# Submodular Maximization in the Presence of Biases

Anonymous Author(s)

## ABSTRACT

Subset selection tasks arise in recommendation systems and search engines. They ask to select a subset of items that maximize the value for the user. The value of subsets often display diminishing returns, and hence, submodular functions have been used to model them. If the inputs defining the submodular function are known, then existing algorithms can be used. In many applications, however, inputs have been observed to have social biases that reduce the utility of the output subset. Hence, interventions to improve the utility are desired. Prior works focus on maximizing linear functions—a special case of submodular functions—and show that fairness constraint-based interventions can not only ensure proportional representation but also achieve near-optimal utility in the presence of biases. We study the maximization of a family of submodular functions that capture functions arising in the aforementioned applications. Our first result is that, unlike with linear functions, constraint-based interventions cannot guarantee any constant fraction of the optimal utility for this family. Our second result is an algorithm for submodular maximization. The algorithm provably outputs subsets that have near-optimal utility for this family and that proportionally represents items from each group. In empirical evaluation, with both synthetic and real-world data, we observe that this algorithm improves the utility of the output subset for this family of submodular functions as well as its extensions.

## KEYWORDS

submodular maximization, recommendation systems, subset selection, algorithmic fairness

**ACM Reference Format:**

Anonymous Author(s). 2018. Submodular Maximization in the Presence of Biases. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 37 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Subset selection arises in many web-based applications including different types of content recommendation systems [10, 17, 49, 58] and search engines [2]. Generally speaking, in all of these applications, given a set of $n$ items (e.g., posts, products, videos, or websites), the task is to select a subset $S$ of size $k$ that is the most valuable for the user. Submodular functions are often used to capture the utility of subsets of items because of the diminishing returns property that arises in the above applications [2, 19, 36]. Diminishing returns arise because of the fact that an item's value to a user depends on the other shown to the user [2, 9, 17, 39, 58]. For instance, in a recommendation system, where items belong to different categories (such as genres or product types), each additional item from the same category gives a diminishing value to the user [17, 49, 58]. Similarly, in web search, results belong to different categories (based on, e.g., relevance to technology, news, locations etc.) and each additional result from the same category adds diminishing value [2]. In settings with multiple stakeholders (e.g., multiple users, content creators, and platform), adding item $i$ to a subset containing another item, relevant to the same stakeholder as $i$, has a lower increment in utility than adding $i$ to a subset not containing such an item [19]. Submodular functions is a family of functions modeling the diminishing returns property. Formally, a set function $F$ is said to be submodular if for any two subsets $S \subseteq T$ and item $i \notin T$, the increase in value on adding $i$ to $S$ is at least the increase in value on adding $i$ to $T$, i.e.,

$$F(S \cup \{i\}) - F(S) \geq F(T \cup \{i\}) - F(T).$$

At a high level, the goal in the above applications is to solve the following maximization program for a suitable submodular function $F\colon 2^{[n]} \to \mathbb{R}$

$$\max_{S \subseteq [n]: |S| \leq k} F(S). \tag{1}$$

**A family of submodular functions.** There is a vast literature on maximizing submodular functions [20, 21, 25, 36, 46, 50]. This literature studies various types of submodular functions. We focus on a family of submodular functions that captures many functions used in recommendation and web search. In these applications, each item has $m$ *attributes* and an item $i \in \{1, 2, \ldots, n\}$ generates a value or *utility* $W_{ij} \geq 0$ for a user (or stakeholder) who looking for items with attribute $j$. For instance, on Amazon Music, [49] choose $W_{ij}$ to encode the utility of song $i$ for users interested in songs from genre $j$. They use the submodular function $F(S) = \sum_{j=1}^{m} \log\left(1 + \sum_{i \in S} W_{ij}\right)$ to capture the value of a set $S$ of song recommendations. Another example is [10] who, roughly speaking, set $W_{ij}$ to encode the utility of song $i$ for stakeholder $j$. They use $F(S) = \sum_{i \in S} W_{i1} + \sum_{j=2}^{m} \sqrt{\sum_{i \in S} W_{ij}}$ to capture the value of a playlist $S$ on Spotify. Generalizing these examples, we consider the following family of submodular functions: Given $m$ increasing concave functions $g_1, g_2, \ldots, g_m\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, corresponding to each attribute, and utilities $W$, $F$ is

$$\forall S \subseteq [n], \quad F(S) := \sum_{j=1}^{m} g_j\left(\sum_{i \in S} W_{ij}\right). \tag{2}$$

We denote the family of functions by $\mathscr{F}$. $\mathscr{F}$ captures the function in [49] when $g_j(x) = \log(1 + x)$ for all $x$ and $j$. It captures the function considered by [10] when $g_1(x) = x$ and $g_2(x) = \cdots = g_m(x) = \sqrt{x}$ for all $x$. In Supplementary Material A, we show that $\mathscr{F}$ also captures functions used by [2, 58].

If the utilities $W$ are accurately known, then one can use standard algorithms to approximately solve Program (1) (see Section 2). However, in the above applications, the utilities are often derived from users; either directly from user feedback or indirectly from



*Conference acronym 'XX, June 03–05, 2018, Woodstock, NY*



ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

predictions of learning algorithms trained on user data. Hence, societal biases can creep into such *observed* utilities. Consequently, the subset maximizing the objective defined by these observed utilities can have a sub-optimal value with respect to the objective $F$ defined by the true or *latent* utilities.

**Bias in inputs.** There are various mechanisms through which biases can arise in observed utilities. If certain social groups are over represented in data compared to their proportion in the user-base, then utilities derived from this data will be skewed toward the opinions of these groups. For instance, the IMDB rating of the 2016 remake of Ghostbusters was "sabotaged by a faction of fans who appeared to be upset by its all-female cast" [5, 43]. Apart from such explicit biases, humans also have unconscious implicit biases [6, 8, 23, 38, 60]. Human's implicit biases can manifest in data and, in turn, introduce skews in the utilities [8, 16]. Bias can also arise due to differences in user characteristics across socially-salient groups. For instance, [33] observe that SOTA text summarization algorithms output summaries which under-represent minority dialects of English. [33] postulate that this is due to "structural differences" across dialects (e.g., differences in lengths of sentences or of Tweets). When algorithms are used to estimate utilities, then such algorithmic biases can introduce skews in estimated utilities.

**A model of bias.** To capture such skews, we consider a model that extends the model in [35]. In this model, items belong one of $p$ disjoint groups $G_1, G_2, \ldots, G_p$. Each group $G_\ell$ has an *unknown* and increasing bias function $\phi_\ell : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. The observed utilities of items $i$ in $G_\ell$ are defined as follows

$$\text{for each } 1 \leq j \leq m, \quad \widehat{W}_{ij} := \phi_\ell\left(W_{ij}\right). \tag{3}$$

In the above examples, $G_1, \ldots, G_p$ can correspond to the set of movies whose protagonists are male (respectively non-male) and the set of tweets written in Standard English (respectively African American English). To gain some intuition, consider the special case studied in [35]: for each $\ell$, $\phi_\ell(x) = \beta_\ell \cdot x$ for some $0 < \beta_\ell < 1$. In this case, the observed utilities of items in $G_\ell$ are $\beta_\ell$ times smaller than their latent utilities. A platform that does not observe the latent utilities $W$, naturally outputs the subset $\widehat{S}$ that maximizes the function $\widehat{F}$ defined by observed utilities $\widehat{W}$:

$$\widehat{S} := \operatorname*{argmax}_{T \subseteq [n]: |T| \leq k} \widehat{F}(T), \text{ where } \forall T \subseteq [n],\ \widehat{F}(T) := \sum_{j=1}^{m} g_j\left(\textstyle\sum_{i \in S} \widehat{W}_{ij}\right).$$

Since $\widehat{S}$ optimizes a different objective than $F$, the latent utility of $S$, $F(\widehat{S})$, can be smaller than the optimal latent utility, $\mathrm{OPT} := \max_{S \subseteq [n]: |S| \leq k} F(S)$.

*Given $\widehat{W}$, can we find a set $S$ such that $F(S)$ is close to OPT?*

**Related work.** Recent works [12, 35, 40] have studied the above model in the special case where the objective $F$ is a linear function–a specific type of functions in $\mathscr{F}$–and the bias functions, for each $\ell$, are $\phi_\ell(x) = \beta_\ell \cdot x$ for some parameters $0 < \beta_1, \ldots, \beta_p < 1$. These works explore if requiring the output $S$ to satisfy fairness constraints can improve its latent utility $F(S)$. Various fairness constraints have been considered in practice [4, 59]: Equal representation requires $S$ to have at most $k/p$ items from each group $G_\ell$ and proportional representation requires $S$ to have at most $k \cdot |G_\ell|/n$ items from each group $G_\ell$. Generalizations of these constraints have also been considered, given values $u_\ell$, generalizations require $|S \cap G_\ell| \leq u_\ell \cdot k$ respectively for each $\ell$. There are many reasons to use fairness constraints, including, ethical and legal ones [4, 53, 54, 65]. [12, 35, 40] demonstrate that another benefit of fairness constraints is that they can improve the latent utility of the output. Given $u = (u_1, \ldots, u_p)$, let $S_u$ be the subset maximizing the observed utility $\widehat{F}(S_u)$ subject to satisfying the constraint specified by $u$. In the special case, where $F$ is linear (e.g., $F(S) := \sum_{i \in S} W_{i1}$) and latent utilities $W$ are drawn i.i.d. from some distribution, [12, 40] show that if $u$ captures proportional representation, then $F(S_u) \geq (1 - o_k(1)) \cdot \mathrm{OPT}$. *Thus, a natural question is if there is a $u$ such that $F(S_u)$ is close to OPT for a general function $F \in \mathscr{F}$.*

**Stochasticity in groups.** [12, 35, 40] assume that entries of $W$ are drawn i.i.d. from some distribution. We weaken this assumption: we let $W$ be arbitrary and assume groups $G_1, \ldots, G_p$ are generated stochastically–independent of $W$. This is done by uniformly sampling $|G_1|$ distinct items and assigning them to $G_1$, then sampling $|G_2|$ distinct items (from those remaining) and assigning them to $G_2$, and so on. This captures the belief that their are no systematic differences in the latent utilities of items in different groups. If, in addition, entries of $W$ are also sampled i.i.d., then this is equivalent to the model of [12, 35, 40]. For further discussion of the model and results of [12, 35, 40] appears in Supplementary Material D.

**Our contributions.** Our first result shows that for any $\varepsilon > 0$ and upper bound parameters $u$, there is a submodular function $F$ in $\mathscr{F}$, latent utilities $W$, and functions $\phi_1, \ldots, \phi_p$, such that with high probability, the latent utility of $S_u$ is at most $\varepsilon \cdot \mathrm{OPT}$ (Theorem 4.1). Thus, no choice of $u$ can ensure that $F(S_u)$ is close to OPT. This result holds in the bias functions $\phi_1, \ldots, \phi_p$ studied by [12, 35, 40]. Hence, it contrasts the result of [12, 35, 40] that for linear $F$ and $u$ encoding proportional representation: $F(S_u) \geq (1 - o_k(1)) \cdot \mathrm{OPT}$.

On the positive side, we give an algorithm for submodular maximization in the presence of biases (Algorithm 1). Algorithm 1 can be used with any $F \in \mathscr{F}$. If each item $i$ has a nonzero utility for at most one attribute $1 \leq j \leq m$ (Assumption 1), then it provably outputs a subset with near-optimal latent utility (Theorem 4.3). Assumption 1 is natural in some settings. For instance, [2] used the assumption in the context of web search. Concretely, under Assumption 1, given observed utilities $\widehat{W}$, Algorithm 1 outputs a subset $S$ whose latent utility is at least $(1 - O(\tau^{-1} m^2 k^{-1/4})) \cdot \mathrm{OPT}$, where $\tau > 0$ is the minimum value such that all nonzero entries of $W$ are between $\tau$ and $\tau^{-1}$ (Theorem 4.3).

Algorithm 1 differs from the standard greedy algorithms for constrained monotone submodular maximization [21, 47, 50]. Roughly speaking, in the $t$-th iteration, greedy algorithms select the item with the highest marginal utility (or an approximation of marginal utility) from some subset $S_t$. Algorithm 1 is also iterative, but it first it computes the "right" constraints for the given data and, then, does submodular maximization subject to the computed constraints.

Empirically, we evaluate the performance of Algorithm 1 when Assumption 1 does not hold. We run simulations on the MovieLens 20M [27] and two synthetic datasets and compare against the baseline Uncons, that outputs the subset maximizing the observed utilities. We fix $\phi_1(x) = x$ and $\phi_\ell(x) = \beta \cdot x$ for each $\ell \neq 1$. In simulations with synthetic datasets, we observe that Algorithm 1 achieves a latent utility higher than 0.99 times the latent utility

achieved when $\beta = 1$ (i.e., there is no bias), even for small values of $\beta$ ($\beta \leq 0.01$). Whereas, Uncons outputs subsets whose latent utility decreases with $\beta$ and for $\beta < 0.1$ is up to 12% smaller than OPT. On MovieLens 20M [27], we observe that the predicted relevance scores in the data are disproportionately higher (by up to 3 times) for movies led male actors compared to movies led by non-male actors in genres stereotypically associated with men. In contrast, user ratings for these sets of movies are within 6% of each other in all genres. We use these (biased) relevance scores to recommend movies from sets of men-stereotypical genres and evaluate the performance of recommended movies with user ratings. Algorithm 1 outperforms Uncons by up to 3% on 14/31 genre sets and has a similar as (within 1%) or better performance than Uncons on more than 87% sets of men-stereotypical genres.

## 2 OTHER RELATED WORK

**Fairness in information retrieval and recommendation.** Information retrieval and recommendation systems (such as personalized feed generators, news recommenders, and search engines) have a significant societal influence [51]. They are one of the primary sources of information for individuals [15, 39, 51, 54], who impart significant trust to these systems–tending to agree with the outputs of these systems [14] and to follow their suggestions [24]. Without fairness considerations, the outputs of existing systems have been observed to encode various societal biases [52]–leading to underrepresentation of some social groups [32], polarization of user opinions [44], and denial of economic opportunities available to individuals [26]. Consequently, a growing body of works designs interventions to mitigate the adverse effects of biases [15, 39, 54]. These works can broadly be divided into those mitigating adverse effects on the users [1, 31, 63], those mitigating adverse effects on the items (denoting providers such as journalists in news recommendation or search, artists in song recommendation, and individuals on online hiring platforms) [7, 22, 56, 61, 62, 64], and those which consider both [41, 42]. Works in each of these categories take diverse approaches: from modifying relevance estimation to satisfy fairness criteria [62–64], to requiring the output to satisfy fairness constraints [7, 22, 56, 61], to modifying the objective of system to capture fairness metrics [1, 31, 41, 42]. We focus on harms for the items or providers. We examine the efficacy of fairness constraints to mitigate these harms when output's utility is captured by a submodular function and the inputs are biased. Unlike this work, most prior works assume that the input data is accurate.

**Submodular maximization.** There is a vast literature on maximizing submodular functions subject to different types of constraints [20, 21, 25, 36, 46, 50]. Among these, cardinality constraints are of specific interest. These are constraints of the form $|S| \leq k$ for some fixed $k$, as in Program (1). For cardinality constraints, given an evaluation oracle for $F$, the standard greedy algorithm of [50] selects a subset $S$ of size $k$ such that $F(S) \geq (1 - e^{-1}) \cdot \text{OPT}$ while making at most $nk$ evaluations of $F$ and doing at most $O(nk)$ additional arithmetic operations [50], where $\text{OPT} = \max_{|S| \leq k} F(S)$. Several variants of this algorithm have also been designed. These variants extend the $(1 - 1/e)$-approximation guarantee to other types of constraints (including upper bounds stated in Section 1), improve its running time, and design distributed variants of the algorithm [11, 45, 47, 48]. All of these algorithms, however, assume that one can evaluate the true function $F$. This may not be possible in the presence of biases.

## 3 MODEL

Let there be $n$ items, indexed by the set of values $[n] := \{1, 2, \ldots, n\}$. A set function $F\colon 2^{[n]} \to \mathbb{R}$ is said to be submodular if for each pair of subsets $T \subseteq S \subseteq [n]$ and item $i \in [n]$, $F(S \cup \{i\}) - F(S) \geq F(T \cup \{i\}) - F(T)$. We consider the following family of submodular functions that are studied in the context of content recommendation [10, 17, 49, 58] and web search [2].

