. This type of notation has been employed in diverse scenarios, such as recommendation systems (Beutel et al. 2019), assortment planning (Chen et al. 2022), ranking and regression models (Narasimhan et al. 2020), and predictive risk scores (Kallus and Zhou 2019). Under this notion, it is expected that groups will experience comparable levels of utility. The extent of fairness is determined by a parameter γ. Specifically, we require that for every two groups $t, t' \in [m]$, the difference between their expected utilities is at most γ, i.e., $\sum_{S\in\mathcal{F}} x_S \cdot g_t(S) - \sum_{S\in\mathcal{F}} x_S \cdot g_{t'}(S) \leq \gamma, \forall t, t' \in [m]$. This problem is formally defined in **P.B**. Our algorithmic findings extend to an even more general version of this problem by introducing a distinct γ for each pair of groups. For the sake of simplicity, we do not elaborate on it here.

P.B $\max_{x\in[0,1]^{\mathcal{F}}} \sum_{S\in\mathcal{F}} x_S f(S)$
subject to:

$$\begin{cases} \sum_{S\in\mathcal{F}} x_S \cdot g_t(S) - \sum_{S\in\mathcal{F}} x_S \cdot g_{t'}(S) \leq \gamma, \forall t, t' \in [m]. \\ \sum_{S\in\mathcal{F}} x_S \leq 1. \end{cases}$$

In contrast to our approach for other fairness notations, we do not transform the original problem into a relaxed form for this particular case. Instead, we solve the dual of **P.B** directly. The dual of **P.B** is presented in **Dual of P.B**.

Dual of P.B $\min_{z\in\mathbb{R}^{m\times m}_{\geq 0}, w\in\mathbb{R}_{\geq 0}} \gamma \sum_{t,t'\in[m]} z_{t,t'} + w$
subject to: $w \geq f(S) + \sum_{t,t'\in[m]}(g_{t'}(S) - g_t(S))z_{t,t'}, \forall S \in \mathcal{F}.$

To solve **Dual of P.B** using ellipsoid method, we define its separation oracle in FairMax.

Definition 3 (*FairMax*) Given functions f and $g_1, g_2, \ldots, g_m$, a matrix $z \in \mathbb{R}_{\geq 0}^{m\times m}$ and a set of feasible subsets $\mathcal{F}$, $\textsc{FairMax}(z, \mathcal{F})$ aims to

$$\max_{S\in\mathcal{F}}(f(S) + \sum_{t,t'\in[m]}(g_{t'}(S) - g_t(S))z_{t,t'}). \tag{15}$$

The next theorem builds a connection between $\textsc{FairMax}(z, \mathcal{F})$ and **P.B**. In contrast to our results for other fairness notations, where we can only anticipate a bicriteria solution, we show that solving $\textsc{FairMax}(z, \mathcal{F})$ approximately leads to a *feasible* solution for **P.B**. The proof of the following theorem is moved to appendix.

Theorem 3 *Suppose that for all $z \in \mathbb{R}_{\geq 0}^{m\times m}$, there exists a polynomial-time algorithm that returns a set $A \in \mathcal{F}$ such that*

$$\forall S \in \mathcal{F},\ f(A) + \sum_{t,t'\in[m]}(g_{t'}(A) - g_t(A))z_{t,t'} \geq \rho\cdot f(S)$$
$$+\mu\cdot\sum_{t,t'\in[m]}(g_{t'}(S) - g_t(S))z_{t,t'},$$

for some $\rho, \mu \in [0, 1]$. Then there exists a feasible ρ-approximation algorithm for ***P.B****.*

It should be noted that the above performance bound does not depend on μ. Observe that if f is non-negative monotone and submodular, g_t is a modular function for all $t \in [m]$ (hence $\sum_{t,t'\in[m]}(g_{t'}(\cdot) - g_t(\cdot))z_{t,t'}$ is also a modular function for all $t, t' \in [m]$), and $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ for a positive integer k, then Sviridenko et al. (2017) presented a randomized polynomial-time algorithm that produces a set $A \in \mathcal{F}$ such that, $\forall S \in \mathcal{F}$,

$$f(A) + \sum_{t,t'\in[m]}(g_{t'}(A) - g_t(A))z_{t,t'} \geq (1 - 1/e)f(S)$$
$$+\sum_{t,t'\in[m]}(g_{t'}(S) - g_t(S))z_{t,t'}. \tag{16}$$

The following proposition follows immediately from Theorem 3 and inequality (16).

