arXiv:2304.06596v1 [cs.LG] 13 Apr 2023

超越次模性：具有群体公平性约束的随机集合选择统一框架

Shaojie Tang[1][0000−0001−9261−5210] 和 Jing Yuan[2][0000−0001−6407−834X]

[1] 纳文·金达尔管理学院，达拉斯德克萨斯大学 shaojie.tang@utdallas.edu[2] 北德克萨斯大学计算机科学与工程系 jing.yuan@unt.edu

摘要。机器学习算法在多种重要的决策过程中发挥着重要作用，包括定向广告展示、住房贷款审批和犯罪行为预测。鉴于这些算法影响深远，确保其公平运行、不对人口中的特定群体抱有偏见或歧视至关重要。确保这些算法公正无私，对于促进平等、避免歧视至关重要。为此，我们提出了一个纳入群体公平性约束的随机子集选择统一框架。我们的问题包含一个全局效用函数，以及每个群体各自的群体效用函数；此处的群体是指具有相同属性（例如性别）的一组个体（例如人）。我们的目标是针对可行子集构建一个分布，指定每个可行集合的选择概率，从而在满足每个群体效用函数预先设定的期望配额的同时，使全局效用函数最大化。请注意，全局效用函数与各群体效用函数之间未必存在直接联系。我们证明，该框架统一并推广了机器学习和运筹学中的许多重要应用。我们的算法结果要么改进了已知的最佳结果，要么为新的应用提供了首批近似算法。

1 引言

机器学习算法在决策中的应用日益广泛，这引发了人们对偏见和歧视可能性的担忧。然而，人们正采取各种措施，开发能够确保个人或群体获得公平结果的公平算法，即使是在敏感领域也是如此。这些措施包括为分类 [34], 排序 [11], 聚类 [13], 老虎机学习 [18],投票 [10], 大学招生 [1], 匹配 [14], 影响力最大化 [30], 和多样化数据摘要 [16]创建技术。

文献中提出了各种公平性概念，包括个体公平性、群体公平性和子群公平性。这些概念旨在解决算法决策中的歧视和偏见问题，尤其是在

就业、住房和刑事司法等敏感领域。然而，公平性没有普适的衡量标准，因为它往往取决于具体情境，也会受到多种因素的影响，例如决策过程的目标以及所涉及的敏感属性。本文提出一种通用的群体公平性记法，统一以往研究中的多种记法。我们假设存在 m 个群体，这些群体由一组共同属性定义，例如种族、性别或年龄。为评估特定解集 S 对特定群体 $t \in [m]$ 是否合适，我们引入 m 群体效用函数 $g_1, g_2, \cdots, g_m$：$2^V \to \mathbb{R}\geq 0$。每个群体效用函数 $g_t(S)$ 都会评估群体 t 从解集 S 中获得的效用。例如，在委员会选拔中，$[10]$, $g_t(S)$ 表示解集 S 中选自群体 t 的候选人人数。给定一组可行子集 $\mathcal{F}$，令 $x \in [0,1]^{\mathcal{F}}$ 表示解集 $\mathcal{F}$ 上的一个分布，其中 x_S 是选择 $S \in \mathcal{F}$ 的概率。若分布 x 满足 $\sum S \in \mathcal{F} x_S g_t(S) \geq \alpha_t, \forall t \in [m]$，则称其是公平的；这意味着每个群体 $t \in [m]$ 的期望效用都有下界 α_t。这种方法能够考量最终结果中各群体的代表性，促进多样性，并防止任何特定群体的代表性不足。

此外，还存在一个全局效用函数 f：$2^V \to \mathbb{R}_{\geq 0}$。分布 x 的期望（全局）效用可计算为 $\sum_{S\in\mathcal{F}} x_S f(S)$。我们的目标是找到一个分布 x，使 $\sum_{S\in\mathcal{F}} x_S f(S)$ 最大化，同时确保 $\sum_{S\in\mathcal{F}} x_S g_t(S) \geq \alpha_t, \forall t \in [m]$。该问题的形式化定义列于 P.0。我们将展示，这一表述整合并拓展了机器学习和运筹学领域中的众多重要应用。接下来，我们总结本文的主要贡献。

– 我们基于椭球法，为 P.0 开发了一个多项式时间算法框架。– 一个主要的算法性发现（如定理1中所正式陈述）是：假设对于所有 $z \in \mathbb{R}^m_{\geq 0}$，都存在一个多项式时间算法，能够返回一个集合 $A \in \mathcal{F}$，使得

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

对于某些 ρ、$\mu \in [0,1]$。随后，可以在多项式时间内为 P.0 设计一个（ρ、μ）近似算法。其中，ρ 表示近似比，而μ 表示该解对公平性约束的违反程度至多为 μ 倍。因此，求解 P.0 可归结为寻找一个组合子集选择问题的近似算法。我们利用这一结果处理各种应用，例如公平性感知子模最大化、带群体约束的序贯子模最大化，以及带市场份额约束的商品组合规划。对于其中一些应用，我们的方法优于现有方法；对于另一些应用，我们提出了新颖的应用场景，并开发了首个近似算法。

– 值得注意的是，当 f 为非负单调次模函数，且 g_t 为模函数时，我们的方法能够给出一个可行且最优的（$1-1/e$，1）-近似解。

– 我们探讨了对其他常用公平性指标的扩展，并提出了用于求解这些问题的有效算法。在第一个扩展中，我们对每个群体 $t \in [m]$的期望效用 β_t 引入了额外的上界。具体而言，我们要求 $\alpha_t \leq \sum_{S\in\mathcal{F}} x_S \cdot g_t(S) \leq \beta_t, \forall t \in [m]$。在第二个扩展中，我们探讨了另一种常用的公平性度量，旨在实现群体之间成对效用的均等。公平性程度由参数 γ定义。具体而言，对于任意两个群体 t、$t' \in [m]$，我们要求它们的期望效用之差不超过 γ，即 $\sum_{S\in\mathcal{F}} x_S \cdot g_t(S) - \sum_{S\in\mathcal{F}} x_S \cdot g_{t'}(S) \leq \gamma, \forall t, t' \in [m]$。

其他相关工作。多年来，各领域为开发公平算法投入了大量努力，以解决决策偏见问题。在影响力最大化和分类领域，研究人员一直在积极开发公平算法 [30,34]。同样，在投票系统中，研究重点是确保选举结果公平地代表选民的偏好，并由此推动了公平算法的发展 [10]。老虎机学习涉及基于不确定信息进行序贯决策，该领域对开发公平算法以解决偏见问题的兴趣也日益浓厚 [18]。此外，数据摘要领域也越来越重视公平算法的开发 [9] ，以均衡地呈现数据并避免有偏见的决策。

以往研究中公平性度量指标的选择取决于具体情境及所要解决的偏见类型，因此产生了一系列针对各类应用特定需求定制的优化问题和公平算法。我们采用一般性的群体效用函数，从各群体的角度评估解的效用。我们的框架具有足够的通用性，可以涵盖许多现有的公平性定义，包括 80% 规则 [4], 、统计均等 [15], 和比例代表性 [22]。尽管以往大多数关于公平性感知算法设计的研究（例如 [9,32,31,21]）旨在寻找一个确定性解集，我们的目标是计算一个随机化解，使其在平均意义上满足群体公平性约束。这种方法在实现群体公平性方面提供了更大的灵活性。我们的框架具有足够的通用性，可以涵盖各种通过随机化实现群体公平性的现有研究，例如[2,12,29]中探讨的研究。