Definition 3.1 (**A family of submodular functions**). *$\mathscr{F}$ is the family of all submodular functions $F$ that are parameterized by a number $m$, increasing concave functions $g_1, g_2, \ldots, g_m\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, and matrix $W \in \mathbb{R}_{\geq 0}^{n \times m}$ as follows:*

$$F(S) := \sum_{j=1}^{m} g_j \left( \sum_{i \in S} W_{ij} \right).$$

Setting $g_j(x) = \log(1 + x)$ for all $x$ and $j$, we get the submodular function tested by [49] on Amazon Music. With $g_1(x) = x$ and $g_j(x) = \sqrt{x}$ for all $x$ and $j \neq 1$, we get the submodular function used by [10] to measure the quality of song playlists. Further examples appear in Supplementary Material A.

Given a suitable $F \in \mathscr{F}$, the goal in our motivating applications is to solve the following maximization program:

$$\max_{S \subseteq [n]: |S| \leq k} F(S).$$

If $W$ and, hence, $F$ is known, then one can hope to find a subset $S$ such that $F(S)$ is close to the optimal value, OPT, of the above program. However, as discussed, in many contexts, the utilities observed by a platform $\widehat{W}$ can encode societal biases. Hence, $\widehat{W}$ can be different from the true or *latent* utilities $W$. Here, we consider a model of bias that builds on [35]. In this model, items belong to one of $p$ disjoint groups $G_1, G_2, \ldots, G_p$. We weaken the assumption in [35] (and related models in [12, 40]) by allowing $W$ to be arbitrary and requiring $G_1, G_2, \ldots, G_p$ to be generated stochastically. In particular, given sizes $|G_1|, \ldots, |G_p|$, $G_1$ is constructed by selecting $|G_1|$ items uniformly without replacement, $G_2$ is constructed by selecting $|G_2|$ items uniformly from those remaining, and so on. We define $\gamma > 0$ to be a constant such that for each $1 \leq \ell \leq p$

$$|G_\ell| \geq \gamma n.$$

The model of bias is as follows.

Definition 3.2 (**Model of bias in utilities**). *For each $\ell$, there is an unknown and increasing bias function $\phi_\ell\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that the observed utilities of items $i \in G_\ell$ are: For all $1 \leq j \leq m$*

$$\widehat{W}_{ij} := \phi_\ell \left( W_{ij} \right).$$

Supplementary Material B extends this to overlapping groups. The specific groups vary with application and could, for instance, be defined by socially salient attributes (e.g., gender, race, or age) of associated with each item (see examples in Section 1.) For any $F \in \mathscr{F}$, parameterized by $W$, let $\widehat{F}$ be the corresponding function parameterized by $\widehat{W}$.

Problem 1. *Given $m$ functions $g_1, \ldots, g_m$ and an $n \times m$ matrix $\widehat{W}$, parameterizing a monotone submodular function $\widehat{F}$, without knowledge of the specific bias functions $\phi_1, \ldots, \phi_p$, find a subset $S$ of size at most $k$ such that $F(S) \approx OPT$.*

# 4 THEORETICAL RESULTS

## 4.1 Upper bound on latent utility achieved with fairness constraints

Our first result considers a family of fairness constraints and shows that no fairness constraints in this family can guarantee a constant fraction of the optimal latent utility.

Let $\gamma_\ell$ be the fraction of items in $G_\ell$, i.e., $\gamma_\ell := \frac{|G_\ell|}{n}$, for each $\ell$. Given $u$ and $v$, the constraint requires the output subset $S$ to satisfy

$$\forall \ell \in [p], \quad |S \cap G_\ell| \leq (u_\ell + v_\ell \gamma_\ell) \cdot k.$$

This family captures equal representation when $u_\ell = \frac{1}{p}$ and $v_\ell = 0$ for each $\ell$ and proportional representation when $u_\ell = 0$ and $v_\ell = 1$ for each $\ell$. Given $(u, v)$, let $S_{uv}$ be the subset that maximizes the observed utility $\widehat{F}(\cdot)$ subject to satisfying the constraints specified by $(u, v)$.

Theorem 4.1 (**Fairness constraints do not guarantee any fraction of OPT**). *Define $\phi_1(x) = \beta_1 \cdot x$ and $\phi_2(x) = \beta_2 \cdot x$ for all $x$. For any $0 < \varepsilon < 1$, $u = (u_1, u_2)$, and $v = (v_1, v_2)$ there exists*

- *a submodular function $F \in \mathscr{F}$,*
- *$0 \leq \gamma_1, \gamma_2 \leq 1$,*
- *$0 < \beta_1, \beta_2 \leq 1$, and*
- *family of $n \times m$ matrices $\mathcal{W}$ parameterized by $n$,*

*such that, for any $k \geq \text{poly}(\varepsilon^{-1})$, $n \geq k \cdot \text{poly}(\varepsilon^{-1})$, and $W \in \mathcal{W}(n)$, it holds that*

$$\Pr\left[F(S_{uv}) \leq \varepsilon \cdot OPT\right] \geq 1 - \varepsilon,$$

*where the probability is over the randomness in $G_1$ and $G_2$.*

Thus, for any fairness constraint $(u, v)$ in the above family and any $\varepsilon > 0$, there is a submodular function $F \in \mathscr{F}$ and family of utilities for which the subset $S_{uv}$ outputted by fairness constrained-maximization has utility at most $\varepsilon \cdot$ OPT with high probability. Theorem 4.1 straightforwardly generalizes to $p > 2$ groups by adding empty groups. It also generalizes to $m > 3$ attributes by fixing utilities so that $W_{ij} = 0$ for each item $i$ and $j \in [m] \backslash [3]$. The proof of Theorem 4.1 appear in Supplementary Material F.1 available in the anonymized PDF file here.

[12, 18, 40] suggest to use the proportional representation constraints with (1) a linear function $F$ and (2) the bias functions of the form $\phi_\ell(x) = \beta_\ell \cdot x$ for all $x$ and $j$, for some parameters $0 < \beta_1, \ldots, \beta_p < 1$. Since the above family captures proportional representation constraint and Theorem 4.1 holds for this choice of $\phi_\ell$, Theorem 4.1 shows that while requiring the output to satisfy the proportional representation constraints guarantees representation, it can lead to significant loss in latent utility.

## 4.2 Algorithmic result

In this section, we present our algorithmic result (Algorithm 1). Algorithm 1 can be used for any submodular function in $\mathscr{F}$ and outputs subsets that proportionally represent items from each group. At the same time, the output subsets provably have a near-optimal latent utility for a subset of functions in $\mathscr{F}$ that satisfy an *algorithmically verifiable* assumption (Assumption 1).

For each $j$, let $C_j$ be the set of items $i$ which have positive utility for the $j$-th attribute, i.e.,

$$C_j := \left\{i \in [n] : W_{ij} > 0\right\}.$$

Intuitively, $C_j$ is the set of items that are relevant to "attribute $j$." More concretely, in many of our motivating submodular functions, attributes correspond to different categories of item (e.g., genres, topics, or retail-types) [2, 34, 49]. Here, $C_j$ is the set of items in the $j$-th category.

Assumption 1 (**Disjoint categories**). *The matrix $W \in \mathbb{R}^{n \times m}$ is such that $C_1, \ldots, C_m$ are disjoint.*

Assumption 1 holds in any context where the relevant categories of items are disjoint. In the context, of web search, the above assumption is identified and studied by [2].

Deviations from Assumption 1 can deteriorate the performance of Algorithm 1. In Section 5, we evaluate the performance of Algorithm 1 on MovieLens 20M data [27] that does not satisfy Assumption 1 (see Figure 2). In the MovieLens 20M data the sets $C_1, \ldots, C_m$ denote genders of movies and can be non-disjoint when a movie has more than one genres. We also evaluate Algorithm 1 on synthetic data that violates Assumption 1 and is inspired by the deployment of submodular-maximization based algorithms by [10] (see Figure 1(b)).

The performance of Algorithm 1 also depends on the range of the non-zero entries of $W$. This dependence arises because the concentration of the sum $\sum_{i \in S \cap G_\ell} W_{ij}$ around its mean depends on this range (where the randomness is due to $G_\ell$). $\tau$ is a parameter that captures this range.

Definition 4.2. *Let $\tau > 0$ be the smallest constant such that for each item $i \in [n]$ and attribute $j \in [m]$, either $W_{ij} = 0$ or $\tau < W_{ij} < 1/\tau$.*

Theorem 4.3. *Suppose $\phi_1(x) = x$ for each $x$. There is an algorithm (Algorithm 1) that, given observed utilities $\widehat{W} \in \mathbb{R}^{n \times m}$ and evaluation oracles for $g_1, \ldots, g_m : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, outputs a set $S$ of size $k$ with the following property (under Assumption 1): For any $\varepsilon, \tau, \gamma > 0$ and any increasing functions $\phi_2, \ldots, \phi_p : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, there is a large enough $k_0$ such that for any $k \geq k_0$, with probability at least $1 - \varepsilon$*

$$F(S) \geq OPT \cdot (1 - \varepsilon).$$

*Where the probability is over the randomness in the choice of $G_1, \ldots, G_p$. The algorithm makes $O(nk)$ evaluations of each $g_1, \ldots, g_m$ and does $O(nk \log n)$ additional arithmetic operations.*

Thus, for large $k$, Algorithm 1 achieves near-optimal latent utility *without knowing* the bias functions $\phi_2, \ldots, \phi_p$.
Theorem 4.3 assumes $\phi_1(x) = x$. Up to re-indexing the groups and bias functions, it suffices to have $\phi_\ell(x) = x$ for all $x$ for at least one $1 \leq \ell \leq p$. This encodes the assumption that utilities of items in at least one group does not have any skew. Theorem 4.3 can also be extended (with the same proof) to a generalization of the bias model where the bias parameter of item $i$ not only depends on the protected group(s) which $i$ is in but also on the attribute $j$: For each $1 \leq \ell \leq p$ and attribute $j$ there is an increasing function

$\phi_{\ell j}: \mathbb{R} \to \mathbb{R}$. The observed utilities of an item $i \in G_t$ are

$$(\widehat{W}_{i1}, \widehat{W}_{i2}, \ldots, \widehat{W}_{im}) = (\phi_{t1}(W_{i1}), \phi_{t2}(W_{i2}), \ldots, \phi_{tm}(W_{im})). \quad (4)$$

This reduces to the model of bias in Definition 3.2 when $\phi_{t\ell} = \phi_\ell$. The parameter $k_0$ is $\widetilde{O}\left(\varepsilon^{-4}\tau^{-4}\gamma^{-10}m^8\right)$. The proof of Theorem 4.3 appears in Supplementary Material F.2 available in the anonymized PDF file here.

For a general submodular function, it is **NP**-hard to output a set $S$ with $F(S) \geq (1 - e^{-1}) \cdot \mathrm{OPT}$ [20]. Algorithm 1 gives a stronger guarantee (for large $k$) because this hardness result does not hold under Assumption 1: if $\phi_1, \ldots, \phi_p$ are known and Assumption 1 holds, then there is an efficient algorithm to find a subset $S$ with $F(S) = \mathrm{OPT}$.

Finally, we remark that in the statements of our results, we have tried to capture the dependence on each parameter as cleanly as possible and not tried to optimize the constants.

*4.2.1 Description of Algorithm 1.* It has two parts, where Part 1's output fixes Part 2's parameters.

- *(Part 1)* The first part uses a greedy algorithm to select a subset $S' \subseteq G_1$ of size $k \cdot \frac{|G_1|}{n}$ (where $\beta_1 = 1$) that maximizes a re-scaled version of the observed utility (Equation (5)) subject to selecting at least $\sqrt{k}$ elements from each set $C_j$ ($j \in [m]$). Using $S'$ it constructs values $k'_j := |S' \cap C_j|$ ($j \in [m]$), which are used to specify the constraints in the next part.
- *(Part 2)* This part solves $m$ different optimization problems, one for each attribute: For each attribute $j$, it selects a subset $S_j$ of at most $k'_j$ items that maximize the observed utility for the $j$-th attribute ($g_j(\sum_{i \in S} W_{ij})$) while satisfying proportional representation constraint. Algorithm 1 outputs $\cup_{j=1}^m S_j$.

## 4.3 Proof overviews of theoretical results

*4.3.1 Technical challenges in extending approaches for linear $F$.* Let us consider $F(S) = \sum_{i \in S} W_{i1}$ (i.e., $m = 1$ and $g(x) = x$) and the subset $S_u$ which maximizes the observed utility for $F$ subject to satisfying proportional representation constraints. [40] show that $F(S_u) \geq (1 - o_k(1)) \cdot \mathrm{OPT}$. This result relies on two properties of any linear function $F$:

(1) For any subset $S$,

$$F(S) = \sum_{\ell \in [p]} F(S \cap G_\ell)$$

(2) For any increasing functions $\phi_1, \ldots, \phi_p$, $\ell \in [p]$, and $R \subseteq [n]$,

$$\operatorname*{argmax}_{S \subseteq R \cap G_\ell : |S| \leq k} \widehat{F}(S) = \operatorname*{argmax}_{S \subseteq R \cap G_\ell : |S| \leq k} F(S).$$

Suppose $S_{\mathrm{OPT}}$ satisfies proportional representation. Then, using property 1, one can show that $S_u = \bigcup_{\ell=1}^p \widehat{T}_\ell$ and $S_{\mathrm{OPT}} = \bigcup_{\ell=1}^p T_\ell$, where

$$\widehat{T}_\ell := \operatorname{argmax}_{T \subseteq G_\ell, |T| \leq \frac{k}{n} \cdot |G_\ell|} \widehat{F}(T)$$

and

$$T_\ell := \operatorname{argmax}_{T \subseteq G_\ell, |T| \leq \frac{k}{n} \cdot |G_\ell|} F(T).$$

Further, because of property 2 (with $R = [n]$), it follows that for any $\ell \in [p]$, $\widehat{T}_\ell = T_\ell$ and, hence, $S_u = S_{\mathrm{OPT}}$. This relies on the assumption that $S_{\mathrm{OPT}}$ satisfies proportional representation. This may not be true, but one can show that with high probability $S_{\mathrm{OPT}}$ nearly-satisfies proportional representation. Using this and an approximate version of the above argument one can show that $F(S_u) \approx F(S_{\mathrm{OPT}})$. However, unfortunately straightforward examples show that neither property holds for submodular functions in $\mathscr{F}$.