Proposition 6 *If the global utility function f is a non-negative monotone submodular function; m group utility functions $g_1, g_2, \ldots, g_m$ are modular functions; $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ for a positive integer k, then there exists a feasible $(1 - 1/e)$-approximation algorithm for* ***P.B****.*

Remark A recent work by Chen et al. (2022) proposed an ellipsoid-based method for the assortment planning problem that includes pairwise fairness constraints. In page 14 of Chen et al. (2022) (the October 28th, 2022 version), they discussed the case where all items have uniform revenues. They showed that for this special case, their separation oracle is to maximize the summation of a non-negative monotone submodular function and a (not necessarily positive) modular function. They claimed that this objective function is a non-monotone submodular function and suggested using the continuous double greedy algorithm proposed in Buchbinder et al. (2014) to obtain a $[1/e + 0.004, 1/2]$-approximation solution. However, it should be noted that the algorithm proposed by Buchbinder et al. (2014) only applies when the objective function is non-negative. Unfortunately, in general, the sum of a non-negative monotone submodular function and a (not necessarily positive) modular function can yield a negative value, rendering (Buchbinder et al. 2014)'s algorithm inapplicable. On the other hand, our separation oracle in Definition 3 also has an objective function in this format, but using Sviridenko et al. (2017)'s algorithm as a subroutine to solve it leads to a $(1 - 1/e)$-approximation solution of our original problem. It is easy to verify that our framework (i.e., employing Sviridenko et al. 2017's algorithm to solve the separation oracle) can be applied to the problem examined by Chen et al. (2022) to obtain a $(1 - 1/e)$-approximation solution by conducting an analogous analysis to Theorem 3.

6 Conclusion

In this paper, we introduce a general group fairness notation that unifies many notations used in previous works. We formulate the problem of finding a distribution over solution sets that maximizes global utility while satisfying fairness constraints. We develop a polynomial-time algorithmic framework based on the ellipsoid method to solve this problem. We also develop an optimal $(1-1/e)$-approximation algorithm for a special case of our problem, where f is monotone and submodular, and g_t is a modular function. This solution satisfies all fairness constraints strictly. Our work shows that this formulation brings together and extends numerous noteworthy applications in both machine learning and operations research.

Funding The authors have not disclosed any funding.

Data availability Enquiries about data availability should be directed to the authors.

Declarations

Conflict of interest The authors have not disclosed any competing interests.

7 Appendix

7.1 Proof of Theorem 2

To prove this theorem, it suffices to present a polynomial (ρ, μ)-approximation algorithm for **P.A**, using a polynomial-time approximation algorithm for FAIR-MAX$(z, u, \mathcal{F})$ as a subroutine. We first introduce a relaxed form of **P.A** in **RP.A** where the lower bound constraint is replaced with $\mu\alpha_t$.

RP.A $\max_{x\in[0,1]^{\mathcal{F}}} \sum_{S\in\mathcal{F}} x_S f(S)$
subject to:

$$\begin{cases} \mu\alpha_t \leq \sum_{S\in\mathcal{F}} x_S \cdot g_t(S) \leq \beta_t, \forall t \in [m]. \\ \sum_{S\in\mathcal{F}} x_S \leq 1. \end{cases}$$

The dual of **RP.A** is listed in **Dual of RP.A**.