2 预备知识与问题陈述

我们考虑一个包含 n 个物品 V 和 m 个群体的集合。存在一个全局效用函数f: $2^V \to \mathbb{R}_{\geq 0}$，以及 m 个群体效用函数 g_1、g_2、$\cdots$、g_m: $2^V \to \mathbb{R}_{\geq 0}$。给定物品子集 $S \subseteq V$，我们使用 $g_t(S)$ 从各群体 t 的角度评估 S 的效用。每个群体的效用期望值都有一个最低要求，用 $\alpha \in \mathbb{R}^m_{\geq 0}$ 表示，它构成了公平性约束。这一建模使我们能够设计兼顾各群体偏好的公平算法，确保决策过程不带偏见，并能带来

所有群体的公平结果。请注意， f 与 g_t 之间可能不存在任何联系。正如第4.1节所述，定义 g_t 的一种常见方法是统计从群体 t 中选出的物品数量。

假设 $\mathcal{F}$ 包含所有可行子集。例如，如果存在一个约束，将所选项目限制为 k，那么 $\mathcal{F}$ 可以定义为 $\{S \subseteq V \ |S| \leq k\}$。我们的问题目标记为 P.0，即找到一个在 $\mathcal{F}$ 上的分布 $x \in [0,1]^{\mathcal{F}}$，使全局效用的期望最大化，同时确保每个群体的最低期望效用均得到满足，以符合公平性约束。这里，决策变量 $x_S \in [0,1]$ 指定每个可行子集 $S \in \mathcal{F}$ 的选择概率。P.0 的形式化定义如下。

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

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

该线性规划包含 $m+1$ 个约束，不包括对所有 $S \in \mathcal{F}$的标准约束 $x_S \geq 0$。然而，该线性规划问题中的变量数量等于 $\mathcal{F}$的大小，这可能会随 n 呈指数增长。因此，传统的线性规划求解器无法高效地求解该线性规划问题。

3 P.0 的近似算法

在介绍我们的算法之前，我们先引入一个名为 FairMax 的组合优化问题。该问题的一个解将作为我们最终算法的子程序。

定义1（FairMax）。给定函数 f 和 g_1、g_2、$\cdots$、g_m，一个向量 $z \in \mathbb{R}^m_{\geq 0}$ 以及可行子集集合 $\mathcal{F}$、FairMax（z、$\mathcal{F}$），其目标是

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

即，FairMax（z，$\mathcal{F}$）旨在找到可行子集 $S \in \mathcal{F}$，使 f（S）$+ \sum_{t \in [m]} g_t$（S）$\cdot z_t$ 最大化。

接下来，我们给出本文的主定理。若解 $y \in [0,1]^{\mathcal{F}}$ 满足以下条件：$\sum_{S \in \mathcal{F}} y_S \leq 1$, $\sum_{S \in \mathcal{F}} y_S f(S) \geq a \times OPT$，其中 OPT表示 P.0 的最优解，且 $\sum_{S \in \mathcal{F}} y_S g_t(S) \geq b \times \alpha_t, \forall t \in [m]$，则称该解针对 P.0 达到 (a,b)-近似。这里，a 表示近似比，而 b 表示该解对公平性约束的违反程度最多为 μ 倍。以下定理建立了 FairMax$(z, \mathcal{F})$ 与 P.0 之间的联系。

定理1. 假设对于所有 $z \in \mathbb{R}^m_{\geq 0}$，都存在一个多项式时间算法，返回一个集合 $A \in \mathcal{F}$，使得

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

对于某些 ρ、$\mu \in [0,1]$。于是，存在一个针对 P.0 的多项式时间 (ρ, μ)-近似算法。

为了证明该定理，只需给出一个针对 P.0 的多项式 (ρ, μ)-近似算法，并将针对 FAIR-MAX$(z, \mathcal{F})$ 的多项式时间近似算法作为子程序。为此，我们将研究 P.0 的一种松弛形式，并将其称为 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 是通过将 P.0 中的公平性约束 α_t 放宽 $\mu \in [0,1]$ 倍得到的，其中 μ 在定理1中定义。通过求解 RP.0，我们可以得到一个对 P.0 近似可行的解。基于定理1中的假设，下面我们将重点寻找 RP.0 的一个解，并证明该解构成原问题 P.0 的一个双准则 (ρ, μ)-近似解。

注意，RP.0 中的变量数量等于 $\mathcal{F}$ 中元素的数量；当 n 较大时，这个数量可能会变得非常庞大。因此，标准线性规划求解器无法高效地求解这一线性规划问题。为应对这一挑战，我们转而考虑 RP.0 的对偶问题，并使用椭球算法 [17] 求解。

RP.0 的对偶问题（标记为 RP.0 的对偶问题）为每个群体 t 分配一个权重 $z_t \in \mathbb{R}_{\geq 0}$，并引入一个额外变量 $w \in \mathbb{R}_{\geq 0}$。

RP.0 的对偶问题 $\min_{z\in\mathbb{R}^m_{\geq 0}\ w\in\mathbb{R}_{\geq 0}} \sum_{t\in[m]} -\mu\alpha_t z_t + w$
满足约束：$w \geq f(S) + \sum_{t\in[m]} g_t(S) \cdot z_t, \forall S \in \mathcal{F}.$

注意，RP.0 的对偶问题中的约束数量可能关于 n 呈指数增长。从总体上看，我们旨在通过将约束数量减少到多项式级别来简化该问题，同时不显著改变最优解。

现在，我们将正式介绍用于 RP.0 对偶问题的算法，该算法包含一系列迭代，称为椭球算法。在每次迭代中，椭球算法都会通过近似求解一个 FAIRMAX 实例，来判断当前解是否可行。这个问题充当

可行性测试，并作为分离预言机，用于判断当前解位于所求问题可行域的内部还是外部。令 $C(L)$ 表示满足以下条件的 $(z \in \mathbb{R}^m_{\geq 0}, w \in \mathbb{R}_{\geq 0})$ 集合：

$$\sum_{t\in[m]} -\mu\alpha_t z_t + w \leq L,$$

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

容易验证，当且仅当 $C(L)$ 非空时，L 关于 RP.0 的对偶问题是可实现的。为了找到使 $C(L)$ 非空的 L 的最小值，我们采用二分搜索算法。

给定 L z w $\sum_{t\in[m]} -\mu\alpha_t z_t +$ 和 $w \leq L$ $\mathcal{A}$ $(,)$，我们首先评估不等式。如果不等式成立，算法将运行一个子程序来求解 FairMax$(z,\mathcal{F})$。具体而言，$\mathcal{A}$ 旨在找到可行子集 $S \in \mathcal{F}$，使 $f(S) + \sum_{t\in[m]} g_t(S)\cdot z_t$ 最大化。令 A 表示 $\mathcal{A}$ 返回的集合。