**Algorithm 1** Near-optimal latent utility under Assumption 1

**Input:** A matrix $\widehat{W} \in \mathbb{R}^{n \times m}$, a number $k$, groups $G_1, \ldots, G_p$, and sets $C_1, \ldots, C_m$ (with guarantee that $\phi_1(x) = x$).
**Output:** A subset $S$ with $|S| \leq k$

▷ *Part 1: Compute data-dependent representation constraints*

Initialize $\widetilde{S} = \emptyset$
Define the following function

$$\forall T \subseteq [n], \quad \widetilde{F}(T) := \sum_{j=1}^m g_j\left(\frac{n}{|G_1|} \cdot \sum_{i \in T \cap G_1} \widehat{W}_{ij}\right). \quad (5)$$

**for** $j \in [m]$ **do**
  Sort items $i$ in $C_j \cap G_1$ in decreasing order of $\widetilde{F}(i) - \widetilde{F}(\emptyset)$
  Add the first $\min\{\sqrt{k}, |C_j \cap G_1|\}$ items in $\widetilde{S}$
**end for**
**while** $|\widetilde{S}| < k \cdot \frac{|G_1|}{n}$ **do**
  Set $\widetilde{S} := \widetilde{S} \cup \{i\}$ where $i := \operatorname{argmax}_{i \in G_1} \widetilde{F}(\widetilde{S} \cup \{i\}) - \widetilde{F}(\widetilde{S})$
**end while**
Let $k_j := |\widetilde{S} \cap C_j|$ for each $1 \leq j \leq m$

▷ *Part 2: For each $1 \leq j \leq m$, select a set $S_j \subseteq C_j$ with $k_j$ items that maximizes observed utilities while satisfying proportional representation*

Initialize $S_j := \emptyset$ for each $1 \leq j \leq m$
**for** $j \in [m]$ and $t \in [k_j]$ **do**
  Define $C := \left\{i \in C_j : \forall \ell \in [p], \ |(\{i\} \cup S_j) \cap G_\ell| \leq \frac{k_j |G_\ell|}{n}\right\}$ ▷ The set of items which can be added without violating prop. repr.
  Let $i := \operatorname{argmax}_{i \in C} g_j(S_j \cup \{i\}) - g_j(S_j)$ ▷ Where $g_j(T)$ is shorthand for $g_j(\sum_{i \in T} W_{ij})$
  Update $S_j := S_j \cup \{i\}$
**end for**
**return** $S = \cup_{j=1}^m S_j$

*4.3.2 Proof overview of Theorem 4.3.* Under Assumption 1, we prove the following variants of properties 1 and 2 stated on the previous page:

- For any subset $S$,

$$F(S) = \sum_{j \in [m]} \sum_{\ell \in [p]} F(S \cap G_\ell \cap C_j)$$

- For any increasing functions $\phi_1, \ldots, \phi_p$, $\ell \in [p]$, and $j \in [m]$

$$\operatorname*{argmax}_{S \subseteq C_j \cap G_\ell : |S| \leq k} \widehat{F}(S) = \operatorname*{argmax}_{S \subseteq C_j \cap G_\ell : |S| \leq k} F(S).$$

*Observation.* Let $S_{\mathrm{OPT}}$ be the subset such that $F(S_{\mathrm{OPT}}) = \mathrm{OPT}$. Suppose we know $k_j := |S_{\mathrm{OPT}} \cap C_j|$ for all $j$. If for all $j$, $k_j \geq \sqrt{k}$, then using the above properties (and the analysis of [40]), we can show that, with high probability, $S := \bigcup_{j=1}^m S_j(k_j)$ has latent utility $(1 - o_k(1)) \cdot F(S_{\mathrm{OPT}})$. Where, for each $j$, $S_j(k_j) \subseteq C_j$ is the subset that maximizes the observed utility subject to selecting at most $k_j$ items and satisfying proportional representation constraint.

Part 1 of Algorithm 1 estimates $k_1, \ldots, k_j$ as $\tilde{k}_1, \ldots, \tilde{k}_j$. For this, it relies on the fact that $\phi_\ell(x) = x$ for some (known) group $\ell$ and that $G_1, \ldots, G_p$ are constructed stochastically. The estimated $\tilde{k}_1, \ldots, \tilde{k}_j$ serve as constraints for Part 2. Part 2 of Algorithm 1 outputs $S := \bigcup_{j=1}^{m} S_j(\tilde{k}_j)$. It is possible that for some $j$, $\left|S_{\text{OPT}} \cap C_j\right| < \sqrt{k}$ and, hence, this observation does not apply. Algorithm 1 avoids this by overestimating $k_j$ so that $\tilde{k}_j \geq \sqrt{k}$ for all $j$. At a high level, this guarantees that $S$ contains any "high-utility" item in $\left|S_{\text{OPT}} \cap C_j\right|$ with high probability.

# 5 EMPIRICAL RESULTS

We evaluate Algorithm 1's performance on both synthetic and real-world data.[1]

## 5.1 Baselines and setup

We compare Algorithm 1's performance against two baselines: Uncons and ProportionalRepr. Uncons, given $k$ and observed utilities $\widehat{W}$, runs the standard greedy algorithm of [50] to find a subset of size $k$ that approximately maximizes the observed utility (Algorithm 2). ProportionalRepr, given $k$ and observed utilities $\widehat{W}$, uses a variant of the greedy algorithm that works with proportional representation constraints (Algorithm 3). It outputs the subset of size $k$ that approximately maximizes the observed utility subject to selecting a proportional number of items from each group.

In all simulations, we generate latent and observed utilities $W$ and $\widehat{W}$ (as explained in subsequent sections) and run algorithms with the following inputs: Algorithm 1 are given $\widehat{W}$, $k$, and the protected groups and Uncons is given $\widehat{W}$ and $k$.

## 5.2 Simulations with synthetic datasets

In this simulation, we show that Algorithm 1 can achieve high latent utility even in some cases where Assumption 1 does hold. We consider two synthetic datasets inspired by the recommendation algorithms used on Spotify [10] and tested on Amazon music [49]. Among these, the first dataset and the corresponding submodular function does not satisfy Assumption 1.

**Setup.** In both simulations, the task is to recommend a set of $k := 50$ songs to the current user. We set $n := 250$, $m := 3$, and consider two disjoint groups.

*Synthetic dataset 1.* In the first simulation, we fix the objective as

$$F(S) := \sum_{i \in S} W_{i1} + \lambda \sum_{j=2}^{3} \sqrt{\sum_{i \in S} W_{ij}}$$

with $\lambda = \frac{1}{20}$.[2] Here, $W_{i1} \geq 0$, $W_{i2} \in \{0, 1\}$, and $W_{i3} \in \{0, 1\}$ denote the a some measure of the song's popularity $i$, whether $i$ is from an "emerging artist," and whether the song has not been heard by the current user respectively. Intuitively, among songs $i$ with similar popularity (i.e., similar $W_{i1}$), songs from emerging-artists and those not heard by the current user have a higher marginal utility. We draw the the entries of $W$ i.i.d. from natural distributions that can arise in these contexts (see Supplementary Material C for details). We divide the items into two groups $G_1$ and $G_2$; and vary the size of $G_1$ among $\{0.25n, 0.5n, 0.75n\}$. Here, $G_2$ can denote songs from artists that users want to hear, but which are nevertheless under recommended due to biases in the recommendation pipeline, e.g., as have been recently observed for regional music on Spotify in India [55]. Given $G_1, G_2$ and a parameter $\beta \in [0, 1]$, we generate observed utilities $\widehat{W}$ as follows:

$$\widehat{W}_{i1} = \begin{cases} W_{i1}, & \text{if } i \in G_1 \\ \beta \cdot W_{i1} & \text{if } i \in G_2. \end{cases}, \qquad \widehat{W}_{i2} = W_{i2}, \quad \text{and} \quad \widehat{W}_{i3} = W_{i3}.$$

For the first attribute, this corresponds to using $\phi_1(x) = x$ and $\phi_2(x) = \beta \cdot x$ for all $x$. For the last two attributes, we do not apply bias because they encode values that may be known by the platform. This violates the model in Definition 3.2 (which assume that the same bias function acts on all attributes), hence, is a hard case for Algorithm 1. It, however, falls into the extension of this model discussed in Equation (4).

*Synthetic dataset 2.* In the second simulation, we fix the objective:[3]

$$F(S) := \sum_{j=1}^{3} \log(1 + \sum_{i \in S} W_{ij}).$$

Here, attributes denote genre and the sets $C_1, C_2, C_3$ are disjoint. For any item $i$ in genre $h(i)$, $W_{ih(i)}$ is number of times users played song $i$ and $W_{ij} = 0$ for $j \neq h(i)$. At a high level, this promotes the content to be diverse across genres as between two items $i$ and $j$ with a similar number of user plays, $i$ has a higher marginal utility if $\sum_{r \in S} W_{rh(i)} < \sum_{r \in S} W_{rh(j)}$. Like the previous simulation, we draw entries of $W$ i.i.d. from natural distributions that can arise in these contexts (see Supplementary Material C). We divide items into two protected groups $G_1$ and $G_2$; and vary the size of $G_1$ among $\{0.25n, 0.5n, 0.75n\}$.

The complete implementation details of these simulations appear in Supplementary Material C. In addition, additional simulations with more than two protected groups are appear in the anonymized PDF file here.

**Results and discussion.** We vary $\beta \in [0, 1]$ and $\frac{|G_1|}{n} \in \{0.25, 0.5, 0.75\}$, and report the normalized latent utilities of different algorithms. Figure 1 presents the results for $\frac{|G_1|}{n} = 0.5$. The results with $\frac{|G_1|}{n} \in \{0.25, 0.75\}$ appear in Figures 3 and 4 in the anonymized PDF file here. Across all figures and both synthetic datasets, Algorithm 1 outputs subsets with NLU $> 0.99$ (even for small values of $\beta$). In contrast, when $\beta$ approaches 0 and $\frac{|G_1|}{n} = 0.5$, Uncons achieves NLU $\leq 0.8$.

Thus, we observe that Algorithm 1 can achieve high latent utility. Since this observation also holds on the first synthetic dataset–where Assumption 1 does not hold–we observe that Algorithm 1 can also achieve high latent utility in (some) cases where Assumption 1 is violated.

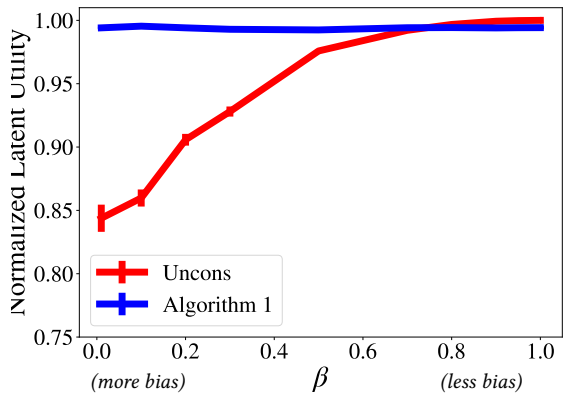
## 5.3 Results on real-world data

In this simulation, we evaluate the performance of Algorithm 1 on real-world data with pre-existing bias (as discussed below) and

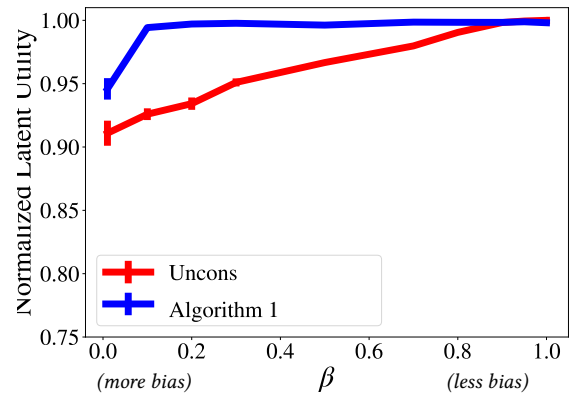
[1] Anonymized code for our simulations is available at https://github.com/submodular-bias/code

[2] This is the same as objective function used by the recommendation algorithm, Mostra, on Spotify [10]. Except that Mostra considers more than $m = 3$ attributes and instead $W_{i1}$ encoding the number of song-plays, it encodes a "relevance score" predicted by a learning algorithm.

[3] This is the same as objective function used by [49] except that they allow $m \geq 3$.




(a) $\phi_1(x) = x$, $\phi_2(x) = \beta \cdot x$, $\frac{|G_1|}{n} = 0.5$, and $\delta = 2.0$




(b) $\phi_1(x) = x$, $\phi_2(x) = \beta \cdot x$, $\frac{|G_1|}{n} = 0.5$, and $\delta = 2.0$

**Figure 1: *Simulation on synthetic data:* We run Algorithm 1 and Uncons (Figures 1(a) and 1(b)) on synthetic data and report their normalized latent utility (error bars denote standard error of the mean). The results show that Uncons can lose significant fraction of the optimal latent utility in the presence of bias (up to 15% for $\beta < 0.1$) while Algorithm 1 have a normalized latent utility higher than 0.99.**

show that it can outperform Uncons and ProportionalRepr, even when the data does not satisfy Assumption 1.

*Data.* MovieLens 20M [27] contains 20 million user ratings (on a scale of 0 to 5) for 27,000 movies submitted by 138,000 users of the movielens.org. For each movie $i$, apart from user-rating, the data has a set of genres of $i$ and relevance scores $r_{ig} \in [0, 1]$ of each genre $g$ encoding "how strongly [movie $i$] exhibits particular properties represented by [genre $g$]." In addition, we gathered information about the lead actor (the first-listed cast member) of each movie from movielens.org.

*Preprocessing.* For each movie $i$, we predict the (probable) gender of its lead actor using the Genderize API (gender-api.com) and remove all movies where this prediction has confidence less than 0.9; this leaves 6612 and 1990 movies led by male and non-male actors respectively.[4] Among the remaining movies, movies led by male actors have a disproportionately higher relevance-scores on genres that are stereotypically associated with men (e.g., "Action" or "War") compared to movies led by non-male actors (differing by up to 300%; see Table 1). In contrast to relevance scores, user ratings are relatively balanced across movies led by male and non-male actors across all genres (the difference is at most 6% across all genres, see Table 2). (This observation also holds if we consider all 27,000 movies; see Table 3.) Hence, compared to the user ratings, the relevance scores are systematically lower for movies led by an actor of a non-stereotypical gender in many genres.

*Setup.* Given a subset $T$ of genres, the task is to recommend $k$ movies from genres $T$ to the users. For each genre $g$, let $R_g$ be the ratio of the average relevance score of movies in this genre led by male actors and non-male actors. We select all genres $g$ where $R_g \geq 2$, these are $B = \{\texttt{action}, \texttt{adventure}, \texttt{crime}, \texttt{western}, \text{and } \texttt{war}\}$. Given a set $M$ of movies and a subset of selected genres $T \subseteq B$, we recommend a subset of movies $S$ that maximizes the following "observed utility:"

$$\widehat{F}(S) := \sum_{g \in T} \sqrt{\sum_{i \in S} r_{ig}}.$$

This function captures that benefit of recommending movies that are relevant to the selected genres $T$ and the $\sqrt{\cdot}$ captures the diminishing return of recommending multiple movies from the same genre $g \in T$. We use the user ratings to evaluate the quality of the recommended movies (or their "latent utility"): Given a set of movies $S$, we say its latent utility is

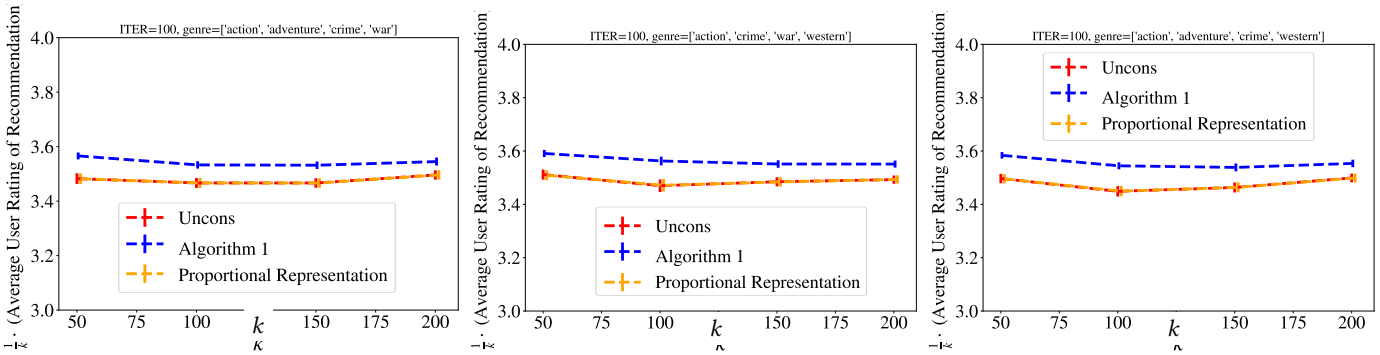
$$F(S) := \frac{1}{|S|} \sum_{i \in S} \text{rat}_i,$$

where $\text{rat}_i$ is the average user rating of movie $i$.

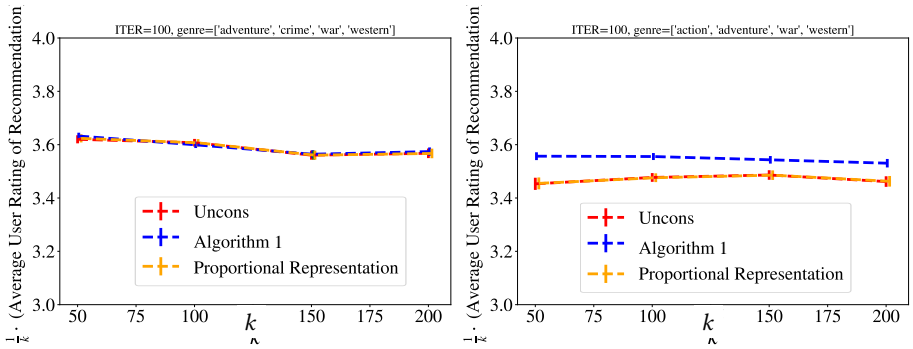
In the simulation, we vary $k \in \{50, 100, 150, 200\}$ and select different nonempty subsets $T \subseteq B$ (31 such subsets). For each $k$ and $T$, we repeat the simulation 100 times. In each iteration, we draw a user $u$ uniformly at random from the set of users who have rated at least 200 movies (19% of the total users). We set $M$ to be the set of movies rated by user $u$; intuitively, this ensures that $M$ contains movies that are watched by users on the platform. Assumption 1 requires that each movie has a unique genre. This does not hold, but we still use Algorithm 1 (without modifications).