Dual of RP.A $\min_{z\in\mathbb{R}^m_{\geq 0}, u\in\mathbb{R}^m_{\geq 0}, w\in\mathbb{R}_{\geq 0}} \sum_{t\in[m]}(\beta_t u_t - \mu\alpha_t z_t) + w$
subject to: $w \geq f(S) + \sum_{t\in[m]} g_t(S)\cdot(z_t - u_t), \forall S \in \mathcal{F}$.

Let $C(L)$ denote the set of $(z \in \mathbb{R}^m_{\geq 0}, u \in \mathbb{R}^m_{\geq 0}, w \in \mathbb{R}_{\geq 0})$ satisfying that

$$\sum_{t\in[m]} (\beta_t u_t - \mu\alpha_t z_t) + w \leq L,$$
$$w \geq f(S) + \sum_{t\in[m]} g_t(S)\cdot(z_t - u_t), \quad \forall S \in \mathcal{F}.$$

It is easy to verify that L is achievable with respect to **Dual of RP.A** if and only if $C(L)$ is non-empty. To find the minimum value of L such that $C(L)$ is non-empty, we use a binary search algorithm.

For a given L and (z, u, w), we first evaluate the inequality $\sum_{t\in[m]}(\beta_t u_t - \mu\alpha_t z_t) + w \leq L$. If the inequality holds, the algorithm runs a subroutine $\mathcal{A}$ to solve FairMax$(z, u, \mathcal{F})$. Assuming that $\mathcal{A}$ is a μ-approximation algorithm for FairMax$(z, u, \mathcal{F})$, let A denote the set returned by $\mathcal{A}$.

- If the condition $f(A) + \sum_{t\in[m]} g_t(A)\cdot(z_t - u_t) \leq w$ holds, we mark $C(L)$ as a non-empty set. In such a scenario, we proceed to try a smaller value of L.
- If $f(A) + \sum_{t\in[m]} g_t(A)\cdot(z_t - u_t)\cdot z_t > w$, this means that $(z, u, w) \notin C(L)$, and hence, A is a separating hyperplane. We search for a smaller ellipsoid with a center that satisfies this constraint. We repeat this process until we either find a feasible solution in $C(L)$, in which case we attempt a smaller L, or until the volume of the bounding ellipsoid becomes so small that it is considered empty with respect to $C(L)$. In the latter case, we conclude that the current objective is unattainable and will therefore try a larger L.

Define L^* to be the smallest value of L for which $C(L)$ is marked as non-empty by our algorithm. We next show that the optimal solution of **P.A** is at most L^*/ρ. To avoid trivial cases, let us assume that $\rho > 0$.

Because $C(L^*)$ is marked as non-empty, there exists a (z^*, u^*, w^*) such that

$$\sum_{t\in[m]} (\beta_t u^*_t - \mu\alpha_t z^*_t) + w^* \leq L^* \tag{17}$$

and

$$f(A) + \sum_{t\in[m]} g_t(A)\cdot(z^*_t - u^*_t) \leq w^*. \tag{18}$$

By the assumption made regarding A in Theorem 2, we have $\forall S \in \mathcal{F}$,

$$f(A) + \sum_{t\in[m]} g_t(A) \cdot z_t^* \geq \rho \cdot f(S) + \mu \cdot \sum_{t\in[m]} g_t(S) \cdot (z_t^* - u_t^*). \tag{19}$$

It follows that $\forall S \in \mathcal{F}$,

$$\begin{aligned} & f(S) + \frac{\mu}{\rho} \cdot \sum_{t\in[m]} g_t(S) \cdot (z_t^* - u_t^*) \\ & \quad \leq (f(A) + \sum_{t\in[m]} g_t(A) \cdot (z_t^* - u_t^*))/\rho \leq w^*/\rho \end{aligned} \tag{20}$$

where the first inequality follows from (19) and the second inequality is by inequality (18).