- 如果条件 $f(A) + \sum_{t\in[m]} g_t(A)\cdot z_t \leq w$ 成立，我们将 $C(L)$ 标记为非空集（请注意，即使在这种情况下，$C(L)$ 仍可能为空，因为 A 只是 FairMax$(z,\mathcal{F})$ 的近似解。不过，正如我们稍后将要说明的，这不会对最终解产生显著影响）。在这种情况下，我们得出结论：L 是可实现的，并继续尝试更小的 L 值。
- 如果 $f(A) + \sum_{t\in[m]} g_t(A)\cdot z_t > w$，这意味着 $(z,w) \notin C(L)$，因此，A是一个分离超平面。为了继续优化过程，我们搜索一个中心满足此约束的更小椭球。我们重复这一过程，直到在 $C(L)$ 中找到可行解，此时我们尝试一个更小的 L；或者直到边界椭球的体积变得非常小，以至于相对于 $C(L)$ 可视为空集。在后一种情况下，我们得出结论：当前目标无法实现，因此将尝试一个更大的 L。

如需详细了解使用分离预言机运行椭球法并获得（乘法和加法）近似保证所需的各个步骤，我们建议参阅 [5] 的第2章。获得上述椭球法的结果后，后续流程将包括两个主要步骤。首先，计算 P.0 的最优解上界。随后，计算 P.0 的 (ρ,μ)-近似解。

为 P.0 的最优解建立上界。定义 L^* 为使 $C(L)$ 被我们的算法标记为非空的最小 L 值。接下来，我们证明 P.0 的最优解至多为 L^*/ρ。为避免平凡情况，假设 $\rho > 0$。

由于 $C(L^*)$ 被标记为非空，因此存在一个 (z^*, w^*)，使得

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

and

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

根据定理1中关于 A 的假设，我们有

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

由于 $\rho > 0$，因此

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

其中，第一个不等式由 (4) 得出，第二个不等式则由不等式 (3) 得出。

考虑 P.0 的对偶问题（标记为 P.0的对偶），不等式 (5) 意味着 $(\frac{\mu}{\rho}\cdot z^*, \frac{1}{\rho}\cdot w^*)$ 是 P.0的对偶的一个可行解。

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: $w \geq f(S) + \sum_{t\in[m]} g_t(S)\cdot z_t, \forall S\in\mathcal{F}.$

将 $(\frac{\mu}{\rho}\cdot z^*, \frac{1}{\rho}\cdot w^*)$ 代入 P.0的对偶的目标函数，可推知 P.0的对偶的值至多为 $\sum_{t\in[m]} -\mu\alpha_t z_t^*/\rho + w^*/\rho \leq L^*/\rho$，其中该不等式由 (2) 得出。根据强对偶，P.0 的值至多为L^*/ρ。因此，通过求得 RP.0 的一个值为 L^*的解，我们可以为原问题 P.0 实现 (ρ, μ)-近似。

求值为 $L^* - \epsilon$ 的 RP.0 解。假设 $L^* - \epsilon$ 是算法判定 $C(L)$ 为空的 L 的最大值。此处，ϵ 表示二分搜索的精度。接下来，我们重点寻找一个值为 $L^* - \epsilon$ 的 RP.0 解。为此，我们只能利用 $\mathcal{F}$ 中与分离预言机获得的分离超平面相对应的可行子集。为实现这一点，我们定义 $\mathcal{F}$ 的一个子集，记为 $\mathcal{F}'$，其中包含在 $C(L^* - \epsilon)$ 上执行椭球算法期间对偶约束被违反的所有可行子集。由于仅有多项式数量的可行子集会违反对偶约束，因此 $\mathcal{F}'$ 的大小是多项式级的。我们可以利用 $\mathcal{F}'$ 中的可行子集，构造 RP.0 的多项式规模对偶线性规划（标记为“多项式规模的P.0对偶”）。

RP.0 的多项式化对偶问题 $\min_{z\in\mathbb{R}^m_{\geq 0}, w\in\mathbb{R}_{\geq 0}} \sum_{t\in[m]} \mu\alpha_t z_t + w$
满足： $w \geq f(S) + \sum_{t\in[m]} g_t(S)\cdot z_t, \forall S\in\mathcal{F}'$.

Poly-sized Dual of RP.0 的目标是在满足 $\mathcal{F}'$ 中由可行子集定义的约束条件下，最大化对偶目标函数。由于 $C(L^*-\epsilon)$ 为空，Poly-sized Dual of RP.0 的值至少为 $L^*-\epsilon$。因此，Poly-sized Dual of RP.0 的对偶问题（记为 Poly-sized RP.0）的最优解至少为 $L^*-\epsilon$。

多项式规模 RP.0
$\max_{x\in[0,1]^{\mathcal{F}'}} \sum_{S\in\mathcal{F}'} x_S f(S)$
满足以下条件:
$$\begin{cases}\sum_{S\in\mathcal{F}'} x_S\cdot g_t(S) \geq \mu\alpha_t, \forall t\in[m].\\ \sum_{S\in\mathcal{F}'} x_S \leq 1.\end{cases}$$

需要注意的是，Poly-sized RP.0 的规模是多项式的，因为$\mathcal{F}'$ 仅包含多项式数量的可行子集。因此，我们可以高效地求解 Poly-sized RP.0，并获得一个值为 $L^*-\epsilon$ 的解。该解是 P.0 的一个 (ρ,μ)-近似（加性误差为 ϵ）。

4 应用

本节介绍了我们的框架的一系列应用，其中一些应用的结果优于以往已知方法。在其他情况下，我们提出了新的应用，并为其提供了首个近似算法。

4.1 带群体公平性约束的次模最大化

在这个问题中，我们假设全局效用函数 f 和 m群体效用函数 $g_1, g_2, \cdots, g_m$ 是次模的[3]。这一问题设定足够通用，能够涵盖许多可通过次模全局效用函数建模的优化问题。此类应用包括数据摘要 [16]、影响力最大化 [20]、以及信息检索 [33]。值得注意的是，这种情形将次模覆盖下的次模最大化 [25] 作为一个特例纳入其中。其目标是在单个次模群体效用函数的最低配额约束下，最大化单调次模函数。不过，他们关注的是寻找一个确定性解集，而我们的方法在实现群体公平性方面更加灵活。接下来，我们将讨论假设

[3] 如果对于任意集合 X 和 Y（它们是 V 的子集，且满足 $X\subseteq Y$）以及任意元素 $e\in V\setminus Y$，以下不等式成立：$h(X\cup\{e\}) - h(X) \geq h(Y\cup\{e\}) - h(Y)$，则称函数 $h\colon 2^V\to\mathbb{R}$ 为次模函数。如果对于任意集合 $X\subseteq V$ 和任意元素 $e\in V\setminus X$，满足 $h(X\cup\{e\}) - h(X)\geq 0$，则称其为单调函数。

群体效用函数为次模函数的依据。在许多先前的研究中 [16,10], , 相对于敏感属性（例如种族或性别）的均衡，一直是评估公平性优化算法所得解的一项常用标准。这种均衡概念通常要求，解在敏感属性方面不会显著损害任何特定群体的利益，同时也能在全局目标函数 f上取得良好表现。下面给出一个具体示例。考虑一个由 n 个对象（例如人）构成的集合V，其中每个对象都关联一个敏感属性。令 $V_1, \ldots, V_m$ 表示属性相同的对象群体 m。我们将解 $x \in [0,1]^{\mathcal{F}}$定义为公平的，其中该解编码了从 $\mathcal{F}$中选择每个集合的选择概率；当从每个群体中 V_t 选出的对象的期望数量至少为 α_t时，即满足公平性，其中 α_t 通常设为 V_t 在整个集合 V中所占比例的倍数。在这种情况下，我们对每个 $t \in [m]$定义 $g_t(S) = |S \cap V_t|$。不难验证， g_t 是一个单调次模函数。