*Results.* We report the normalized latent utilities of Uncons, ProportionalRepr, and Algorithm 1 in Figure 2 (for all subsets $T$ of size 4). The results for the remaining choices of $T$ appear in Figures 5 to 9 in the anonymized PDF file here. Across all choices of $T$, we observe that Algorithm 1 achieves 3% higher quality than Uncons and ProportionalRepr for 14/31 subsets of stereotypical genres (>45%), has similar quality (within 1%) as Uncons or ProportionalRepr in 13/31 subsets of stereotypical genres (42%), and has up to 2% lower quality than Uncons in 4/31 choices of genres (13%).

[4] We choose a high threshold of 0.9 to ensure that the gender predictions have a low error rate among the remaining movies. We repeated the simulation with thresholds 0.7 and 0.8 and observed similar results.




(a) Action, Adventure, Crime, War (b) Action, Crime, Was, Western (c) Action, Adventure, Crime, Western




(d) Adventure, Crime, War, Western (e) Action, Adventure, War, Western

**Figure 2: *Simulation on MovieLens 20M data:* The relevance scores in the data are disproportionately higher (up to 3 times) for movies led male actors compared to movies led by non-male actors, in genres stereotypically associated with men. In contrast, user ratings for these sets of movies are within 6% for each genre. We use relevance scores to recommend $k \in \{50,100,150,200\}$ movies from different subsets of men-stereotypical genres and use user ratings to estimate the latent utility of the recommended movies. Figures 2(a) to 2(e) present results all subsets of size 4–we observe that Algorithm 1 achieves 1.40% and 0.28% higher normalized latent utility than Uncons and ProportionalRepr for all $k$. (Results for other genre subsets of other sizes appear in Figures 5 to 9 in the anonymized PDF file here.)**

## 6 LIMITATIONS AND CONCLUSION

This work studies maximization of a family of submodular functions that have been used to capture the utility of item-subsets in recommendation systems and web search. In particular, it studies the setting where the inputs defining the submodular function have social biases–modeled by an extension of [35]'s bias model–and these biases lead to reduction in the latent utility of the output subset. Our first result shows that maximizing the observed utility subject to fairness constraints is not sufficient to recover any fraction of the optimal latent utility (Theorem 4.1). On the positive side, we give an algorithm (Algorithm 1) for submodular maximization that works for the family of submodular functions we consider. Under mild assumptions, it provably outputs a subset with near-optimal latent utility (Theorem 4.3). Empirically, the subsets output by this algorithm have higher latent utility than baselines even when the assumptions required by the theoretical results do not hold (Figures 1 and 2).

Our work raises several questions for future work. Empirical results on real-world data, showed that our algorithm can outperform baselines, even in some cases where data does not follow the theoretical model considered. However, a careful assessment of our algorithms' performance on application-specific data, both pre- and post- deployment, would be important to avoid any unintended harms. Further, submodular maximization is one part in the larger information retrieval or recommendation system; examining the effect of biases in the input on other parts of the system and evaluating the efficacy of our algorithm in conjunction with the broader system are interesting directions.

# REFERENCES


[1] Himan Abdollahpouri. *Popularity bias in recommendation: a multi-stakeholder perspective.* PhD thesis, University of Colorado at Boulder, 2020.

[2] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying Search Results. In *WSDM*, pages 5–14. ACM, 2009.

[3] Rémi Bardenet and Odalric-Ambrym Maillard. Concentration Inequalities for Sampling Without Replacement. *Bernoulli*, 21(3):1361–1385, 2015.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[5] Alyssa Bereznak. The Problem With IMDb's Rating System, January 2019. https://www.theringer.com/tv/2019/6/12/18661850/imdb-rating-system-problems-chernobyl.

[6] Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American economic review*, 94(4):991–1013, 2004.

[7] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *SIGIR*, pages 405–414. ACM, 2018.

[8] Iris Bohnet. *What Works: Gender Equality by Design.* Harvard University Press, 2016.

[9] Bert Boyce. Beyond topicality: A two stage view of relevance and the retrieval process. *Information Processing & Management*, 18(3):105–109, 1982.

[10] Emanuele Bugliarello, Rishabh Mehrotra, James Kirk, and Mounia Lalmas. Mostra: A Flexible Balancing Framework to Trade-off User, Artist and Platform Objectives for Music Sequencing. In *WWW*, pages 2936–2945. ACM, 2022.

[11] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a Monotone Submodular Function Subject to a Matroid Constraint. *SIAM Journal on Computing*, 40(6):1740–1766, 2011.

[12] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. Interventions for Ranking in the Presence of Implicit Bias. In *FAT**, pages 369–380. ACM, 2020.

[13] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-Law Distributions in Empirical Data. *SIAM review*, 51(4):661–703, 2009.

[14] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. Is seeing believing? how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, page 585–592, New York, NY, USA, 2003. Association for Computing Machinery.

[15] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, 16(1-2):1–177, 2022.

[16] Michael D. Ekstrand and Daniel Kluver. Exploring author gender in book rating and recommendation. *User Modeling and User-Adapted Interaction*, 31(3):377–420, 2021.

[17] Khalid El-Arini, Gaurav Veda, Dafna Shahaf, and Carlos Guestrin. Turning down the Noise in the Blogosphere. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 289–298, New York, NY, USA, 2009. Association for Computing Machinery.

[18] Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. On Fair Selection in the Presence of Implicit and Differential Variance. *Artificial Intelligence*, 302:103609, 2022.

[19] Spotify Engineering. Reach for the Top: How Spotify Built Shortcuts in Just Six Months, April 2020. https://engineering.atspotify.com/reach-for-the-top-how-spotify-built-shortcuts-in-just-six-months/.

[20] Uriel Feige. A Threshold of $\ln n$ for Approximating Set Cover. *J. ACM*, 45(4):634–652, jul 1998.

[21] Marshall L Fisher, George L Nemhauser, and Laurence A Wolsey. An Analysis of Approximations for Maximizing Submodular Set Functions —II. In *Polyhedral combinatorics*, pages 73–87. Springer, 1978.

[22] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *KDD*, pages 2221–2231. ACM, 2019.

[23] Anthony G Greenwald and Linda Hamilton Krieger. Implicit Bias: Scientific Foundations. *California Law Review*, 94(4):945–967, 2006.

[24] Ulrike Gretzel and Daniel R. Fesenmaier. Persuasion in recommender systems. *International Journal of Electronic Commerce*, 11(2):81–100, 2006.

[25] Anupam Gupta, Aaron Roth, Grant Schoenebeck, and Kunal Talwar. Constrained Non-monotone Submodular Maximization: Offline and Secretary Algorithms. In Amin Saberi, editor, *Internet and Network Economics*, pages 246–257, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[26] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1914–1933, 2017.

[27] F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, December 2015.

[28] Lu Hong and Scott E Page. Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.

[29] Don Hush and Clint Scovel. Concentration of the Hypergeometric Distribution. *Statistics & Probability Letters*, 75(2):127–132, 2005.

[30] Lars Bo Jeppesen and Karim R Lakhani. Marginality and Problem-Solving Effectiveness in Broadcast Search. *Organization science*, 21(5):1016–1033, 2010.

[31] Toshihiro Kamishima and Shotaro Akaho. Considerations on recommendation independence for a find-good-items task. 2017.

[32] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In Bo Begole, Jinwoo Kim, Kori Inkpen, and Woontack Woo, editors, *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 3819–3828, Seoul, Republic of Korea, 2015. ACM.

[33] Vijay Keswani and L. Elisa Celis. Dialect Diversity in Text Summarization on Twitter. In *WWW*, pages 3802–3814. ACM / IW3C2, 2021.

[34] Jon Kleinberg and Maithra Raghu. Team Performance with Test Scores. *ACM Trans. Econ. Comput.*, 6(3–4), oct 2018.

[35] Jon M. Kleinberg and Manish Raghavan. Selection Problems in the Presence of Implicit Bias. In *ITCS*, volume 94 of *LIPIcs*, pages 33:1–33:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

[36] Andreas Krause and Daniel Golovin. Submodular Function Maximization. *Tractability*, 3:71–104, 2014.

[37] Hui Lin, Jeff Bilmes, and Shasha Xie. Graph-Based Submodular Selection for Extractive Summarization. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 381–386. IEEE, 2009.

[38] Karen S Lyness and Madeline E Heilman. When Fit Is Fundamental: Performance Evaluations and Promotions of Upper-Level Female and Male Managers. *Journal of Applied Psychology*, 91(4):777, 2006.

[39] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

[40] Anay Mehrotra, Bary S. R. Pradelski, and Nisheeth K. Vishnoi. Selection in the Presence of Implicit Bias: The Advantage of Intersectional Constraints. In *FAccT*, page To appear. ACM, 2022.

[41] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. Auditing search engines for differential satisfaction across demographics. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 626–633, 2017.

[42] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2243–2251, New York, NY, USA, 2018. Association for Computing Machinery.

[43] Scott Mendelson. 'Ocean's 8' Can Afford To Ignore The Trolls Who Sabotaged 'Ghostbusters', May 2018. https://www.forbes.com/sites/scottmendelson/2018/05/17/box-office-oceans-8-sandra-bullock-cate-blanchett-anne-hathaway-ghostbusters-rihanna/?sh=76cfd561596d.

[44] Christopher Mims. Why social media is so good at polarizing us, October 2020. https://www.wsj.com/articles/why-social-media-is-so-good-at-polarizing-us-11603105204.

[45] Michel Minoux. Accelerated Greedy Algorithms for Maximizing Submodular Set Functions. In J. Stoer, editor, *Optimization Techniques*, pages 234–243, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg.

[46] Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, and Amin Karbasi. Fast Constrained Submodular Maximization: Personalized Data Summarization. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1358–1367, New York, New York, USA, 20–22 Jun 2016. PMLR.

[47] Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrak, and Andreas Krause. Lazier Than Lazy Greedy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015.

[48] Baharan Mirzasoleiman, Amin Karbasi, Rik Sarkar, and Andreas Krause. Distributed Submodular Maximization. *The Journal of Machine Learning Research*, 17(1):8330–8373, 2016.

[49] Houssam Nassif, Kemal Oral Cansizlar, Mitchell Goodman, and S. V. N. Vishwanathan. Diversifying Music Recommendations. *CoRR*, abs/1810.01482, 2018.

[50] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An Analysis of Approximations for Maximizing Submodular Set Functions - I. *Mathematical Programming*, 14(1):265–294, 1978.

[51] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press, 2018.

[52] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers Big Data*, 2:13, 2019.

[53] Gourab K. Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. Fair Ranking: A Critical Review, Challenges, and Future Directions. In *FAccT*, page To appear. ACM, 2022.

[54] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. Fairness in Rankings and Recommendations: An Overview. *The VLDB Journal*, 2021.

[55] Soumyajit Saha. Spotify Adopts Indian Habits to Avoid the 'Netflix Problem', May 2022. https://the-ken.com/story/spotify-adopts-indian-habits-to-avoid-the-netflix-problem/.
[56] Ashudeep Singh and Thorsten Joachims. Fairness of Exposure in Rankings. In *KDD*, pages 2219–2228. ACM, 2018.
[57] Michael Tauberg. Power Law in Popular Media, June 2018. https://michaeltauberg.medium.com/power-law-in-popular-media-7d7efef3fb7c.
[58] Choon Hui Teo, Houssam Nassif, Daniel Hill, Sriram Srinivasan, Mitchell Goodman, Vijai Mohan, and S.V.N. Vishwanathan. Adaptive, Personalized Diversity for Visual Discovery. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 35–38, New York, NY, USA, 2016. Association for Computing Machinery.
[59] Sahil Verma and Julia Rubin. Fairness Definitions Explained. In Yuriy Brun, Brittany Johnson, and Alexandra Meliou, editors, *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, pages 1–7, Gothenburg, Sweden, 2018. ACM.
[60] Christine Wennerås and Agnes Wold. Nepotism and Sexism in Peer-Review. *Nature*, 387(6631):341–343, May 1997.
[61] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. Balanced ranking with diversity constraints. In *IJCAI*, pages 6035–6042. ijcai.org, 2019.
[62] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. Causal intersectionality and fair ranking. In *FORC*, volume 192 of *LIPIcs*, pages 7:1–7:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.
[63] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems*, 30, 2017.
[64] Meike Zehlike and Carlos Castillo. Reducing disparate exposure in ranking: A learning to rank approach. In *WWW*, pages 2849–2855. ACM / IW3C2, 2020.
[65] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in Ranking, Part I: Score-Based Ranking. *ACM Comput. Surv.*, apr 2022. Just Accepted.

**Anonymized Supplementary:** Anonymized supplementary with full proofs of the theoretical results and additional simulation plots is available at this link https://anonymous.4open.science/r/Submodular-Maximization-in-the-Presence-of-Biases-53A3

**Code:** The anonymized code for all simulations is available at https://github.com/submodular-bias/code

# A ADDITIONAL EXAMPLES OF SUBMODULAR FUNCTIONS USED BY PRIOR WORKS

The diminishing returns property, of submodular functions, arises in many applications, including content recommendation [10, 49], web search [2], text summarization [37], and team selection [28, 30, 34]. This property is one of the key reasons for the use of submodular functions in the above applications [36].

For instance, submodular functions are used when a recommendation system has different objectives and each additional item satisfying the same objective has a diminishing return. Concretely, [10] explain that "a fundamental requirement of [Spotify's] music recommender system is its ability to accommodate considerations from the users (e.g. short-term satisfaction objectives), artists (e.g. exposure of emerging artists) and platform (e.g. facilitating discovery and boosting strategic content), when surfacing music content to users" [10] design the recommendation system for Spotify that uses a submodular objective function. For each item $i$ (e.g., song or podcast), let $W_{i1} \geq 0$ denote its relevance to a user (predicted by a learning algorithm) and $W_{i2}, \ldots, W_{im} \in \{0, 1\}$ indicate artist and platform specific metrics (e.g., if $i$ is created by an emerging artist or if the platform wants to promote $i$). [10] use the following objective to capture the utility of a playlist $S$ for the user

$$F(S) \coloneqq \sum_{i \in S} W_{i1} + \sqrt{\sum_{i \in S} W_{i2}} + \cdots + \sqrt{\sum_{i \in S} W_{im}}. \quad (6)$$

Submodular functions also arise in web search, where each query $q$ can have multiple interpretations: For instance, the query "flash" can refer to the Adobe Flash player, the superhero "The Flash", or the village Flash with the highest elevation in Great Britain. Irrespective of the intended interpretation, each additional result related to the same interpretation offers a smaller marginal utility to the user [2]. Suppose a query $q$ (e.g., "Flash") belongs to category $j \in [m]$ (e.g., technology, movies, or location) with probability $\Pr[j \mid q]$ and, conditioned on the event that $q$ belongs to category $j$, result $i$ satisfies the user with probability $\Pr[i \mid j, q]$ (independent of other items). [2] observe that search results $S$ that maximize the following submodular objective have a higher quality than search results of a commercial search engine.

$$F(S) \coloneqq \sum_{j=1}^{m} \Pr[j \mid q] \left(1 - \prod_{i \in S} (1 - \Pr[i \mid j, q])\right). \quad (7)$$

This is in the family $\mathscr{F}$. To see this set, $W_{ij} = \log \frac{1}{1-\Pr[i|j,q]}$ and $g_j(x) = -\Pr[j \mid q] \cdot e^{-x}$ for all $i$ and $1 \leq j \leq m$, and $g_{j+1}(x) = 1$ for all $x$.

Given weights $w_1, \ldots, w_{m-1}$ and scores $s_1, \ldots, s_n$, [58] consider the following submodular function $F \in \mathscr{F}$ defined by (1) $g_j(x) = w_j \cdot \log(1 + x)$ for all $1 \leq j \leq m - 1$ and $g_m(x) = x$, (2) for each $1 \leq j \leq m - 1$, $W_{ij} = 1$ if $i$ has attribute $j$ and $W_{ij} = 0$ otherwise, and (3) $W_{im} = s_i$.