Consider the dual of **P.A** (labeled as **Dual of P.A**), inequality (20) implies that $(\frac{\mu}{\rho} \cdot z^*, \frac{\mu}{\rho} \cdot u^*, \frac{1}{\rho} \cdot w^*)$ is a feasible solution of **Dual of P.A**.

Dual of P.A $\min_{z\in\mathbb{R}^m_{\geq 0}, u\in\mathbb{R}^m_{\geq 0}, w\in\mathbb{R}_{\geq 0}} \sum_{t\in[m]}(\beta_t u_t - \alpha_t z_t) + w$
subject to: $w \geq f(S) + \sum_{t\in[m]} g_t(S) \cdot (z_t - u_t), \forall S \in \mathcal{F}$.

Plugging $(\frac{\mu}{\rho} \cdot z^*, \frac{\mu}{\rho} \cdot u^*, \frac{1}{\rho} \cdot w^*)$ into the objective function of **Dual of P.A**, we can infer that the value of **Dual of P.A** is at most

$$\begin{aligned} \frac{\mu}{\rho} \cdot \sum_{t\in[m]} (\beta_t u_t^* - \alpha_t z_t^*) + w^*/\rho \leq & \sum_{t\in[m]} (\beta_t u_t^*/\rho - \mu\alpha_t z_t^*/\rho) \\ & + w^*/\rho \leq L^*/\rho \end{aligned} \tag{21}$$

where the first inequality is by the observations that $\beta_t u_t^*/\rho \geq 0$ and $\mu \in [0, 1]$, and the second inequality is by (17). By strong duality, the value of **P.A** is at most L^*/ρ. Hence, by finding a solution to **RP.A** with a value of L^*, we can achieve an (ρ, μ)-approximation for the original problem **P.A**.

Suppose $L^* - \epsilon$ is the largest value of L for which the algorithm identifies that $C(L)$ is empty. Here, ϵ denotes the precision of the binary search. We next focus on finding a solution to **RP.A** with a value of $L^* - \epsilon$. Define $\mathcal{F}'$ as the set that contains all the feasible subsets for which the dual constraint is violated during the implementation of the ellipsoid algorithm on $C(L^* - \epsilon)$. We use $\mathcal{F}'$ to construct a polynomial sized dual linear program of **RP.A** (labeled as **Poly-sized Dual of P.A**).

Poly-sizsed Dual of RP.A $\min_{z\in\mathbb{R}^m_{\geq 0}, u\in\mathbb{R}^m_{\geq 0}, w\in\mathbb{R}_{\geq 0}} \sum_{t\in[m]}(\beta_t u_t - \mu\alpha_t z_t) + w$
subject to: $w \geq f(S) + \sum_{t\in[m]} g_t(S) \cdot (z_t - u_t), \forall S \in \mathcal{F}'$.

Because $C(L^* - \epsilon)$ is empty, the value of **Poly-sized Dual of RP.A** at least $L^* - \epsilon$. Hence, the optimal solution to the dual of **Poly-sized Dual of RP.A** (labeled as **Poly-sized RP.A**) is at least $L^* - \epsilon$.

Poly-sized RP.A $\max_{x\in[0,1]^{\mathcal{F}'}} \sum_{S\in\mathcal{F}'} x_S f(S)$
subject to:

$$\begin{cases}\mu\alpha_t \leq \sum_{S\in\mathcal{F}'} x_S \cdot g_t(S) \leq \beta_t, \forall t \in [m].\\ \sum_{S\in\mathcal{F}'} x_S \leq 1.\end{cases}$$

Solving **Poly-sized RP.A** obtains a solution with a value of $L^* - \epsilon$. This solution is a (ρ, μ)-approximation (with additive error ϵ) for **P.A**.