回顾一下，求解 P.0 可归约为求解 FairMax$(z, \mathcal{F})$（定义1）。这里，FairMax$(z, \mathcal{F})$ 的目标是 $\max_{S \in \mathcal{F}}(f(S) + \sum_{t \in [m]} g_t(S) \cdot z_t)$。幸运的是，如果 f 和 g_t 是次模函数，那么根据次模函数的线性组合仍然是次模函数这一事实，$f(S) + \sum_{t \in [m]} g_t(S) \cdot z_t$ 也是次模函数。如果我们假设族 $\mathcal{F}$ 是基于基数约束定义的，具体而言为 $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$，其中 k 为正整数，那么 FairMax$(z, \mathcal{F})$ 问题可归约为在满足基数约束的同时最大化一个次模函数。根据[24], , 如果目标函数是非负、单调且次模的，则该问题存在一个最优的 $(1 - 1/e)$-近似算法，即$\rho = \mu = 1 - 1/e$。另一方面，如果目标函数是非单调的（此外还满足非负和次模），则可以实现 0.385 的近似 [6], ，即 $\rho = \mu = 0.385$。这与定理1一起蕴含了以下命题。

命题 1.如果 f 和 g_t 是非负单调次模函数，并且 $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ 对于正整数 k成立，那么对于 P.0，存在一个最优的 $(1 - 1/e, 1 - 1/e)$-近似算法。如果 f 和 g_t 是非负非单调子模函数，则对于 P.0，存在一个 (0.385, 0.385)-近似算法。

单调次模 f 和模函数 g_t 的改进结果接下来，我们研究这一应用的一个重要特例，并假设全局效用函数 f 是非负单调且次模的； m 群体效用函数 $g_1, g_2, \cdots, g_m$ 是模函数。很容易验证，前文讨论的示例（其中 $g_t(S) = |S \cap V_t|$）满足模群组效用函数的性质。我们证明，对于这一特例，存在一种可行的最优 $(1 - 1/e)$-近似算法。

注意，如果 f 是非负单调且次模的， g_t 是模函数（因此$\sum_{t \in [m]} g_t(\cdot) \cdot z_t$ 也是模函数），并且 $\mathcal{F} = \{S \subseteq V \, |S| \leq k\}$ 对于正整数 k，那么 [27] 提出了一种随机多项式时间算法，该算法生成一个集合 $A \in \mathcal{F}$，使得对于每个 $S \in \mathcal{F}$，都有

that

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

即， $\rho = 1-1/e$ 和 $\mu = 1$。将这些值代入定理1，可得以下命题。注意，$\mu = 1$ 表明我们的解严格满足所有公平性约束。

命题2. 若全局效用函数 f 是非负单调次模函数； m 群体效用函数 $g_1, g_2, \cdots, g_m$ 是模函数； $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ 对于正整数 k，则存在一个适用于 P.0 的 $(1-1/e, 1)$-近似算法。

请注意，由于为简洁起见省略了 [27], 中原始结果（不等式 (6)）所包含的类似误差，因此上述结果可能存在较小的加性误差。

4.2 序贯子模最大化

该问题最早由 [2] 研究，其目标是确定一组物品的最优排序，以最大化多个子模函数的线性组合。此类子模最大化问题源于这样一种场景：平台向用户展示一份产品列表。用户会查看列表中的前 l 件商品，其中 l 从给定分布中随机选取；用户是否购买列表中的商品取决于选择模型。因此，平台的目标是最大化购物者的参与度，该参与度定义为购买概率。形式化地，给定单调次模函数 $h_1, \cdots, h_n\colon 2^V \to \mathbb{R}_{\geq 0}$ 和 $h_1^t, \cdots, h_n^t\colon 2^V \to \mathbb{R}_{\geq 0}$，对每个群体 $t \in [m]$均有定义；以及非负系数 $\lambda_1, \cdots, \lambda_n$ 和 $\lambda_1^t, \cdots, \lambda_n^t$，对每个群体 $t \in [m]$均有定义。略去符号上的区别，令 S 为物品 V 上的一个排列，并令 $S_{[l]}$ 表示 S中前 l 件物品。将全局效用函数定义为 $f(S) = \sum_{l\in[n]} \lambda_l h_l(S_{[l]})$。在产品排序的语境下， λ_l 表示耐心程度为 l的用户所占比例，而 h_l 对应于耐心程度为 l的用户的总体购买概率函数。类似地，对于每个用户群体 $t \in [m]$，群体效用函数定义为 $g_t(S) = \sum_{l\in[n]} \lambda_l^t h_l^t(S_{[l]})$，其中 λ_l^t 应解释为群体 t中耐心程度为 l 的用户所占比例。尽管 f 和 g_t 定义在排列而非集合上，但很容易验证定理1仍然成立。具体而言，假设对于所有 $z \in \mathbb{R}^m_{\geq 0}$，都存在一个多项式时间算法，能够返回一个排列 $A \in \mathcal{F}$，使得

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

对于某些 ρ、 $\mu \in [0,1]$。这里， $\mathcal{F}$是V中所有可能排列的集合。因此，存在一个用于序贯次模最大化问题的多项式时间（ρ， μ）近似算法。

注意，在这种情况下， FairMax$(z, \mathcal{F})$ 的目标函数可以写为

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

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

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

根据单调次模函数的线性组合仍为单调次模函数这一性质，如果对于所有 $l\in[n]$ 和 $t\in[m]$，h_l 和 h_l^t 均为单调次模函数，那么对于所有 $l\in[n]$，$\lambda_l h_l(\cdot) + \sum_{t\in[m]} z_t\lambda_l^t h_l^t(\cdot)$ 也具有单调性和次模性。根据 [2]，中定理1的分析，如果对于所有 $l\in[n]$，$\lambda_l h_l(\cdot)+\sum_{t\in[m]} z_t\lambda_l^t h_l^t(\cdot)$ 都是单调次模函数，那么寻找能使方程（9）右侧最大化的排列 S，即可转化为带有（层状）拟阵约束的次模最大化问题。因此，存在一种针对 FairMax$(z, \mathcal{F})$ 的 $(1-1/e)$-近似算法 [8]，即 $\rho = \mu = 1-1/e$。利用定理1，我们得到了针对 P.0 的 $(1-1/e, 1-1/e)$-近似算法。值得注意的是，目前最先进的结果 [2] 为 P.0 提供了双准则 $((1-1/e)^2, (1-1/e)^2)$-近似。与其结果相比，我们提出的框架在近似比和可行性方面均有显著改进。

命题3. 存在一种用于序贯子模最大化的最优（$1-1/e, 1-1/e$）-近似算法。

4.3 带群体市场份额约束的随机商品组合规划

第三个应用涉及商品组合规划，这是运筹学领域广泛关注的问题。带群体市场份额约束的商品组合规划，旨在确定向顾客展示的最佳产品组合，同时确保满足每个群体的最低市场份额要求。现有研究主要关注如何找到满足单一群体最低市场份额要求的确定性解。我们将这一研究拓展至包含多个群体的随机化设置。形式上，该问题以一个包含 n 个产品的集合 V 作为输入，并将其划分为 m 个（可能不互斥的）群体，记为 V_1、V_2、$\cdots$、V_m。在著名的多项式 Logit（MNL）模型下，每个产品 $i\in V$ 都有一个偏好权重 ν_i，并令 ν_0 表示不购买的偏好。令 r_i 表示销售产品 $i\in V$ 所获得的收入。给定一个商品组合 $S\subseteq V$，任意产品 $i\in S$ 的购买概率为 $\frac{\nu_i}{\nu_0+\sum_{i\in S}\nu_i}$。因此，提供 S 所获得的期望收入为 $f(S) = \frac{\sum_{i\in S} r_i\nu_i}{\nu_0+\sum_{i\in S}\nu_i}$，而由此产生的群体市场份额为