# B OVERLAPPING GROUPS AND NEED FOR STOCHASTICITY IN GROUPS

Consider two overlapping groups $G_1, G_2$, these divide the items into four disjoint "intersections": $I_a \coloneqq G_1 \cap G_2$, $I_b \coloneqq G_1 \backslash G_2$, $I_c \coloneqq G_2 \backslash G_1$, and $I_d \coloneqq [n] \backslash (G_1 \cup G_2)$. Similarly, $p$ overlapping groups $G_1, G_2, \ldots, G_p$ divide the items into up to $2^p$ intersections. [12] extend the bias model of [35] to overlapping groups. In their model, each item $i$ in a different intersection faces a different amount of bias: The observed utility of $i$ is $\prod_{\ell : G_\ell \ni i} \beta_\ell$ times smaller than its latent utility.

Remark B.1 (**Overlapping groups**). *Instead of $p$ overlapping groups $G_1, \ldots, G_p$, one can consider the groups as the intersections formed by $G_1, \ldots, G_p$ For example, with $p = 2$, one can consider $I_a$, $I_b$, $I_c$, and $I_d$ as the groups with bias functions $\phi_a = \phi_1 \circ \phi_2$, $\phi_b = \phi_1$,$\phi_c = \phi_2$, and $\phi_d(x) = x$ respectively. Since there are at most $\min(n, 2^p)$ non-empty intersections, this does not increase the running time of the algorithm proposed in the paper (Theorem 4.3). For simplicity, in this paper we assume that the groups are disjoint.*

*The need of stochasticity in groups.* In the model, we assume that the latent utilities are deterministically chosen and the protected groups are stochastic. This generalizes the model of [12, 35, 40] which is equivalent to the model that draws latent utilities i.i.d. from some distribution and also constructs protected groups stochastically. A further generalization, could consider the case where both the latent utilities and protected groups are stochastic. However, in this model it is information theoretically impossible to output a subset whose latent utility is to guaranteed to be at least a positive fraction of OPT.

Formally, we can show the following result if both latent utilities and protected groups can be arbitrary then no algorithm can recover any constant factor of approximation of the optimal utility. Let $S_A$ be the subset output by algorithm $A$ when given $\widehat{W}$ and $G_1, \ldots, G_p$ as input. For any $\varepsilon > 0$ and any algorithm $A$, there are two disjoint protected groups $G_1$ and $G_2$ and bias functions $\phi_1, \phi_2 : \mathbb{R} \to \mathbb{R}$ such that $F(S_A) \leq \varepsilon$.

# C IMPLEMENTATION DETAILS OF SIMULATIONS

*Code.* The anonymized code for all simulations is available at https://github.com/submodular-bias/code.

*Synthetic dataset 1.* Figure 2(a) presents results with this data. This data has $m \coloneqq 3$ attributes and $n \coloneqq 250$ songs, and uses the objective $F(S) \coloneqq \sum_{i \in S} W_{i1} + \lambda \sum_{j=2}^{3} \sqrt{\sum_{i \in S} W_{ij}}$ (with $\lambda = \frac{1}{20}$).

- First, we select a subset $S_E$ of songs with size $|S_E| = 0.8n$ and label all songs in $S_E$ to be from an emerging artist and all other songs as songs from non-emerging artists. This implies that $W_{i2} = 1$ if $i \in S_E$ and $W_{i2} = 0$ otherwise.
- Next, for each song $i \in [n]$, with probability $p_{\text{NH}} \coloneqq 0.9$, we label it as "`not heard by the current user`" and set $W_{i3} = 1$ and, otherwise, we label it as "`heard by the current user`" and set $W_{i3} = 0$.
- Finally, for each song $i \notin S_E$, we independently draw a value $X$ from the power-law distribution with exponent $\delta$ and set $W_{i1} = X \cdot 1000$.[5] For each $i \in S_E$, we independently draw a value $X$

from the power-law distribution with exponent $\delta$ *conditioned on* $X \leq 2$ and set $W_{i1} = X \cdot 1000$. The conditioning encodes the fact that emerging artists do not have any "popular" song yet.

We fix $\lambda := \frac{1}{20}$, $|S_E| = 0.8n$, and $p_{\text{NH}} = 0.9$ and vary $\beta \in [0, 1]$, $\delta \in \{1, 1.5, 2, 2.5, 3\}$, and $\frac{|G_1|}{n} \in \{0.25, 0.5, 0.75\}$.

Remark C.1. *We also repeated the simulation with* $|S_E| \in \{0.4n, 0.9n\}$, $p_{\text{NH}} \in \{0.4, 1.0\}$, *and* $\lambda \in \left\{\frac{1}{10}, \frac{1}{5}\right\}$, *and observed similar results as Figures 1, 3 and 4.*

*Synthetic dataset 2.* The second synthetic dataset corresponds Figure 2(b), has $m := 3$ attributes and $n := 250$ songs, and uses the objective function $F(S) := \sum_{j=1}^{3} \log(1 + \sum_{i \in S} W_{ij})$.

- We generate sets $C_1, C_2, C_3$ uniformly at random: For each item $i$, with probability $\frac{1}{3}$ we assign it to $C_1$, otherwise with probability $\frac{1}{3}$ we assign it to $C_2$, and otherwise we assign it to $C_3$.
- For each $h \in [3]$ and $i \in C_h$, we independently draw a value $X$ from the power-law distribution with exponent $\delta$ and set $W_{i1} = X \cdot 1000$.[5]

Like the first synthetic dataset, we generate groups $G_1$ and $G_2$ by assigning $|G_1|$ items chosen uniformly at random without replacement to $G_1$ and the remaining items to $G_2$ Given $\beta \in [0, 1]$, we generate observed utilities $\widehat{W}$ as in Definition 3.2 with $\phi_1(x) = x$ and $\phi_2(x) = \beta x$. The simulation on this data fixed varies $\beta \in [0, 1]$, $\delta \in \{1, 2, 3\}$, and $\frac{|G_1|}{n} \in \{0.25, 0.5, 0.75\}$.

# D FURTHER DISCUSSION OF RELATED WORKS

Recent works [12, 18, 35, 40] have demonstrated the benefit of imposing fairness constraints on the output of subset selection on the latent utility of the output when the objective is additive or *linear*. [35] introduce the mathematical model of bias mentioned in Section 1. They consider two groups, with $G_2$ being the disadvantaged group, and study conditions on $\beta$, group sizes, $k$, and the distribution of $W$, where requiring the output $S$ to satisfy $|S \cap G_2| \geq 1$ increases the latent utility of the output. [12] study a generalization of subset selection, ranking, where the selected individuals also need to be ordered. Specializing their work to subset selection: They consider two groups $G_1, G_2$ and show that if $U$ encodes proportional representation and $W$ is drawn i.i.d. from the uniform distribution on $[0, 1]$, then $F(S_U) \geq (1 - o_k(1)) \cdot \text{OPT}$.[6] [18] study selection under a different model of bias, where the observed utility has higher than average noise for individuals in one group. They show give a family of fairness constraints (including proportional representation) which increases the output's latent utility. [40] study subset selection under the same model as [12], but with multiple and overlapping groups. They compare the efficacy of the fairness constraints applied on groups $G_1, \ldots, G_p$ to fairness constraints applied on the disjoint intersections formed by these groups. Most relevant to this work, they extend the result of [12] for proportional representation constraints to multiple disjoint groups $G_1, \ldots, G_p$. Unlike all of these works, we study the efficacy of fairness constraints when the objective of subset selection is submodular.

[5] This is natural as powerlaw distributions have been observed to arise in the performance of musicians [57] and in the performance of other creative professionals [13].

[6] To be precise they required the output to have at least $L_\ell$ items from group $G_\ell$. For two groups, this is equivalent to requiring the output subset to have at most $U_\ell = k - L_{\neg\ell}$ items from $G_\ell$.

# E STANDARD ALGORITHMS FOR SUBMODULAR MAXIMIZATION

In this section, we present standard greedy algorithm by [50] for maximizing monotone submodular functions with, e.g., cardinality constraints.

**Algorithm 2** The standard greedy algorithm ([50])

**Input:** An evaluation oracle for $F$, a number $k$, and a set of feasible sets $I \subseteq 2^{[n]}$
**Output:** A subset $S \in I$ with $|S| \leq k$

```
Initialize S = ∅
while |S| < k do
    Let i* be the item in [n] that maximizes F(S ∪ {i*}) − F(S) subject to S ∪ {i*} ∈ I
    Set S = S ∪ {i*}
end while
return S
```

**Algorithm 3** A variant of the standard greedy algorithm that satisfies given upper bound constraints

**Input:** An evaluation oracle for $F$, a numbers $k, U_1, \ldots, U_p$, and groups $G_1, \ldots, G_p$
**Output:** A subset $S$ with $|S| \leq k$ and $|S \cap G_\ell| \leq U_\ell$

```
Initialize S = ∅
while |S| < k do
    Let i* be the item in [n] that maximizes F(S ∪ {i*}) − F(S) subject to |(S ∪ {i*}) ∩ G_ℓ| ≤ U_ℓ (for each ℓ)
    Set S = S ∪ {i*}
end while
return S
```

## Contents

# F PROOFS

## F.1 Proof of Theorem 4.1

In this section, we show that a family of fairness constraints is insufficient to guarantee any constant fraction of the optimal latent utility. This family of fairness constraints is parameterized by $2p$ parameters $u_1, u_2, \ldots, u_p \geq 0$ and $v_1, v_2, \ldots, v_p \geq 0$. Given $u_1, \ldots, u_p$ and $v_1, v_2, \ldots, v_p$, the constraints require any output subset $S$ to satisfy

$$\forall \ell \in [p], \quad |S \cap G_\ell| \leq (u_\ell + v_\ell \gamma_\ell) \cdot k. \tag{8}$$

Where for each $\ell \in [p]$

$$\gamma_\ell \coloneqq \frac{|G_\ell|}{n}.$$

This family captures equal representation when $u_\ell = \frac{1}{p}$ and $v_\ell = 0$ (for each $\ell \in [p]$), proportional representation when $u_\ell = 0$ and $v_\ell = 1$ (for each $\ell \in [p]$), and it ensures that each output subset satisfies the the 4/5ths rule when $u_\ell = 0$ and $\frac{\min_\ell v_\ell}{\max_\ell v_\ell} \geq \frac{4}{5}$. Given $U \coloneqq (u_1, \ldots, u_p)$ and $V \coloneqq (v_1, \ldots, v_p)$, let $S_{UV} \subseteq [n]$ be the subset of size at most $k$ that maximizes observed utility subject to satisfying the constraint specified by $(U, V)$, i.e.,

$$S_{UV} \coloneqq \underset{|S| \leq k\colon S \text{ satisfies Equation (8)}}{\operatorname{argmax}} \widehat{F}(S).$$

In this section, we prove the following theorem.

Theorem F.1. *For any $0 < \varepsilon < 1$, $U = (u_1, u_2)$, and $V = (v_1, v_2)$ there exists*

- *a submodular function $F \in \mathscr{F}$,*
- *$0 \leq \gamma_1, \gamma_2 \leq 1$,*
- *$0 < \beta_1, \beta_2 \leq 1$, and*
- *family of $n \times 3$ matrices $\mathcal{W}$ parameterized by $n$,*

*such that, for any $k \geq \frac{1}{\varepsilon^{13}} \log \frac{e}{\varepsilon}$, $n \geq \frac{4k}{\varepsilon^4} \log \frac{e}{\varepsilon}$, and $W = \mathcal{W}(n)$, it holds that*

$$\Pr\left[F(S_{UV}) \leq \varepsilon \cdot OPT\right] \geq 1 - \varepsilon,$$

*where the probability is over the randomness in $G_1$ and $G_2$.*

The above theorem straightforwardly extends to $p > 2$ groups by adding empty groups. It also extends to $m > 2$ attributes by fixing utilities so that $W_{ij} = 0$ for each item $i$ and $j \in [m] \setminus \{1, 2\}$. We divide the proof into cases depending on the value of $u_1$ and $v_1$.

*F.1.1 Case A: ($v_1 > \varepsilon^{-1}$ and $u_1 = 0$).* Set the following parameters.

- For any subset $S$ define $F(S) = \sum_{i \in S} W_{i1}$.
- Let $\gamma_1 = \varepsilon$ and $\gamma_2 = 1 - \varepsilon$.
- Let $\beta_1 = 1$ and $\beta_2 = \varepsilon^2$.
- Given $n$, let $W = \mathcal{W}(n)$ be the following matrix: $W_{i1} = 1$ for each $1 \leq i \leq k$ and $W_{i1} = \varepsilon$ otherwise. ($W_{ij} = 0$ for any $j \neq 1$ and $i \in [n]$.)

Let $\mathscr{E}$ be the event that $|G_1 \cap [k]| \leq 2k\varepsilon$. By the construction of $G_1$, $|G_1 \cap [k]|$ is a hypergeometric random variable. Using the standard concentration inequality ([29, Theorem 1]) for hypergeometric random variables, it follows that

$$\Pr\left[\left||G_1 \cap [k]| - \gamma_1 k\right| \geq \sqrt{\frac{k}{2} \log \frac{1}{\varepsilon}}\right] \leq \varepsilon \quad \overset{k\varepsilon \geq \sqrt{\frac{k}{2}\log 1/\varepsilon}}{\Longrightarrow} \quad \Pr\left[|G_1 \cap [k]| \geq 2k\varepsilon\right] \geq 1 - \varepsilon.$$

Thus, $\Pr[\mathscr{E}] \geq 1 - \varepsilon$. By the construction of $W$, $\beta_1$, and $\beta_2$, the following holds

$$\forall i \in [G_1], \forall j \in [G_2], \quad \widehat{W}_{i1} \geq \widehat{W}_{j1}.$$

This implies that $|S_{UV} \cap G_1| = k$ and $|S_{UV} \cap G_2| = 0$. If not, then we can swap any item in $S_{UV} \cap G_2$ with an item from $G_1 \setminus S_{UV}$ to increase $S_{UV}$'s observed utility. $G_1 \setminus S_{UV}$ is non empty as $|G_1| = n\varepsilon > k \geq |S_{UV}|$. The resulting set satisfies the constraints as in this case $u_1 + v_1\gamma_1 \geq k$. Conditioned on $\mathscr{E}$, we have

$$\begin{aligned}
F(S_{UV}) &= \sum\nolimits_{i \in S_{UV} \cap G_1} W_{i1} && \text{(Since } |S_{UV} \cap G_2| = 0\text{)} \\
&\leq \sum\nolimits_{i \in G_1 \cap [k]} W_{i1} + \sum\nolimits_{i \in (S_{UV} \cap G_1) \setminus [k]} W_{i1} \\
&\leq 2k\varepsilon + k\varepsilon && \text{(Using that conditioned on } \mathscr{E}\text{, } |G_1 \cap [k]| \leq 2k\varepsilon\text{, } |S_{UV}| \leq k\text{, and the value of } W\text{)} \\
&\leq 3k\varepsilon.
\end{aligned}$$

From construction, OPT $= k$. Thus, it holds that $\Pr\left[F(S_{UV}) \leq 3\varepsilon \cdot \text{OPT}\right] \geq \Pr\left[F(S_{UV}) \leq 3\varepsilon \cdot \text{OPT} \mid \mathscr{E}\right] \Pr[\mathscr{E}] \geq 1(1 - \varepsilon)$. The claim follows by scaling down $\varepsilon$ by a factor of 3 in the construction.

*F.1.2 Case B: ($v_1 < \varepsilon$ and $u_1 = 0$).* Set the following parameters.

- For any subset $S$ define $F(S) = \sum_{i\in S} W_{i1}$.
- Let $\gamma_1 = 1-\varepsilon$ and $\gamma_2 = \varepsilon$.
- Let $\beta_1 = 1$ and $\beta_2 = \varepsilon^2$.
- Given $n$, let $W = \mathcal{W}(n)$ be the following matrix: $W_{i1} = 1$ for each $1 \le i \le k$ and $W_{i1} = \varepsilon$ otherwise. ($W_{ij} = 0$ for any $j \ne 1$ and $i \in [n]$.)