7.2 Proof of Theorem 3

To prove this theorem, it suffices to present a feasible ρ-approximation algorithm for **P.B**, using a polynomial-time approximation algorithm for FAIRMAX$(z, \mathcal{F})$ as a subroutine. Let $C(L)$ denote the set of $(z \in \mathbb{R}^{m\times m}_{\geq 0}, w \in \mathbb{R}_{\geq 0})$ satisfying that

$$\gamma \sum_{t,t'\in[m]} z_{t,t'} + w \leq L,$$

$$w \geq f(S) + \sum_{t,t'\in[m]} (g_{t'}(S) - g_t(S))z_{t,t'}, \forall S \in \mathcal{F}.$$

It is easy to verify that L is achievable with respect to **Dual of P.B** if and only if $C(L)$ is non-empty. To find the minimum value of L such that $C(L)$ is non-empty, we use a binary search algorithm.

For a given L and (z, w), we first evaluate the inequality $\gamma \sum_{t,t'\in[m]} z_{t,t'} + w \leq L$. If the inequality holds, the algorithm runs a subroutine $\mathcal{A}$ to solve FAIRMAX$(z, \mathcal{F})$. Let A denote the set returned by $\mathcal{A}$.

- If the condition $f(A) + \sum_{t,t'\in[m]}(g_{t'}(A) - g_t(A))z_{t,t'} \leq w$ holds, we mark $C(L)$ as a non-empty set. In such a scenario, we proceed to try a smaller value of L.
- If $f(A) + \sum_{t,t'\in[m]}(g_{t'}(A) - g_t(A))z_{t,t'} > w$, this means that $(z, w) \notin C(L)$, and hence, A is a separating hyperplane. We search for a smaller ellipsoid with a center that satisfies this constraint. We repeat this process until we either find a feasible solution in $C(L)$, in which case we attempt a smaller L, or until the volume of the bounding ellipsoid becomes so small that it is considered empty with respect to $C(L)$. In the latter case, we conclude that the current objective is unattainable and will therefore try a larger L.

Define L^* to be the smallest value of L for which $C(L)$ is marked as non-empty by our algorithm. We next show that the optimal solution of **P.B** is at most L^*/ρ. To avoid trivial cases, let us assume that $\rho > 0$.

Because $C(L^*)$ is marked as non-empty, there exists a (z^*, w^*) such that

$$\gamma \sum_{t,t'\in[m]} z^*_{t,t'} + w^* \leq L^* \tag{22}$$

and

$$f(A) + \sum_{t,t' \in [m]} (g_{t'}(A) - g_t(A))z^*_{t,t'} \le w^*. \tag{23}$$

By the assumption made regarding A in Theorem 3, we have $\forall S \in \mathcal{F}$,

$$\begin{aligned} f(A) &+ \sum_{t,t' \in [m]} (g_{t'}(A) - g_t(A))z^*_{t,t'} \\ &\ge \rho \cdot f(S) + \mu \cdot \sum_{t,t' \in [m]} (g_{t'}(S) - g_t(S))z^*_{t,t'}. \end{aligned} \tag{24}$$

It follows that $\forall S \in \mathcal{F}$,

$$\begin{aligned} f(S) &+ \frac{\mu}{\rho} \cdot \sum_{t,t' \in [m]} (g_{t'}(S) - g_t(S))z^*_{t,t'} \\ &\le (f(A) + \sum_{t,t' \in [m]} (g_{t'}(A) - g_t(A))z^*_{t,t'})/\rho \le w^*/\rho \end{aligned} \tag{25}$$

where the first inequality follows from (24) and the second inequality is by inequality (23).

Inequality (25) implies that $(\frac{\mu}{\rho} \cdot z^*, \frac{1}{\rho} \cdot w^*)$ is feasible for **Dual of P.B**.

Dual of P.B $\min_{z \in \mathbb{R}^{m \times m}_{\ge 0}, w \in \mathbb{R}_{\ge 0}} \gamma \sum_{t,t' \in [m]} z_{t,t'} + w$
subject to: $w \ge f(S) + \sum_{t,t' \in [m]} (g_{t'}(S) - g_t(S))z_{t,t'}, \forall S \in \mathcal{F}$.