$t \in [m]$ 为 $g_t(S) = \frac{\sum_{i \in S \cap V_t} \nu_i}{\nu_0 + \sum_{i \in S} \nu_i}$。假设 $\mathcal{F}$ 由 V 的所有可能子集构成，P.0 的目标是计算每种产品组合 $S \subseteq V$ 的选择概率 x_S，使期望收入 $\sum_{S \in \mathcal{F}} x_S f(S)$ 最大化，同时每个群体的期望市场份额至少达到 $\sum_{S \in \mathcal{F}} x_S g_t(S) \geq \alpha_t$。为解决这一问题，我们考虑其对应的 FairMax$(z, \mathcal{F})$，其目标是对于给定向量 $z \in \mathbb{R}^m_{\geq 0}$，求出一个使 $f(S) + \sum_{t \in [m]} g_t(S) \cdot z_t$ 最大化的 $S \subseteq V$。根据 f 和 g_t 的定义，FairMax$(z, \mathcal{F})$ 的目标函数可写为

$$f(S) + \sum_{t \in [m]} g_t(S) \cdot z_t = \frac{\sum_{i \in S} r_i \nu_i}{\nu_0 + \sum_{i \in S} \nu_i} + \sum_{t \in [m]} \frac{\sum_{i \in S \cap V_t} \nu_i}{\nu_0 + \sum_{i \in S} \nu_i} \cdot z_t \tag{10}$$

$$= \frac{\sum_{i \in S} (r_i + \sum_{t \in [m]} z_t \cdot \mathbf{1}_{i \in V_t}) \nu_i}{\nu_0 + \sum_{i \in S} \nu_i}, \tag{11}$$

其中，$\mathbf{1}_{i \in V_t} \in \{0, 1\}$ 是一个指示变量，当且仅当$i \in V_t$时，$\mathbf{1}_{i \in V_t} = 1$。

因此，FairMax$(z, \mathcal{F})$ 的目标是

$$\max_{S \subseteq V} \frac{\sum_{i \in S} (r_i + \sum_{t \in [m]} z_t \cdot \mathbf{1}_{i \in V_t}) \nu_i}{\nu_0 + \sum_{i \in S} \nu_i}. \tag{12}$$

该问题可以视为一个无约束的商品组合规划问题，其中每种商品 $i \in V$ 的收入为 $(r_i + \sum_{t \in [m]} z_t \cdot \mathbf{1}_{i \in V_t})$，偏好权重为 ν_i，而不购买的偏好权重为 ν_0。根据 [28], , 该问题的最优解是按收入排序的商品组合。换言之，使收入最大化的商品组合由收入最高的 l 种商品 $(r_i + \sum_{t \in [m]} z_t \cdot \mathbf{1}_{i \in V_t})$ 构成，其中 $l \in [n]$。因此，FairMax$(z, \mathcal{F})$（问题 (12)）至多通过考察 n 个候选商品组合，即可在多项式时间内求得最优解。因此，定理1与能够在多项式时间内最优求解 FairMax$(z, \mathcal{F})$ 的能力（即 $\rho = \mu = 1$）相结合，意味着 P.0 存在一个最优解。

命题 4. 存在一种最优且可行的算法，用于带有群体市场份额约束的商品组合规划。

5 关于其他公平性表示法的讨论

本节将考察文献中常用的另外两种公平性表示法。

5.1 纳入公平性上界约束 β_t

扩展我们在 P.0 中引入的公平性记法，一种自然的方式是对每个群体的期望效用施加额外的上界 β_t。形式上，

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

满足：

$$\begin{cases} \alpha_t \le \sum_{S\in\mathcal{F}} x_S \cdot g_t\ S\ \le \beta_t\ \forall t \in [m]. 0\,, \\ \sum_{S\in\mathcal{F}} x_S \le 1. \end{cases}$$

这种一般形式可见于以往关于公平性感知优化的若干研究，例如 [9,16]。幸运的是，我们仍可使用椭球法求解 P.A，从而得到一种双准则算法。正如我们稍后将看到的，以下 FairMax 问题可充当分离预言机，帮助判断当前解位于待求解问题的可行域之内还是之外。

定义2（FairMax）。给定函数 f 和 g_1、g_2、$\cdots$、g_m，两个向量 $z \in \mathbb{R}^m_{\ge 0}$ 和 $u \in \mathbb{R}^m_{\ge 0}$，以及一个可行子集集合 $\mathcal{F}$，FairMax（z，u，$\mathcal{F}$）旨在

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

与式（1）中的目标函数不同，上述效用函数中 g_t 的系数可能为负值。下一个定理建立了 FairMax$(z,u,\mathcal{F})$ 与 P.A. 之间的联系。在此，我们扩展 (a,b)-近似的定义：若解 $y \in \mathbb{R}^m_{\ge 0}$ 满足以下条件，则称其对 P.A. 实现了 (a,b)-近似：$\sum_{S\in\mathcal{F}} y_S \le 1$,$\sum_{S\in\mathcal{F}} y_S f(S) \ge a \times OPT$，其中 OPT 表示 P.A. 的最优解，且 $\beta_t \ge \sum_{S\in\mathcal{F}} y_S g_t(S) \ge b \times \alpha_t, \forall t \in [m]$。以下定理的证明见附录。

定理 2。假设对于所有的 $z \in \mathbb{R}^m_{\ge 0}$ 和 $u \in \mathbb{R}^m_{\ge 0}$，都存在一个多项式时间算法，返回一个集合 $A \in \mathcal{F}$，使得

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

对于某些 ρ、$\mu \in [0,1]$成立。则存在一个针对 P.A. 的多项式时间（ρ，μ）-近似算法。

注意，如果 f 是非负单调次模函数，g_t 是一个模函数（因此 $\sum_{t\in[m]} g_t(\cdot)\cdot(z_t - u_t)$ 也是一个模函数），并且对于一个正整数 k，$\mathcal{F} = \{S \subseteq V \mid |S| \le k\}$，那么 [27] 提出了一种随机化多项式时间算法，该算法生成一个集合 $A \in \mathcal{F}$，使得对于每个 $S \in \mathcal{F}$，都有

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

以下命题可由定理 2 和不等式 (14) 立即推出。请注意，该结果重现了来自 [29] 的发现。

命题 5. 若全局效用函数 f 是非负单调子模函数； m 群体效用函数 g_1、 g_2、 $\cdots$、 g_m 均为模函数；$\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ 对于正整数 k，则存在一种针对 P.A. 的可行 $(1 - 1/e, 1)$-近似算法。

对于 f 为非负、非单调且次模， g_t 为模函数，并且 $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ 对于正整数k的情况， [26] 提出了一种用于 FairMax 的算法，可实现 $\rho = \frac{te^{-t}}{t+e^{-t}} - \epsilon$ 和$\mu = \frac{t}{t+e^{-t}}$，对任意常数 $t \in [0, 1]$ 均成立。结合定理 2，这表明 P.A. 存在一个（$\frac{te^{-t}}{t+e^{-t}} - \epsilon$， $\frac{t}{t+e^{-t}}$）-近似算法。