Let $\mathscr{E}$ be the event that $|G_2 \cap [k]| \le 2k\varepsilon$. By replacing $G_1$ and $\gamma_1$ by $G_2$ and $\gamma_2$ in the calculation of $\Pr[\mathscr{E}]$ in Case A, it follows that $\Pr[\mathscr{E}] \ge 1-\varepsilon$. Since $u_1 + v_1\gamma_1 < \varepsilon$ in this case, it holds that $|S_{UV} \cap G_1| \le k\varepsilon$. Conditioned on $\mathscr{E}$, we have

$$\begin{aligned}
F(S_{UV}) &= \sum_{i\in S_{UV}\cap G_1} W_{i1} + \sum_{i\in S_{UV}\cap G_2\cap[k]} W_{i1} + \sum_{i\in (S_{UV}\cap G_2)\setminus[k]} W_{i1} \\
&\le k\varepsilon + \sum_{i\in S_{UV}\cap G_2\cap[k]} W_{i1} + \sum_{i\in (S_{UV}\cap G_2)\setminus[k]} W_{i1} && \text{(Using that } W_{ij}\le 1 \text{ and } |S_{UV}\cap G_1|\le k\varepsilon\text{)} \\
&\le 3k\varepsilon + \sum_{i\in (S_{UV}\cap G_2)\setminus[k]} W_{i1} && \text{(Using that conditioned on } \mathscr{E}, |G_2\cap[k]|\le 2k\varepsilon \text{ and } W_{ij}\le 1\text{)} \\
&\le 4k\varepsilon. && \text{(Using that for any } i\notin[k], W_{i1}=\varepsilon \text{ and } |S_{UV}|\le k\text{)}
\end{aligned}$$

From construction, $\mathrm{OPT} = k$. Thus, it holds that $\Pr[F(S_{UV}) \le 4\varepsilon\cdot\mathrm{OPT}] \ge \Pr[F(S_{UV}) \le 4\varepsilon\cdot\mathrm{OPT} \mid \mathscr{E}]\Pr[\mathscr{E}] \ge 1(1-\varepsilon)$. The claim follows by scaling down $\varepsilon$ by a factor of 4 in the construction.

*F.1.3 Case C: ($\varepsilon < v_1 < \varepsilon^{-1}$ and $u_1 = 0$).* Set the following parameters.

- For any subset $S$ define
$$F(S) = \left(\sum_{i\in S} W_{i1}\right)^{1/3} + \varepsilon\sqrt{\sum_{i\in S} W_{i2}} \tag{9}$$
- Let $\gamma_1 = \varepsilon^4$ and $\gamma_2 = 1-\varepsilon^4$.
- Let $\beta_1 = 1$ and $\beta_2 = \beta$.
- We divide items into two types. Let $A := [n/2]$ be the set of type $A$ items and $B := [n]\setminus A$ be the set of type $B$ items. Given $n$, let $W = \mathcal{W}(n)$ be the following matrix:
  - for each type $A$ item $i \in A$, $W_i = (1,0)$, and
  - for each type $B$ item $i \in B$, $W_i = (0,1)$.

Let $\mathscr{E}$ be the event that (1) $A$ has at least $k$ items from each of $G_1$ and $G_2$ and (2) $B$ has at least $k$ items from each of $G_1$ and $G_2$. By the construction of $G_1$ and $G_2$, $|A\cap G_\ell|$ and $|B\cap G_\ell|$ are hypergeometric random variables (for each $\ell\in[2]$). Using the standard concentration inequality ([29, Theorem 1]) for hypergeometric random variables, it follows that for each $\ell\in[p]$

$$\Pr\left[\left||A\cap G_\ell| - \gamma_1|A|\right| \ge \sqrt{\frac{|A|}{2}\log\frac{1}{\varepsilon}}\right] \le \varepsilon \overset{\gamma_1,\gamma_2\ge\varepsilon^4,\ |A|=n/2}{\Longrightarrow} \Pr\left[|A\cap G_1| \ge \frac{\varepsilon^4 n}{2} - \sqrt{\frac{n}{4}\log\frac{1}{\varepsilon}}\right] \ge 1-\varepsilon. \tag{10}$$

Using that $\frac{\varepsilon^4 n}{2} - \sqrt{\frac{n}{4}\log\frac{1}{\varepsilon}}$ is an increasing function of $n$ for $n \ge \frac{1}{4\varepsilon^8}\log\frac{1}{\varepsilon}$ and by construction $n \ge \frac{1}{\varepsilon^8}\log\frac{1}{\varepsilon}$, it follows that

$$\frac{\varepsilon^4 n}{2} - \sqrt{\frac{n}{4}\log\frac{1}{\varepsilon}} \overset{n\ge\frac{4k}{\varepsilon^4}\log\frac{e}{\varepsilon}}{\ge} \left(2k - \varepsilon^{-2}\sqrt{k}\right)\cdot\log\frac{1}{\varepsilon} \ge k\cdot\log\frac{1}{\varepsilon}. \qquad \text{(Using that } k\ge\varepsilon^{-4} \text{ by construction)}$$

Substituting this in Equation (10), it follows that for each $\ell\in[p]$, $\Pr[|A\cap G_\ell|\ge k] \ge 1-\varepsilon$. Replacing $A$ by $B$ in the above argument, $\Pr[|B\cap G_\ell|\ge k] \ge 1-\varepsilon$. Thus, by union bound $\Pr[\mathscr{E}] \ge 1-4\varepsilon$.

We claim that conditioned on $\mathscr{E}$

$$|S_{UV}\cap G_2\cap A| \le \varepsilon^3 k. \tag{11}$$

Suppose that the above claim is true. Conditioned on $\mathscr{E}$ we have that

$$\begin{aligned}
F(S_{UV}) &= \left(|S_{UV}\cap G_1\cap A| + |S_{UV}\cap G_2\cap A|\right)^{1/3} + \varepsilon\sqrt{|S_{UV}\cap G_1\cap B| + |S_{UV}\cap G_1\cap B|} && \text{(Using Equation (9) and construction of } W\text{)} \\
&= \left(|S_{UV}\cap G_1\cap A| + \varepsilon^3 k\right)^{1/3} + \varepsilon\sqrt{|S_{UV}\cap G_1\cap B| + |S_{UV}\cap G_1\cap B|} && \text{(Using Equation (11))} \\
&= \left(2\varepsilon^3 k\right)^{1/3} + \varepsilon\sqrt{k} && \text{(Using that } S_{UV} \text{ satisfies the constraints, hence, } |S_{UV}\cap G_1\cap A| \le (u_1+\gamma_1 v_1)k = \varepsilon^3 k \text{ and } |S_{UV}|\le k\text{)} \\
&\le 3\varepsilon k. && \text{(Using Equation (9))}
\end{aligned}$$

By construction, $\mathrm{OPT} \ge \sqrt{k}$. Thus, $\Pr[F(S_{UV}) \le 3\varepsilon\cdot\mathrm{OPT}] \ge \Pr[F(S_{UV}) \le 3\varepsilon\cdot\mathrm{OPT} \mid \mathscr{E}]\Pr[\mathscr{E}] \ge 1(1-4\varepsilon)$. The theorem's claim follows by scaling down $\varepsilon$ by a factor of 4 in the construction.

It remains to prove the claim above (Equation (11)). Towards a contradiction assume that

$$|S_{UV}\cap G_2\cap A| > \varepsilon^3 k. \tag{12}$$

Consider a subset $S$ that is the same as $S_{UV}$ except that compared to $S_{UV}$ it selects one less item from $G_2 \cap A$ and one more item from $G_2 \cap B$. $S$ exists because, conditioned on $\mathscr{E}$, $|G_2 \cap B| \geq k$. $S$ satisfies the constraints specified by $(U, V)$ as $|S \cap G_\ell| = |S_{UV} \cap G_\ell|$ for each $\ell \in [2]$ and $S_{UV}$ satisfies the constraints specified by $(U, V)$. It holds that

$$\begin{aligned}\widehat{F}(S) - \widehat{F}(S_{UV}) = &\left(|S_{UV} \cap A \cap G_1| + \beta|S_{UV} \cap A \cap G_2| - \beta\right)^{1/3} + \varepsilon\sqrt{|S_{UV} \cap B \cap G_1| + \beta|S_{UV} \cap B \cap G_2| + \beta} \\ &- \left(\left(|S_{UV} \cap A \cap G_1| + \beta|S_{UV} \cap A \cap G_2|\right)^{1/3} + \varepsilon \cdot \sqrt{|S_{UV} \cap B \cap G_1| + \beta|S_{UV} \cap B \cap G_2|}\right)\end{aligned}$$

The RHS of the above equation is a decreasing function of $|S_{UV} \cap B \cap G_1| + \beta|S_{UV} \cap B \cap G_2|$. Since $S_{UV}$ satisfies the constraints specified by $(U, V)$, $|S_{UV} \cap B \cap G_1| + \beta|S_{UV} \cap B \cap G_2| \leq (\varepsilon^3 + \beta)k = 2\beta k$. Consequently

$$\begin{aligned}\widehat{F}(S) - \widehat{F}(S_{UV}) \geq &\left(|S_{UV} \cap A \cap G_1| + \beta|S_{UV} \cap A \cap G_2| - \beta\right)^{1/3} + \varepsilon\left(\sqrt{2\beta k + \beta} - \sqrt{2\beta k}\right) \\ &- \left(|S_{UV} \cap A \cap G_1| + \beta|S_{UV} \cap A \cap G_2|\right)^{1/3}\end{aligned}$$

RHS of the above inequality is an increasing function of $|S_{UV} \cap A \cap G_2|$ and, from Equation (12), $|S_{UV} \cap A \cap G_2| > \varepsilon^3 k$. Hence,

$$\widehat{F}(S) - \widehat{F}(S_{UV}) \geq \left(|S_{UV} \cap A \cap G_1| + \beta\varepsilon^3 k - \beta\right)^{1/3} + \varepsilon\left(\sqrt{2\beta k + \beta} - \sqrt{2\beta k}\right) - \left(|S_{UV} \cap A \cap G_1| + \beta\varepsilon^3 k\right)^{1/3}$$

RHS of the above inequality is an increasing function of $|S_{UV} \cap A \cap G_1|$. In this case, $(u_1 + v_1\gamma_1) \geq \varepsilon\gamma_1 k = \varepsilon^5 k$. We claim this implies that $|S_{UV} \cap A \cap G_1| \geq \varepsilon^5 k$. If this is not true, then one can increase the observed utility of $S_{UV}$ by removing one item in $S_{UV}$ from $A \cap G_2$ and adding one item in $S_{UV}$ from $A \cap G_1$. This is possible as (1) in this case, $|S_{UV} \cap A \cap G_2| > 0$ and (2) conditioned on $\mathscr{E}$, $|A \cap G_1| \geq k$. Combining $|S_{UV} \cap A \cap G_1| \geq \varepsilon^5 k$ with previous observation, it follows that

$$\begin{aligned}\widehat{F}(S) - \widehat{F}(S_{UV}) &= \left((\varepsilon^5 + \beta\varepsilon^3)k - \beta\right)^{1/3} - \left((\varepsilon^5 + \beta\varepsilon^3)k\right)^{1/3} + \varepsilon\left(\sqrt{2\beta k + \beta} - \sqrt{2\beta k}\right) \\ &\geq -\frac{\beta}{\left(\varepsilon^3\beta k\right)^{2/3}} + \varepsilon\left(\sqrt{2\beta k + \beta} - \sqrt{2\beta k}\right) && \text{(Using that for all } 0 \leq x \leq 1, (1-x)^{1/3} \geq 1 - x \text{ and } \tfrac{\beta}{\varepsilon^5 k + \beta\varepsilon^3 k} \leq 1) \\ &\geq -\frac{\beta}{\left(\varepsilon^3\beta k\right)^{2/3}} + \varepsilon\frac{\beta}{\sqrt{6\beta k}} && \text{(Using that for all } 0 \leq x \leq 3, \sqrt{1+x} \geq 1 + \tfrac{x}{3} \text{ and } k \geq 1) \\ &= \frac{\beta}{\sqrt{k}}\left(\sqrt{\frac{\varepsilon}{6}} - \frac{1}{\varepsilon^{5/3}k^{1/6}}\right) && \text{(Using Equation (9) and that } \beta = \varepsilon) \\ &> 0. && \text{(Using that } k > 216\varepsilon^{-13}, \beta > 0, \text{ and } k \geq 1)\end{aligned}$$

## F.2 Proof of Theorem 4.3

In this section, we prove Theorem 4.3. For the reader's convenience, we restate Theorem 4.3 and the assumption used in Theorem 4.3, below.

Assumption 2 (**Disjoint attributes**). *The matrix $W \in \mathbb{R}^{n\times m}$ is such that $C_1, \ldots, C_m$ are disjoint.*

Theorem F.2. *Suppose $\phi_1(x) = x$ for each $x$. There is an algorithm (Algorithm 1) that, given observed utilities $\widehat{W} \in \mathbb{R}^{n\times m}$ and evaluation oracles for $g_1, \ldots, g_m : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, outputs a set $S$ of size $k$ with the following property (under Assumption 1): For any $\varepsilon, \tau, \gamma > 0$ and any increasing functions $\phi_2, \ldots, \phi_p : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, there is a large enough $k_0$ such that for any $k \geq k_0$, with probability at least $1 - \varepsilon$*

$$F(S) \geq OPT \cdot (1 - \varepsilon).$$

*Where the probability is over the randomness in the choice of $G_1, \ldots, G_p$. The algorithm makes $O(nk)$ evaluations of each $g_1, \ldots, g_m$ and does $O(nk \log n)$ additional arithmetic operations.*

*F.2.1 Overview of the proof of Theorem 4.3.* We have to prove that the subset $\widetilde{S}$ output by Algorithm 1 has latent utility at least $\text{OPT} \cdot (1 - \widetilde{O}(k^{-1/4}))$ with high probability, provided that Assumption 1 holds. One can show that because of Assumption 1, $S_{\text{OPT}} \coloneqq \text{argmax}_{S\subseteq[n]:|S|\leq k} F(S)$ is of the form

$$S_{\text{OPT}} \coloneqq \bigcup_{j=1}^{m} S_j^\star,$$

where for each $j$,

$$S_j^\star \coloneqq \underset{S\subseteq C_j : |S| \leq |S_{\text{OPT}} \cap C_j|}{\text{argmax}} g_j\left(\sum_{i\in S} W_{ij}\right).$$

Here, the function in the RHS $H(T) \coloneqq g_j\left(\sum_{i\in T} W_{ij}\right)$ is a function in $\mathscr{F}$ with one attribute. For such functions, there is an algorithm which, given *observed* utilities as input, outputs a subset $S$ with *latent* utility $F(S)$ at least $\text{OPT} \cdot \left(1 - k^{-1/2}\right)$ with high probability (Lemma F.7). Hence, we can use Lemma F.7 to compute (an approximation to) $S_j^\star$ and, hence, (an approximation to) $S_{\text{OPT}}$. The catch is that the definition of $S_j^\star$, itself, depends upon $S_{\text{OPT}}$ because of the $|S_{\text{OPT}} \cap C_j|$ term in its definition.

Our key idea is to estimate $\left|S_{\mathrm{OPT}} \cap G_j\right|$ without actually computing $S_{\mathrm{OPT}}$. This uses the stochasticity in the protected groups $G_1, \ldots, G_p$ and that $\phi_1(x) = x$ for all $x$. In particular, given a function $F \in \mathscr{F}$, $F(T) \coloneqq \sum_{j=1}^{m} g_j\left(\sum_{i \in S} W_{ij}\right)$, we define the following alternate version of $F$:

$$\forall T \subseteq [n], \quad \widetilde{F}(T) \coloneqq \sum\nolimits_{j=1}^{m} g_j\left(\frac{n}{|G_1|} \cdot \sum\nolimits_{i \in T \cap G_1} W_{ij}\right). \tag{13}$$

There are two differences between $\widetilde{F}$ and $F$: First, given a set $T$ as input $\widetilde{F}$ only sums the utilities over $T \cap G_1$. Second it scales the computed sums by a factor of $\frac{n}{|G_1|}$. Intuitively, because $G_1$ is constructed by sampling elements uniformly at random without replacement, for any subset $T$ independent of $G_1$, the two changes cancel each other: For any subset $T$ independent of $G_1$, under some mild assumptions on $T$, we prove that $F(T) \in \widetilde{F}(T) \cdot \left(1 \pm k^{-1/4}\right)$ (Lemma F.4).