Plugging $(\frac{\mu}{\rho} \cdot z^*, \frac{1}{\rho} \cdot w^*)$ into the objective function of **Dual of P.B**, we can infer that the value of **Dual of P.B** is at most

$$\mu\gamma \sum_{t,t' \in [m]} (z^*_{t,t'}/\rho) + w^*/\rho \le \gamma \sum_{t,t' \in [m]} (z^*_{t,t'}/\rho) + w^*/\rho \le L^*/\rho \tag{26}$$

where the first inequality is by the observations that $\gamma \ge 0$, $z^*_{t,t'}/\rho \ge 0$, $\mu \in [0, 1]$, and the second inequality is by (22). By strong duality, the value of **P.B** is at most L^*/ρ. Hence, by finding a solution to **P.B** with a value of L^*, we can achieve a ρ-approximation for **P.B**.

Suppose $L^* - \epsilon$ is the largest value of L for which the algorithm identifies that $C(L)$ is empty. Define $\mathcal{F}'$ as the set that contains all the feasible subsets for which the dual constraint is violated during the implementation of the ellipsoid algorithm on $C(L^* - \epsilon)$. We use $\mathcal{F}'$ to construct a polynomial sized dual linear program of **P.B** (labeled as **Poly-sized Dual of P.B**).

Poly-sized Dual of P.B $\min_{z \in \mathbb{R}^{m \times m}, w \in \mathbb{R}_{\ge 0}} \gamma \sum_{t,t' \in [m]} z_{t,t'} + w$
subject to: $w \ge f(S) + \sum_{t,t' \in [m]} (g_{t'}(S) - g_t(S))z_{t,t'}, \forall S \in \mathcal{F}'$.

Because $C(L^* - \epsilon)$ is empty, the value of **Poly-sized Dual of P.B** at least $L^* - \epsilon$. Hence, the optimal solution to the dual of **Poly-sized Dual of P.B** (labeled as **Poly-sized P.B**) is at least $L^* - \epsilon$.

Poly-sized P.B $\max_{x \in [0,1]^{\mathcal{F}'}} \sum_{S \in \mathcal{F}} x_S f(S)$
subject to:

$$\begin{cases} \sum_{S \in \mathcal{F}'} x_S \cdot g_t(S) - \sum_{S \in \mathcal{F}'} x_S \cdot g_{t'}(S) \le \gamma, \forall t, t' \in [m]. \\ \sum_{S \in \mathcal{F}'} x_S \le 1. \end{cases}$$

Solving **Poly-sized P.B** obtains a solution with a value of $L^* - \epsilon$. This solution is a feasible ρ-approximation (with additive error ϵ) for **P.B**.

References

Abdulkadiroğlu A (2005) College admissions with affirmative action. Int J Game Theory 33:535–549

Asadpour A, Niazadeh R, Saberi A, Shameli A (2022) Sequential submodular maximization and applications to ranking an assortment of products. Oper Res

Beutel A, Chen J, Doshi T, Qian H, Wei L, Wu Y, Heldt L, Zhao Z, Hong L, Chi EH et al. (2019): Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. pp 2212–2220

Biddle D (2017) Adverse impact and test validation: a practitioner's guide to valid and defensible employment testing. Routledge, London

Bubeck S et al (2015) Convex optimization: algorithms and complexity. Found Trends® Mach Learn 8(3–4):231–357

Buchbinder N, Feldman M (2019) Constrained submodular maximization via a nonsymmetric technique. Math Oper Res 44(3):988–1005

Buchbinder N, Feldman M, Naor J, Schwartz R (2014) Submodular maximization with cardinality constraints. In: Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms. SIAM, pp 1433–1452

Calinescu G, Chekuri C, Pál M, Vondrák J (2007) Maximizing a submodular set function subject to a matroid constraint. In: International conference on integer programming and combinatorial optimization. Springer, pp 182–196