5.2 成对公平

现在，我们将探讨另一种常用的公平性定义，该定义依赖于实现不同群体之间成对效用的均等。这种定义已应用于各种场景，例如推荐系统 [3],、商品组合规划 [12], 、排序和回归模型 [23], ，以及预测风险评分 [19]。根据这一概念，各群体的效用水平应当相近。公平程度由参数γ决定。具体而言，对于任意两个群体 t、 $t' \in [m]$，其期望效用之差至多为 γ，即 $\sum_{S\in\mathcal{F}} x_S \cdot g_t(S) - \sum_{S\in\mathcal{F}} x_S \cdot g_{t'}(S) \leq \gamma$、 $\forall t$、 $t' \in [m]$。P.B. 对该问题进行了形式化定义。通过为每一对群体引入不同的 γ，我们的算法研究结果可推广到这一问题的更一般版本。为简洁起见，此处不再详述。

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

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

与处理其他公平性表示法的方法相比，对于这一特殊情况，我们不会将原始问题转换为松弛形式，而是直接求解 P.B 的对偶问题。P.B 的对偶问题见“P.B 的对偶问题”一节。

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

要使用椭球法求解 P.B 的对偶问题，我们在 FairMax 中定义其分离预言机。

定义 3（FairMax）。给定函数 f和 $g_1, g_2, \cdots, g_m$、矩阵 $z \in \mathbb{R}^{m\times m}_{\geq 0}$ 以及可行子集集合 $\mathcal{F}$, FairMax$(z, \mathcal{F})$，其目标是

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

下一个定理建立了 FairMax$(z, \mathcal{F})$ 与 P.B. 之间的联系。与我们针对其他公平性表示法的结果不同，在那些情况下我们只能预期得到一个双准则解；我们证明，近似求解 FairMax$(z, \mathcal{F})$ 可得到 P.B. 的一个可行解。以下定理的证明移至附录。

定理 3. 假设对于所有 $z \in \mathbb{R}_{\geq 0}^{m \times m}$，都存在一个多项式时间算法，能够返回集合 $A \in \mathcal{F}$，使得

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

对于某些 ρ、$\mu \in [0, 1]$。那么，对于 P.B.，存在一个可行的 ρ-近似算法。

需要注意的是，上述性能界不依赖于μ。注意，如果 f 是非负单调次模函数，g_t 对于所有 $t \in [m]$ 都是模函数（因此 $\sum_{t,\ t' \in [m]} (g_{t'}(\cdot) - g_t(\cdot)) z_{t,\ t'}$ 对于所有 t、$t' \in [m]$ 也都是模函数），并且 $\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ 是一个正整数，那么 [27] 提出了一种随机化多项式时间算法，该算法生成集合 $A \in \mathcal{F}$，使得 $\forall S \in \mathcal{F}$，

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

以下命题可直接由定理 3 和不等式（16）推出。

命题 6. 如果全局效用函数 f 是一个非负单调子模函数；m 群体效用函数 g_1、g_2、$\cdots$、g_m 是模函数；$\mathcal{F} = \{S \subseteq V \mid |S| \leq k\}$ 对于正整数 k，则存在一个针对 P.B. 的可行（$1 - 1/e$）近似算法。

备注： [12] 最近提出了一种基于椭球的方法，用于解决带有成对公平性约束的商品组合规划问题。在 [12] 第 14 页（2022 年 10 月 28 日版本）中，他们讨论了所有商品收益均为统一收益的情况。他们指出，对于这一特殊情况，其分离预言机需要最大化一个非负单调次模函数与一个（不一定为正的）模函数之和。他们声称，该目标函数是非单调次模函数，并建议使用 [7] 提出的连续双贪心算法来获得一个$[1/e + 0.004, 1/2]$近似解。然而，需要注意的是， [7] 提出的算法仅适用于目标函数非负的情况。遗憾的是，一般而言，非负单调次模函数与一个（不一定为正的）模函数之和可能取负值，

因此 [7]的算法并不适用。另一方面，我们在定义 3 中的分离预言机也具有这种形式的目标函数，但将 [27]的算法用作子程序来求解，可得到原问题的 $(1-1/e)$ 近似解。容易验证，我们的框架（即使用 [27] 的算法求解分离预言机）也可应用于 [12] 研究的问题；通过对定理 3 进行类似分析，可以得到一个 $(1-1/e)$ 近似解。

6 结论

本文提出了一种通用的群体公平性记法，统一了以往研究中使用的多种记法。我们将问题表述为：在满足公平性约束的同时，寻找一个解集上的分布，使全局效用最大化。我们基于椭球法开发了一个多项式时间算法框架来解决这一问题。我们还针对问题的一个特殊情形开发了一种最优的（$1-1/e$）近似算法，其中 f 是单调次模函数，而 g_t 是模函数。该解严格满足所有公平性约束。我们的研究表明，这一表述汇集并拓展了机器学习和运筹学领域中的众多重要应用。

7 附录

7.1 定理 2 的证明

为了证明该定理，只需给出一个用于 P.A 的多项式 (ρ,μ)-近似算法，并将用于 FAIR-MAX$(z,u,\mathcal{F})$ 的多项式时间近似算法作为子程序。我们首先介绍 P.A 在 RP.A 中的一种松弛形式，其中下界约束被 $\mu\alpha_t$ 取代。

$$\begin{array}{|l|}\hline \textbf{RP.A } \max_{x\in[0,1]^{\mathcal{F}}} \sum_{S\in\mathcal{F}} x_S f(S) \\ \textbf{subject to:} \\ \quad \begin{cases} \mu\alpha_t \le \sum_{S\in\mathcal{F}} x_S \cdot g_t(S) \le \beta_t, \forall t\in[m]. \\ \sum_{S\in\mathcal{F}} x_S \le 1. \end{cases} \\ \hline \end{array}$$

RP.A 的对偶问题列于“RP.A 的对偶问题”一节中。

$$\begin{array}{|l|}\hline \textbf{Dual of RP.A } \min_{z\in\mathbb{R}^m_{\ge0},u\in\mathbb{R}^m_{\ge0},w\in\mathbb{R}_{\ge0}} \sum_{t\in[m]}(\beta_t u_t - \mu\alpha_t z_t) + w \\ \textbf{subject to: } w \ge f(S) + \sum_{t\in[m]} g_t(S)\cdot(z_t-u_t), \forall S\in\mathcal{F}. \\ \hline \end{array}$$

令 $C(L)$ 表示满足以下条件的 $(z\in\mathbb{R}^m_{\ge0}, u\in\mathbb{R}^m_{\ge0}, w\in\mathbb{R}_{\ge0})$ 集合：

$$\sum_{t\in[m]}(\beta_t u_t - \mu\alpha_t z_t) + w \le L,$$

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

容易验证，当且仅当 $C(L)$ 非空时， L 对 RP.A 的对偶问题是可实现的。为了找到使 $C(L)$ 非空的 L 最小值，我们使用二分搜索算法。

对于给定的 L 和 (z,u,w)，我们首先检验不等式 $\sum_{t\in[m]}(\beta_t u_t - \mu\alpha_t z_t) + w \leq L$。如果该不等式成立，算法将运行子程序 $\mathcal{A}$ 来求解 FairMax$(z,u,\mathcal{F})$。假设 $\mathcal{A}$ 是一个用于FairMax$(z,u,\mathcal{F})$的 μ-近似算法，令 A 表示由 $\mathcal{A}$返回的集合。