Algorithm 1 computes a subset $\widetilde{S}$ which selects $k \cdot \frac{|G_1|}{n}$ items from $|G_1|$ and maximizes $\widetilde{F}$. Algorithm 1 is able to find such a subset because of Assumption 1 (see Lemma F.3). We use $k_j \coloneqq \frac{n}{|G_1|} \cdot \left|\widetilde{S} \cap C_j\right|$ as an approximation of $|S_{\mathrm{OPT}} \cap C_j|$ for each $j \in [m]$. Finally, using Lemma F.4 and some additional analysis (Lemma F.5), we show that the subset $S$ defined as,

$$S \coloneqq \bigcup\nolimits_{j=1}^{m} S_j, \qquad \text{where} \qquad \forall j, \quad S_j \coloneqq \operatorname*{argmax}_{T \subseteq C_j : |T| \leq k_j} g_j\left(\sum\nolimits_{i \in S} W_{ij}\right), \tag{14}$$

has latent utility at least $\mathrm{OPT} \cdot (1 - \widetilde{O}(k^{-1/8}))$ (Lemma F.6). This analysis has to be careful to ensure that the assumptions needed by Lemma F.4 are satisfied for all considered subsets. For example, $S_{\mathrm{OPT}}$ may not satisfy these assumptions and we construct a subset $\overline{S_{\mathrm{OPT}}}$ which is "similar" to $S_{\mathrm{OPT}}$ and which satisfies these assumptions (Lemma F.5). Moreover, $\widetilde{S}$ is not independent of $G_1$ as required by Lemma F.4. To bypass this, we show that $\widetilde{S}$ always belongs to a family with at most $k^m$ subsets; and show that Lemma F.4 holds for all subsets in $(k+1)^m$ with high probability and, hence, in particular it holds for $\widetilde{S}$.

### F.2.2 Proof of Theorem 4.3.

Lemma F.3. *Suppose Assumption 1 holds and $\phi_1(x) = x$ for all $x$. For any $W \in \mathbb{R}_{\geq 0}^{m \times n}$, $F \in \mathscr{F}$ and the corresponding function $\widetilde{F}$ (defined in Equation (13)), increasing functions $\phi_2, \ldots, \phi_p$, $k \geq 1$, and protected groups $G_1, \ldots, G_p$, the subset $\widetilde{S}$ computed by Algorithm 1 is an optimal solution to*

$$\begin{aligned} \max_{S \subseteq G_1} \quad & \widetilde{F}(S), \\ \text{s.t.,} \quad & 1 \leq j \leq m, \quad \left|S \cap C_j\right| \geq \min\{\sqrt{k}, \left|C_j\right|\}, \\ & |S| \leq k \cdot \frac{|G_1|}{n}. \end{aligned} \tag{15}$$

Proof. In this proof, we only consider subsets of $G_1$. To simplify notation, we write $C_j$ to denote $C_j \cap G_1$ for each $1 \leq j \leq m$. Under Assumption 1, for each item $i$, there is a unique $1 \leq j(i) \leq m$ such that $W_{ij} \neq 0$. Define $w_i \coloneqq W_{ij(i)}$ for each $1 \leq i \leq n$. For each $j$, let $i(1, j), i(2, j), \ldots, i(\left|C_j\right|, j)$ be items in $C_j$ arranged in non-increasing order of $w_i$.

***Local optimality.*** A subset $S$ is said to be *locally-optimal* if it has the property that: For each $j$, if $S$ selects $r$ elements from $C_j$ then these are the $r$ elements with the highest utility in $C_j$, i.e.,

$$S \cap C_j = \left\{i(r, j) \mid 1 \leq r \leq \left|S \cap C_j\right|\right\}.$$

Let $S^\star$ be an optimal solution of Equation (15). Without loss of generality, $S^\star$ is locally optimal.

***Without loss of generality $S^\star$ is locally optimal.*** If $S^\star$ is not locally-optimal, construct another set $S$ such that $\left|S \cap C_j\right| = \left|S^\star \cap C_j\right|$ and $S$ is locally-optimal. Hence, $\sum_{i \in S \cap C_j} w_i \geq \sum_{i \in S^\star \cap C_j} w_i$ for each $1 \leq j \leq m$. Consequently, as $g_1, \ldots, g_m$ are increasing, $F(S) \geq F(S^\star)$. Because $\left|S \cap C_j\right| = \left|S^\star \cap C_j\right|$ for each $j$, it also follows that $S$ is feasible for Equation (15). Since $S^\star$ is an optimal solution of Equation (15), it must hold that $F(S) = F(S^\star)$. Hence, $S$ is optimal solution of Equation (15) and locally optimal.

One can verify that Algorithm 1 selects items in decreasing order of $w_i$ and, hence, $\widetilde{S}$ is also locally-optimal. Suppose $F(\widetilde{S}) < F(S^\star)$ and, hence, $\widetilde{S} \neq S^\star$. Since $\widetilde{S}$ and $S^\star$ are not equal and both $\widetilde{S}$ and $S^\star$ are locally optimal hence, there exist two attributes $j(1)$ and $j(2)$ such that $|\widetilde{S} \cap C_{j(1)}| > |S^\star \cap C_{j(1)}|$ and $|\widetilde{S} \cap C_{j(2)}| < |S^\star \cap C_{j(2)}|$. Without loss of generality assume that $j(1) = 1$ and $j(2) = 2$. Let

$$x \coloneqq |S^\star \cap C_1| \quad \text{and} \quad y \coloneqq |S^\star \cap C_2|.$$

Since $S^\star$ is feasible for the Equation (15), $x, y \geq \sqrt{k}$. Consider the iteration (of the while loop in Step 7 of Algorithm 1) where $\widetilde{S}$ selects the $(x+1)$-th item from $C_1$, i.e., $i(x+1, 1)$. (This iteration exists because $x > \sqrt{k}$ and, hence, $(x+1)$-th item is not selected in Step 5 of Algorithm 1 and $|\widetilde{S} \cap C_1| > x$, so $(x+1)$-th item is selected in some iteration of Step 7 of Algorithm 1.) Let $\bar{S}$ be the value of $\widetilde{S}$ at this iteration and $z \coloneqq |\bar{S} \cap C_2| \leq |\widetilde{S} \cap C_2| < y$. Since Algorithm 1, selected $i^\star \coloneqq i(x+1, 1)$ instead of $i(z+1, 2)$, it must hold that

$$\widetilde{F}_{i(x+1,1)}(\bar{S}) \geq F_{i(z,2)}(\bar{S}).$$

Equivalently,

$$g_1\left(w_{i(x+1,1)}+\sum_{h=1}^{x} w_{i(h,1)}\right)-g_1\left(\sum_{h=1}^{x} w_{i(h,1)}\right)\geq g_2\left(w_{i(z+1,1)}+\sum_{h=1}^{z} w_{i(h,2)}\right)-g_2\left(\sum_{h=1}^{z} w_{i(h,2)}\right). \tag{16}$$

If the above holds with equality, then Algorithm 1 could have chosen $i(z+1,2)$ instead of $i(x+1,1)$ in this iteration. If Equation (16) holds with equality, suppose Algorithm 1 selects $i(z+1,2)$ instead of $i(x+1,1)$ in this iteration, and find the next iteration where Algorithm 1 selects an item not in $S^\star$. (This will only repeat a finite number of times as part 2 of Algorithm 1 lasts for finite number of iterations.) If there is no such iteration, we have shown $F(\widetilde{S})=F(S^\star)$.

It remains to consider the case where Equation (16) holds with strict inequality, i.e.,

$$g_1\left(w_{i(x+1,1)}+\sum_{h=1}^{x} w_{i(h,1)}\right)-g_1\left(\sum_{h=1}^{x} w_{i(h,1)}\right)> g_2\left(w_{i(z+1,1)}+\sum_{h=1}^{z} w_{i(h,2)}\right)-g_2\left(\sum_{h=1}^{z} w_{i(h,2)}\right). \tag{17}$$

Consider the subset $S' \coloneqq S^\star \cup i(x+1,2)\backslash i(y+1,1)$. We can lower bound $F(S')$ as follows

$$\begin{aligned}
F(S')-F(S^\star) &= \left(g_1\left(w_{i(x+1,1)}+\sum_{h=1}^{x} w_{i(h,1)}\right)-g_1\left(\sum_{h=1}^{x} w_{i(h,1)}\right)\right)-\left(g_2\left(w_{i(y+1,1)}+\sum_{h=1}^{y} w_{i(h,2)}\right)-g_2\left(\sum_{h=1}^{y} w_{i(h,2)}\right)\right)\\
&\geq \left(g_1\left(w_{i(x+1,1)}+\sum_{h=1}^{x} w_{i(h,1)}\right)-g_1\left(\sum_{h=1}^{x} w_{i(h,1)}\right)\right)-\left(g_2\left(w_{i(z+1,1)}+\sum_{h=1}^{z} w_{i(h,2)}\right)-g_2\left(\sum_{h=1}^{z} w_{i(h,2)}\right)\right)\\
&\qquad \text{(Using that } z\leq y \text{ and the } g_2 \text{ is concave)}\\
&> 0. \qquad \text{(Using Equation (17))} \qquad (18)
\end{aligned}$$

Further, $S'$ is feasible for Equation (15). We only need to verify that $S'$ satisfies the constrains on the first two attributes. These hold because $|S'\cap C_1|\geq|\widetilde{S}\cap C_1|\geq L_1$ and $|S'\cap C_2|\geq\left|S^\star\cap C_2\right|\geq L_2$. Since $S^\star$ is optimal for Equation (15), $S'$ is feasible, and $F(S')\geq F(S^\star)$, it must be true that $F(S')\leq F(S^\star)$. But Equation (18) contradicts this. Hence, $F(\widetilde{S})=F(S^\star)$ and, since we already showed that $\widetilde{S}$ is feasible for Equation (15), $\widetilde{S}$ is optimal for Equation (15) and the lemma follows. □

Next, Lemma F.4 shows that, under some assumptions on a subset $T\subseteq[n]$ (Equation (19)), $\widetilde{F}(T)$ is a good approximation of $F(T)$.

Lemma F.4. *For any subset $S$ of size $k$ (independent of $G_1$), numbers $x,\delta>0$, attribute $1\leq j\leq m$, and increasing and concave function $g_j:\mathbb{R}\to\mathbb{R}$, if*

$$\sum_{i\in S\cap C_j} W_{ij}\geq xk, \tag{19}$$

*then for any $\delta\geq\exp\left(-\tau\gamma^6 xk\right)$, with probability at least $1-\delta$, it holds that*

$$\left|g_j\left(\frac{n}{|G_1|}\cdot\sum_{i\in S\cap G_1\cap C_j} W_{ij}\right)-g_j\left(\sum_{i\in S\cap C_j} W_{ij}\right)\right|\leq\sqrt{\frac{1}{\tau\gamma^6 xk}\log\frac{1}{\delta}}\cdot g_j\left(\sum_{i\in S\cap C_j} W_{ij}\right).$$

*Where the probability is over the randomness in the choice of $G_1$.*

Taking a union bound over $1\leq j\leq m$ and using the definition of $\widetilde{F}$ (Equation (13)), it follows that: with probability at least $1-m\delta$ (for any $\delta\geq e^{-\tau\gamma^6 xk}$), it holds that

$$\left|\widetilde{F}(S)-F(S)\right|\leq\sqrt{\frac{1}{\tau\gamma^6 xk}\log\frac{1}{\delta}}\cdot F(S).$$

Proof of Lemma F.4. Fix any $j\in[m]$. For each $i\in[n]$, let $Z_i$ be the discrete random variable that is $W_{ij}$ if $i\in G_1$ and 0 otherwise. We have that $\Pr[Z_i]=\frac{|G_1|}{n}$ and $\sum_{i\in S\cap C_j} Z_i=\sum_{i\in S\cap C_j\cap G_1} W_{ij}$. Using linearity of expectation, it follows that

$$\mu\coloneqq\mathbb{E}\left[\sum_{i\in S\cap C_j} Z_i\right]=\frac{|G_1|}{n}\cdot\sum_{i\in S\cap C_j} W_{ij}.$$

Using a variant of the Hoeffding's inequality which holds for sampling without replacement [3, Proposition 1.2], we get that for any $\delta>0$

$$\Pr\left[\left|\sum_{i\in S\cap C_j} Z_i-\mu\right|>\sqrt{\frac{1}{2}\cdot\left(\sum_{i\in S\cap C_j} W_{ij}^2\right)\cdot\log\frac{1}{\delta}}\right]\leq\delta. \tag{20}$$

Let $\mathscr{E}$ be the event that

$$\left|\sum_{i\in S\cap C_j} Z_i-\mu\right|\leq\sqrt{\frac{1}{2}\cdot\left(\sum_{i\in S\cap C_j} W_{ij}^2\right)\cdot\log\frac{1}{\delta}}.$$

Since $0 \le W_{ij} \le \tau^{-1}$, it follows that $W_{ij}^2 \le W_{ij}/\tau$. Substituting this in the above inequality, we get that, conditioned on $\mathscr{E}$

$$\left|\sum_{i\in S\cap C_j} Z_i - \mu\right| \le \sqrt{\frac{1}{2\tau}\cdot\left(\sum_{i\in S\cap C_j} W_{ij}\right)\cdot\log\frac{1}{\delta}}.$$

Next, using that $\sum_{i\in S\cap C_j} W_{ij} \ge xk$, we get that conditioned on $\mathscr{E}$

$$\frac{\left|\sum_{i\in S\cap C_j} Z_i - \mu\right|}{\mu} \le \frac{n}{|G_1|}\sqrt{(2\tau xk)^{-1}\cdot\log\frac{1}{\delta}} \le \sqrt{(2\tau\gamma^2 xk)^{-1}\cdot\log\frac{1}{\delta}}. \qquad \text{(Using that } |G_1| \ge \gamma n\text{)}$$

Since $g_j$ is an increasing function and $\frac{|G_1|}{n} \ge \gamma$, the above equation implies: Conditioned on $\mathscr{E}$

$$g_j\left(\frac{n\mu}{|G_1|} - \sqrt{\frac{1}{2\tau\gamma^4 xk}\log\frac{1}{\delta}}\right) \le g_j\left(\frac{n}{|G_1|}\sum_{i\in S\cap C_j} Z_i\right) \le g_j\left(\frac{n\mu}{|G_1|} + \sqrt{\frac{1}{2\tau\gamma^4 xk}\log\frac{1}{\delta}}\right). \tag{21}$$

Let

$$\alpha := g_j'\left(\frac{n\mu}{|G_1|}\left(1 - \sqrt{\frac{1}{2\tau\gamma^4 xk}\log\frac{1}{\delta}}\right)\right). \tag{22}$$

Since $g_j$ is concave $g'(y) \le g_j'(z)$ for any $y \ge z$. Hence, from Equation (21), we get that conditioned on $\mathscr{E}$

$$\left|g_j\left(\frac{n}{|G_1|}\sum_{i\in S\cap C_j} Z_i\right) - g_j\left(\frac{n\mu}{|G_1|}\right)\right| \le \alpha\frac{n\mu}{|G_1|}\sqrt{\frac{1}{2\tau\gamma^4 xk}\log\frac{1}{\delta}}. \tag{23}$$

Let

$$\zeta := \sqrt{\frac{1}{2\tau\gamma^6 xk}\log\frac{1}{\delta}}. \tag{24}$$

Observe that

$$\begin{aligned}
g_j\left(\frac{n\mu}{|G_1|}\right) &\ge g_j(\mu) && \text{(Using that } n \ge |G_1| \text{ and that } g_j \text{ is an increasing function)}\\
&= \int_0^{\mu} g_j'(z)dz\\
&\ge \int_0^{\mu(1-\zeta)} g_j'(z)dz && \text{(Using that } \zeta \ge 0 \text{ and } g'(z) \ge 0 \text{ for all } z \ge 0\text{)}\\
&\ge g_j'(\mu(1-\zeta))\cdot\mu(1-\zeta) && \text{(Using that } g_j' \text{ is a decreasing function, as } g_j \text{ is concave)}\\
&= \alpha\mu(1-\zeta) && \text{(Using Equations (22) and (24))}\\
&\ge \alpha\gamma(1-\zeta)\frac{\mu n}{|G_1|}.
\end{aligned}$$

Substituting this in Equation (23), it follows that conditioned on $\mathscr{E}$

$$\left|g_j\left(\frac{n}{|G_1|}\sum_{i\in S\cap C_j} Z_i\right) - g_j\left(\frac{n\mu}{|G_1|}\right)\right| \le \frac{\zeta}{(1-\zeta)}g\left(\frac{n\mu}{|G_1|}\right) \le 2\zeta g\left(\frac{n\mu}{|G_1|}\right). \qquad \text{(Using that } \zeta \le \tfrac{1}{2} \text{ as } \delta \ge e^{-\tau\gamma^6 xk}, \mu \ge x, \text{ and Equation (24))}$$

Substituting the value of $\mu$ and $\zeta$, we get that conditioned on $\mathscr{E}$

$$\left|g_j\left(\frac{n}{|G_1|}\sum_{i\in S\cap C_j} Z_i\right) - g_j\left(\sum_{i\in S\cap C_j} W_{ij}\right)\right| \le g_j\left(\sum_{i\in S\cap C_j} W_{ij}\right)\cdot\sqrt{\frac{2}{\tau xk\gamma^6}\cdot\log\frac{1}{\delta}}.$$

The lemma follows because $\Pr[\mathscr{E}] \ge 1-\delta$ by Equation (20). □

Next, Lemma F.5 lower bounds $\widetilde{F}(\widetilde{S})$ by $\mathrm{OPT}\cdot\left(1 - O(k^{-1/4})\right)$.