Celis E, Keswani V, Straszak D, Deshpande A, Kathuria T, Vishnoi N (2018) Fair and diverse DPP-based data summarization. In: International conference on machine learning. PMLR, pp 716–725

Celis LE, Huang L, Vishnoi NK (2018) Multiwinner voting with fairness constraints. In: Proceedings of the 27th international joint conference on artificial intelligence. pp 144–151

Celis LE, Straszak D, Vishnoi NK (2017) Ranking with fairness constraints. arXiv:1704.06840

Chen Q, Golrezaei N, Susan F, Baskoro E (2022) Fair assortment planning. arXiv:2208.07341

Chierichetti F, Kumar R, Lattanzi S, Vassilvitskii S (2017) Fair clustering through fairlets. Adv Neural Inf Process Syst 30

Chierichetti F, Kumar R, Lattanzi S, Vassilvtiskii S (2019) Matroids, matchings, and fairness. In: The 22nd international conference on artificial intelligence and statistics. PMLR, pp 2212–2220

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference

El Halabi M, Mitrović S, Norouzi-Fard A, Tardos J, Tarnawski JM (2020) Fairness in streaming submodular maximization: algorithms and hardness. Adv Neural Inf Process Syst 33:13609–13622

Grötschel M, Lovász L, Schrijver A (1981) The ellipsoid method and its consequences in combinatorial optimization. Combinatorica 1(2):169–197

Joseph M, Kearns M, Morgenstern JH, Roth A (2016) Fairness in learning: classic and contextual bandits. Adv Neural Inf Process Syst 29

Kallus N, Zhou A (2019) The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. Adv Neural Inf Process Syst 32

Kempe D, Mahdian M (2008) A cascade model for externalities in sponsored search. In: International workshop on internet and network economics. Springer, pp 585–596

Mehrotra A, Celis LE (2021) Mitigating bias in set selection with noisy protected attributes. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp 237–248

Monroe BL (1995) Fully proportional representation. Am Polit Sci Rev 89(4):925–940

Narasimhan H, Cotter A, Gupta M, Wang S (2020) Pairwise fairness for ranking and regression. In: Proceedings of the AAAI conference on artificial intelligence, vol 34. pp 5248–5255

Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functions—I. Math Program 14(1):265–294

Ohsaka N, Matsuoka T (2021) Approximation algorithm for submodular maximization under submodular cover. In: Uncertainty in artificial intelligence. PMLR, pp 792–801

Qi B (2022) On maximizing sums of non-monotone submodular and linear functions. In: 33rd International symposium on algorithms and computation (ISAAC 2022). Schloss Dagstuhl-Leibniz-Zentrum für Informatik

Sviridenko M, Vondrák J, Ward J (2017) Optimal approximation for submodular and supermodular optimization with bounded curvature. Math Oper Res 42(4):1197–1218

Talluri K, Van Ryzin G (2004) Revenue management under a general discrete choice model of consumer behavior. Manage Sci 50(1):15–33

Tang S, Yuan J, Twumasi MB (2023) Achieving long-term fairness in submodular maximization through randomization. In: 19th Cologne-Twente workshop on graphs and combinatorial optimization

Tsang A, Wilder B, Rice E, Tambe M, Zick Y (2019) (2019) Group-fairness in influence maximization. arXiv:1903.00967

Wang Y, Fabbri F, Mathioudakis M (2021) Fair and representative subset selection from data streams. In: Proceedings of the web conference 2021. pp 1340–1350

Yuan J, Tang S (2023) Group fairness in non-monotone submodular maximization. J Comb Optim 45(3):88

Yue Y, Guestrin C (2011) Linear submodular bandits and their application to diversified retrieval. Adv Neural Inf Process Syst 24

Zafar MB, Valera I, Rogriguez MG, Gummadi KP (2017) Fairness constraints: mechanisms for fair classification. In: Artificial intelligence and statistics. PMLR, pp 962–970

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.