– 如果条件 $f(A) + \sum_{t\in[m]} g_t(A)\cdot(z_t - u_t) \leq w$ 成立，我们就将 $C(L)$标记为空集以外的集合。在这种情况下，我们会尝试一个更小的 L。– 如果 $f(A) + \sum_{t\in[m]} g_t(A)\cdot(z_t - u_t)\cdot z_t > w$，这意味着 $(z,u,w) \notin C(L)$，因此 A 是一个分离超平面。我们会寻找一个更小的椭球，其中心满足该约束。我们重复这一过程，直到在 $C(L)$中找到一个可行解，此时我们尝试一个更小的L；或者直到边界椭球的体积小到相对于 $C(L)$可视为空集。在后一种情况下，我们得出结论：当前目标无法实现，因此将尝试一个更大的 L。

定义 L^* 为使 $C(L)$ 被我们的算法标记为非空的最小 L 值。接下来，我们证明 P.A 的最优解至多为 L^*/ρ。为避免平凡情形，假设 $\rho > 0$。

由于 $C(L^*)$ 被标记为非空，因此存在一个 (z^*,u^*,w^*)，使得

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

and

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

根据定理 2 中关于 A 的假设，我们有 $\forall S \in \mathcal{F}$,

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

由此可得 $\forall S \in \mathcal{F}$,

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

其中，第一个不等式由 (19) 得出，第二个不等式则依据不等式 (18)。

考虑 P.A 的对偶问题（标记为 P.A 的对偶问题），不等式 (20) 表明，$(\frac{\mu}{\rho}\cdot z^*, \frac{\mu}{\rho}\cdot u^*, \frac{1}{\rho}\cdot w^*)$ 是 P.A 的对偶问题的可行解。

P.A 的对偶问题 $\min_{z\in\mathbb{R}^m_{\geq 0},\ u\in\mathbb{R}^m_{\geq 0},\ w\in\mathbb{R}_{\geq 0}} \sum_{t\in[m]}(\beta_t u_t - \alpha_t z_t) + w$
约束条件：$w \geq f(S) + \sum_{t\in[m]} g_t(S)\cdot(z_t - u_t)\ \forall S\in\mathcal{F}$.

将 $(\frac{\mu}{\rho}\cdot z^*, \frac{\mu}{\rho}\cdot u^*, \frac{1}{\rho}\cdot w^*)$ 代入 P.A 的对偶问题的目标函数中，我们可以推断出 P.A 的对偶问题的值至多为

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

其中，第一个不等式基于以下观察：$\beta_t u_t^*/\rho \geq 0$ 和 $\mu\in[0,1]$；第二个不等式则由（17）得出。根据强对偶，P.A 的值至多为 L^*/ρ。因此，通过找到一个值为 L^*的 RP.A 解，我们可以为原问题 P.A 实现 (ρ,μ)-近似。

假设 $L^*-\epsilon$ 是算法判定 $C(L)$ 为空的 L 的最大值。这里，ϵ 表示二分搜索的精度。接下来，我们将重点放在寻找一个值为 $L^*-\epsilon$的 RP.A 解上。定义 $\mathcal{F}'$ 为这样一个集合，其中包含在对 $C(L^*-\epsilon)$ 实施椭球算法期间违反对偶约束的所有可行子集。我们使用 $\mathcal{F}'$ 构造一个规模为多项式的 RP.A 对偶线性规划（标记为 P.A 的多项式规模对偶问题）。

RP.A 的多项式规模对偶问题 $\min_{z\in\mathbb{R}^m_{\geq 0},\ u\in\mathbb{R}^m_{\geq 0},\ w\in\mathbb{R}_{\geq 0}} \sum_{t\in[m]}(\beta_t u_t - \mu\alpha_t z_t) + w$
满足：$w \geq f(S) + \sum_{t\in[m]} g_t(S)\cdot(z_t - u_t)\ \forall S\in\mathcal{F}'$.

由于 $C(L^*-\epsilon)$ 为空，RP.A 的多项式规模对偶问题的值至少为 $L^*-\epsilon$。因此，RP.A 多项式规模对偶问题的对偶问题（记为多项式规模 RP.A）的最优解至少为 $L^*-\epsilon$。

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

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

求解多项式规模 RP.A 可得到一个值为 $L^*-\epsilon$ 的解。对于 P.A，该解是一个 (ρ,μ)-近似（加性误差为 ϵ）。

7.2 定理 3 的证明

为证明该定理，只需给出一个可行的 ρ-近似算法来求解 P.B，并以 FairMax $(z,\mathcal{F})$ 的多项式时间近似算法作为子程序。令 $C(L)$ 表示满足以下条件的 $(z\in\mathbb{R}^{m\times m}_{\geq 0}, w\in\mathbb{R}_{\geq 0})$ 集合：

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

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

容易验证，当且仅当 $C(L)$ 非空时， L 相对于 P.B 的对偶问题是可实现的。为了找到使 $C(L)$ 非空的 L 最小值，我们使用二分搜索算法。

对于给定的 L z w $\gamma \sum_{t\, t' \in [m]} z_{t\, t'} +$ 和 $w \leq L$ $\mathcal{A}$ (,)，我们首先评估不等式 ,,。如果该不等式成立，算法会运行一个子程序来求解 FairMax$(z, \mathcal{F})$。令 A 表示 $\mathcal{A}$ 返回的集合。

– 如果条件 $f(A) + \sum_{t,t' \in [m]} (g_{t'}(A) - g_t(A)) z_{t,t'} \leq w$ 成立，我们就将 $C(L)$ 标记为非空集合。在这种情况下，我们会尝试取一个更小的 L。– 如果 $f(A) + \sum_{t,t' \in [m]} (g_{t'}(A) - g_t(A)) z_{t,t'} > w$，这意味着 $(z, w) \notin C(L)$，因此 A 是一个分离超平面。我们寻找一个更小的椭球，其中心满足该约束。我们重复此过程，直到在 $C(L)$ 中找到可行解，此时我们会尝试取一个更小的 L；或者直到包围椭球的体积变得极小，以至于相对于 $C(L)$ 可视为为空。在后一种情况下，我们得出结论：当前目标无法实现，因此将尝试取一个更大的 L。

定义 L^* 为使得我们的算法将 $C(L)$ 标记为非空的最小 L 值。接下来，我们证明 P.B 的最优解至多为 L^*/ρ。为避免平凡情况，假设 $\rho > 0$。

由于 $C(L^*)$ 被标记为非空，因此存在一个 (z^*, w^*)，使得

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

and

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

根据定理 3 中关于 A 所作的假设，我们有 $\forall S \in \mathcal{F}$,

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

由此可得 $\forall S \in \mathcal{F}$,

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

其中，第一个不等式由（24）得出，第二个不等式则由不等式（23）得出。

不等式(25) 表明，($\frac{\mu}{\rho} \cdot z^*, \frac{1}{\rho} \cdot w^*$) 对 P.B 的对偶问题是可行的。

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

将 ($\frac{\mu}{\rho} \cdot z^*, \frac{1}{\rho} \cdot w^*$)代入 P.B的对偶问题的目标函数中，我们可以推断 P.B 的对偶问题的值至多为

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

其中，第一个不等式由以下观察得出：$\gamma \geq 0$、$z^*_{t,t'}/\rho \geq 0$、$\mu \in [0,1]$；第二个不等式由（22）得出。根据强对偶，P.B 的值至多为 L^*/ρ。因此，通过找到一个值为 L^*的 P.B 解，我们可以为 P.B 实现 ρ-近似。

假设 $L^* - \epsilon$ 是算法识别出 $C(L)$ 为空时 L 的最大值。定义 $\mathcal{F}'$ 为在对 $C(L^* - \epsilon)$ 执行椭球算法期间违反对偶约束的所有可行子集组成的集合。我们使用 $\mathcal{F}'$ 来构造 P.B 的一个多项式规模的对偶线性规划（标记为多项式规模的 P.B 对偶问题）。

多项式规模的 P.B 对偶问题 $\min_{z \in \mathbb{R}^{m \times m},\ w \in \mathbb{R}_{\geq 0}} \gamma \sum_{t,\ t' \in [m]} z_{t,\ t'} + w$ 满足：$w \geq f(S) + \sum_{t,\ t' \in [m]} (g_{t'}(S) - g_t(S)) z_{t,\ t'}, \forall S \in \mathcal{F}'.$

由于 $C(L^* - \epsilon)$ 为空，多项式规模的 P.B 对偶问题的值至少为$L^* - \epsilon$。因此，多项式规模的 P.B 对偶问题的对偶问题的最优解（标记为多项式规模的 P.B）至少为 $L^* - \epsilon$。

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

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

求解多项式规模的 P.B 可得到值为 $L^* - \epsilon$ 的解。该解是 P.B 的一个可行 ρ-近似（加性误差为 ϵ）。

参考文献

1.Abdulkadi˘glu,A.: 平权行动下的大学招生。国际博弈论杂志 33, 535–549 (2005)

2. Asadpour, A., Niazadeh, R., Saberi, A., Shameli, A.：顺序次模最大化及其在产品组合排序中的应用。《运筹学研究》（2022）3. Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E.H., 等：通过成对比较实现推荐排序中的公平性。载于：第25届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第2212–2220页（2019）4. Biddle, D.：不利影响与测验效度验证：有效且站得住脚的就业测验实践指南。Routledge（2017）5. Bubeck, S., 等：凸优化：算法与复杂性。《机器学习基础与趋势》8(3-4)，231–357（2015）6. Buchbinder, N., Feldman, M.：通过非对称技术实现约束次模最大化。《运筹数学》44(3)，988–1005（2019）7. Buchbinder, N., Feldman, M., Naor, J., Schwartz, R.：基数约束下的次模最大化。载于：第二十五届 ACM-SIAM 离散算法年度研讨会论文集，第1433–1452页。SIAM（2014）8. Calinescu, G., Chekuri, C., P´al, M., Vondr´ak, J.：在拟阵约束下最大化次模集合函数。载于：整数规划与组合优化国际会议，第182–196页。施普林格（2007）9. Celis, E., Keswani, V., Straszak, D., Deshpande, A., Kathuria, T., Vishnoi, N.：公平且多样的基于DPP的数据摘要。载于：国际机器学习会议，第716–725页。PMLR（2018）10. Celis, L.E., Huang, L., Vishnoi, N.K.：具有公平性约束的多赢家投票。载于：第27届国际人工智能联合会议论文集，第144–151页（2018）11. Celis, L.E., Straszak, D., Vishnoi, N.K.：具有公平性约束的排序。arXiv预印本 arXiv:1704.06840（2017）12. Chen, Q., Golrezaei, N., Susan, F., Baskoro, E.：公平的商品组合规划。arXiv预印本 arXiv:2208.07341（2022）13. Chierichetti, F., Kumar, R., Lattanzi, S., Vassilvitskii, S.：通过公平簇实现公平聚类。《神经信息处理系统进展》30（2017）14. Chierichetti, F., Kumar, R., Lattanzi, S., Vassilvtiskii, S.：拟阵、匹配与公平性。载于：第22届人工智能与统计国际会议，第2212–2220页。PMLR（2019）15. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.：通过认知实现公平性。载于：第3届理论计算机科学创新会议论文集（2012）16. El Halabi, M., Mitrovi´c, S., Norouzi-Fard, A., Tardos, J., Tarnawski, J.M.：流式次模最大化中的公平性：算法与困难性。《神经信息处理系统进展》33，13609–13622（2020）17. Gr¨otschel, M., Lov´asz, L., Schrijver, A.：椭球法及其在组合优化中的影响。Combinatorica 1(2)，169–197（1981）18. Joseph, M., Kearns, M., Morgenstern, J.H., Roth, A.：学习中的公平性：经典老虎机与上下文老虎机。《神经信息处理系统进展》29（2016）19. Kallus, N., Zhou, A.：分类之外的风险评分公平性：二部排序与XAUC指标。《神经信息处理系统进展》32（2019）20. Kempe, D., Mahdian, M.：赞助搜索中外部性的级联模型。载于：互联网与网络经济学国际研讨会，第585–596页。施普林格（2008）

21.Mehrotra, A., Celis,L.E.：在受噪声干扰的受保护属性下缓解集合选择中的偏见。载于：2021年ACM公平、问责与透明度会议论文集。第237–248页（2021） 22. Monroe,B.L.：完全比例代表制。《美国政治学评论》89(4)，925–940（1995） 23. Narasimhan,H.,Cotter,A.,Gupta, M., Wang,S.：排序和回归中的成对公平。载于：AAAI人工智能会议论文集。第34卷，第5248–5255页（2020） 24. Nemhauser,G.L.,Wolsey, L.A., Fisher,M.L.：最大化次模集合函数的近似分析-i。《数学规划》14(1)，265–294（1978） 25.Ohsaka, N.,Matsuoka,T.：子模覆盖下子模最大化的近似算法。载于：人工智能中的不确定性。第792–801页。PMLR（2021）26. Qi, B.：关于最大化非单调次模函数与线性函数之和。载于：第33届国际算法与计算研讨会（ISAAC 2022）。Schloss Dagstuhl-Leibniz-Zentrum f¨urInformatik（2022）27. Sviridenko, M.,Vondr´ak,J.,Ward,J.：具有有界曲率的次模与超模优化的最优近似。《运筹数学》42(4)，1197–1218（2017） 28. Talluri,K., Van Ryzin,G.：一般消费者行为离散选择模型下的收益管理。《管理科学》50(1)，15–33（2004）29. Tang, S.,Yuan, J., Twumasi,M.B.：通过随机化实现子模最大化中的长期公平。载于：第19届科隆-特温特图与组合优化研讨会（2023）30. Tsang,A., Wilder, B., Rice, E., Tambe, M., Zick,Y.：影响力最大化中的群体公平性。arXiv预印本arXiv:1903.00967（2019） 31. Wang,Y., Fabbri, F., Mathioudakis,M.：从数据流中进行公平且具有代表性的子集选择。载于：2021年万维网会议论文集。第1340–1350页（2021） 32. Yuan, J.,Tang, S.：非单调次模最大化中的群体公平性。《组合优化杂志》45(3)，88（2023） 33.Yue,Y., Guestrin,C.：线性子模老虎机及其在多样化检索中的应用。神经信息处理系统进展24（2011） 34. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.：公平性约束：公平分类的机制。载于：人工智能与统计。第962–970页。PMLR（2017）