Lemma F.5. *Suppose Assumption 1 holds and $\phi_1(x) = x$ for all $x$. Consider $\widetilde{S}$ computed in Part 1 of Algorithm 1. For any $\delta \ge me^{-8\tau\gamma k}$, with probability $1-\delta$ it holds that*

$$\widetilde{F}(\widetilde{S}) \ge OPT\cdot\left(1 - O\left(\sqrt{\frac{m^4}{\tau^2\gamma^5 k^{1/2}}\log\frac{m}{\delta}}\right)\right).$$

*Where the probability is over the randomness in the choice of protected groups $G_1, G_2, \ldots, G_p$.*

Since $\widetilde{S}$ is an optimal solution for Equation (15), it holds that for any other subset $T$ feasible for Equation (15), $\widetilde{F}(\widetilde{S}) \geq \widetilde{F}(T)$. In particular, if $S_{\text{OPT}}$ is feasible for Equation (15), then $\widetilde{F}(\widetilde{S}) \geq \widetilde{F}(S_{\text{OPT}})$. If $S_{\text{OPT}}$ satisfies the condition in Equation (19) for any constant $x > 0$, the from Lemma F.4 we would get that $\widetilde{F}(S_{\text{OPT}}) \geq \text{OPT}\left(1 - O(k^{-3/2})\right)$. Then the proof would follow from chaining the last two inequalities. However, $S_{\text{OPT}}$ may not be feasible for Equation (15) or satisfy Equation (19) and, hence, we cannot use Lemma F.4.

Instead, we construct a subset $\overline{S_{\text{OPT}}}$ from OPT and show that $\overline{S_{\text{OPT}}}$ satisfies Equation (19) for $x = \Omega(k^{-1/2})$ and $F(\overline{S_{\text{OPT}}}) \geq F(\text{OPT})$. We (for our analysis) construct a subset $\overline{S_{\text{OPT}}}$ from OPT as follows:

**Algorithm to construct $\overline{S_{\text{OPT}}}$ (which is used in the proof of Lemma F.5)**

- **Input:** $S_{\text{OPT}}$, latent utilities $W$, and $\delta > 0$
- **Output:** $\overline{S_{\text{OPT}}}$

(1) Initialize $\overline{S_{\text{OPT}}} = S_{\text{OPT}}$
(2) **For** $j \in [m]$ **do**
  - **If** $\left|\overline{S_{\text{OPT}}} \cap C_j\right| \leq 2\gamma^{-1}k^{1/2}$ **then**
    - Add any $2\gamma^{-1}k^{1/2}$ elements from $C_j \setminus \overline{S_{\text{OPT}}}$ to $\overline{S_{\text{OPT}}}$
(3) Set $j^\star := \text{argmax}_j \left|\overline{S_{\text{OPT}}} \cap C_j\right|$ ▷ By Pigeon hole $\left|\overline{S_{\text{OPT}}} \cap C_{j^\star}\right| \geq \frac{k}{m} \geq \Omega((m+1)k^{1/2})$
(4) Remove $\left|\overline{S_{\text{OPT}}}\right| - k$ elements from $\overline{S_{\text{OPT}}} \cap C_{j^\star}$ with the smallest latent utilities ▷ Here, $0 \leq |\overline{S_{\text{OPT}}}| - k \leq m(2\gamma^{-1}k^{1/2})$
(5) **return** $\overline{S_{\text{OPT}}}$

Proof of Lemma F.5. Set $k_0$ to satisfy the following inequality

$$k_0 \geq \max\left\{2\gamma^{-1}m(m+1), \left(\gamma \log \frac{m}{\delta}\right)^3, 16\gamma^{-2}m^4\right\}. \tag{25}$$

Fix any $1 \leq j \leq m$.

$\overline{S_{\text{OPT}}}$ is a function of just $S_{\text{OPT}}$, $W$, and $\delta$, and, $S_{\text{OPT}}$ itself is just a function of $W$. h Hence, $\overline{S_{\text{OPT}}}$ is independent of $G_1$. Repeating the concentration argument in Lemma F.4, but with $Z_i := \mathbb{I}[i \in G_1]$, we get that for any $\delta > 0$

$$\Pr\left[\left|\left|\overline{S_{\text{OPT}}} \cap C_j \cap G_1\right| - \frac{|G_1|}{n} \cdot \left|\overline{S_{\text{OPT}}} \cap C_j\right|\right| > \sqrt{\frac{1}{2} \cdot k^{1/2} \cdot \log \frac{m}{\delta}}\right] \leq \frac{\delta}{m}.$$

Consequently

$$\Pr\left[\left|\overline{S_{\text{OPT}}} \cap C_j \cap G_1\right| > \frac{|G_1|}{n} \cdot \left|\overline{S_{\text{OPT}}} \cap C_j\right| - \sqrt{\frac{1}{2} \cdot k^{1/2} \cdot \log \frac{m}{\delta}}\right] \leq \frac{\delta}{m}$$

$$\xrightarrow{\left|\overline{S_{\text{OPT}}} \cap C_j\right| \geq 2\gamma^{-1}k^{1/2} \text{ and } |G_1| \geq \gamma n} \Pr\left[\left|\overline{S_{\text{OPT}}} \cap C_j \cap G_1\right| > 2k^{1/2} - \sqrt{\frac{1}{2} \cdot k^{1/2} \cdot \log \frac{m}{\delta}}\right] \leq \frac{\delta}{m}.$$

Since $k \geq k_0$ and $k_0$ satisfies Equation (25), the above inequality implies

$$\Pr\left[\left|\overline{S_{\text{OPT}}} \cap C_j \cap G_1\right| > k^{1/2}\right] \leq \frac{\delta}{m}. \tag{26}$$

From the Hoeffding's bound, we also have that

$$\Pr\left[\left|\overline{S_{\text{OPT}}} \cap G_1\right| \leq k\frac{|G_1|}{n} + \sqrt{\frac{1}{2} \cdot k\frac{|G_1|}{n} \cdot \log \frac{m}{\delta}}\right] \leq \frac{\delta}{m}. \tag{27}$$

Let $\mathscr{E}$ be the event that

$$\left|\overline{S_{\text{OPT}}} \cap G_1\right| \leq k\frac{|G_1|}{n} + \sqrt{k \log \frac{m}{\delta}} \quad \text{and} \quad \left|\overline{S_{\text{OPT}}} \cap C_j \cap G_1\right| > k^{1/2}.$$

From Equations (26) and (27), $\Pr[\mathscr{E}] \geq 1 - 2\delta$. Therefore, conditioned on $\mathscr{E}$, $\overline{S_{\text{OPT}}} \cap G_1$ is feasible for Equation (15). Since $\widetilde{S}$ is an optimal solution for Equation (15), it follows that conditioned on $\mathscr{E}$:

$$\widetilde{F}(\widetilde{S}) \geq \widetilde{F}(\overline{S_{\text{OPT}}}). \tag{28}$$

Further, as for each $j$, $\left|\overline{S_{\text{OPT}}} \cap C_j\right| \geq 2\gamma^{-1}\sqrt{k}$ and $W_{ij} \geq \tau$, it follows that $\sum_{i \in \overline{S_{\text{OPT}}} \cap C_j} W_{ij} \geq 2\tau\sqrt{k}$. Hence, $\overline{S_{\text{OPT}}}$ satisfies Equation (19) with $x \geq 2\tau\gamma^{-1}(k)^{-1/2}$. Using Lemma F.4 (as $\overline{S_{\text{OPT}}}$ is independent of $G_1$), it follows that with probability at least $1-\delta$ (for any $\delta \geq me^{-\tau^2\gamma^5\sqrt{k}}$)

$$\widetilde{F}(\overline{S_{\text{OPT}}}) \geq F(\overline{S_{\text{OPT}}}) \left(1 - \sqrt{\frac{1}{2\tau^2\gamma^5\sqrt{k}} \log \frac{1}{\delta}}\right). \tag{29}$$

Finally, we lower bound $F(\overline{S_{\text{OPT}}})$ by a multiple of $F(\text{OPT})$. First observe that for all $j \neq j^\star$, $\overline{S_{\text{OPT}}} \cap C_j \supseteq \text{OPT} \cap C_j$ and, hence,

$$\forall j \neq j^\star, \quad g_j\left(\sum_{i \in \overline{S_{\text{OPT}}}} W_{ij}\right) \geq g_j\left(\sum_{i \in S_{\text{OPT}}} W_{ij}\right). \tag{30}$$

Further, by construction of $\overline{S_{\text{OPT}}}$, $(\overline{S_{\text{OPT}}} \cap C_{j^\star}) \subseteq (S_{\text{OPT}} \cap C_{j^\star})$ and $(S_{\text{OPT}} \cap C_{j^\star}) \setminus (\overline{S_{\text{OPT}}} \cap C_{j^\star})$ consists of $\alpha$ items with the smallest latent utility in $S_{\text{OPT}} \cap C_{j^\star}$ where $\alpha \leq 2m\gamma^{-1}k^{1/2}$. Let $\Delta := \left|S_{\text{OPT}} \cap C_{j^\star}\right|$. Let $i(1), i(2), \ldots, i(\Delta)$ be items in $S_{\text{OPT}} \cap C_{j^\star}$ such that $W_{i(1)j^\star} \geq W_{i(2)j^\star} \geq \cdots$. We have that

$$\begin{aligned}
\frac{\sum_{i \in \overline{S_{\text{OPT}}}} W_{ij^\star}}{\sum_{i \in S_{\text{OPT}}} W_{ij^\star}} &= \frac{\sum_{i \in \overline{S_{\text{OPT}}} \cap C_{j^\star}} W_{ij^\star}}{\sum_{i \in S_{\text{OPT}} \cap C_{j^\star}} W_{ij^\star}} && \text{(Using Assumption 1)} \\
&= \frac{\sum_{h=1}^{|\overline{S_{\text{OPT}}} \cap C_{j^\star}|} W_{i(h)j^\star}}{\sum_{h=1}^{\Delta} W_{i(h)j^\star}}. && \text{(Using } S_{\text{OPT}} \cap C_{j^\star} = \{i(1), \ldots, i(\Delta)\}\text{)}
\end{aligned}$$

By construction, $\overline{S_{\text{OPT}}}$ drops at most $\alpha$ items from OPT and if it drop $x$ items, then these are the $x$ items with the smallest latent utilities in $S_{\text{OPT}} \cap C_{j^\star}$. Consequently

$$\begin{aligned}
\frac{\sum_{i \in \overline{S_{\text{OPT}}}} W_{ij^\star}}{\sum_{i \in S_{\text{OPT}}} W_{ij^\star}} &\geq \frac{\sum_{h=1}^{\Delta-\alpha} W_{i(h)j^\star}}{\sum_{h=1}^{\Delta} W_{i(h)j^\star}} \\
&\geq \left(1 + \frac{\alpha W_{i(\Delta-\alpha)j^\star}}{\sum_{h=1}^{\Delta-\alpha} W_{i(h)j^\star}}\right)^{-1} \\
&\geq \left(1 + \frac{\alpha}{\Delta - \alpha}\right)^{-1} && \text{(Using that } W_{i(\Delta-\alpha)j^\star} \leq W_{i(h)j^\star} \text{ for all } \alpha \leq j \leq \Delta - \alpha\text{)} \\
&\geq \left(1 + \frac{2m\gamma^{-1}\sqrt{k}}{\frac{k}{2m} - 2m\gamma^{-1}\sqrt{k}}\right)^{-1} && \text{(Using that } \alpha \leq 2m\gamma^{-1}k^{1/2} \text{ and by Pigeon hole principal } \Delta \geq \tfrac{k}{m}\text{)} \\
&= 1 - \frac{4m^2}{\gamma\sqrt{k}}. && (31)
\end{aligned}$$

Since $g_{j^\star}$ is concave and increasing, we have that

$$\begin{aligned}
g_{j^\star}\left(\sum_{i \in \overline{S_{\text{OPT}}}} W_{ij^\star}\right) &\overset{(31)}{\geq} g_{j^\star}\left(\left(1 - 4\gamma^{-1}m^2k^{-1/2}\right) \cdot \sum_{i \in S_{\text{OPT}}} W_{ij^\star}\right) \\
&\geq \left(1 - 4\gamma^{-1}m^2k^{-1/2}\right) \cdot g_{j^\star}\left(\sum_{i \in \overline{S_{\text{OPT}}}} W_{ij^\star}\right). && \text{(Using that } g_{j^\star} \text{ is concave and } g(0) \geq 0\text{)} \quad (32)
\end{aligned}$$

Hence, combining Equations (30) and (32) it follows that

$$F(\overline{S_{\text{OPT}}}) \geq F(S_{\text{OPT}}) \cdot \left(1 - 4\gamma^{-1}m^2k^{-1/2}\right). \tag{33}$$

We get the required result by chaining Equations (28), (29), and (33) and taking the union bound over events in Equations (28) and (29). □

If $\widetilde{S}$ is independent of $G_1$, then Lemmas F.4 and F.5 show that with high probability there is a subset $S_E$ such that $\widetilde{S} = S_E \cap G_1$ and $F(S_E)$ is at least

$$\text{OPT} \cdot (1 - O(k^{-1/4})).$$

However, $\widetilde{S}$ is not independent of $G_1$. Lemma F.6 addresses this.

Lemma F.6. *Suppose $C_1, \ldots, C_m$ are disjoint and $\phi_1(x) = x$ for all $x$. Let $\widetilde{S}$ be as constructed in Part 1 of Algorithm 1 and $k_1, \ldots, k_j$ be as defined in Step 11 of Algorithm 1. For any $\delta > 0$, with probability at least $1-\delta$, there exists a subset $S_E \subseteq [n]$ satisfying*

$$\forall j \in [m], \quad \left|S_E \cap C_j\right| \leq k_j \quad \textit{and} \quad F(S_E) \geq OPT \cdot \left(1 - O\left(\sqrt{\frac{m^4}{\tau^2\gamma^5k^{1/2}} \log \frac{m}{\delta}}\right)\right). \tag{34}$$

Proof. The proof relies on the fact that $\widetilde{S} \subseteq G_1$ is locally optimal. Recall that $S \subseteq G_1$ is said to be locally optimal if for each $j$, $S \cap C_j$ is the set of $\left|S \cap C_j\right|$ items with the highest latent utility in $C_j \cap G_1$. Extend the definition of local optimality to sets which are not necessarily a subset of $G_1$: A subset $S \subseteq [n]$ is said to be locally optimal if for each $j \in [m]$, $S \cap C_j$ contains $\left|S \cap C_j\right|$ items with the highest latent utility in $C_j$ (instead of $C_j \cap G_1$). A locally optimal subset $S$ is uniquely defined by the following values

$$x_{j,S} \coloneqq \left|S \cap C_j\right| \quad \text{for each } 1 \le j \le m.$$

Let $\mathcal{S}$ be the set of all locally optimal subsets $S$ of size at most $k$ satisfying $x_{j,S} \ge 2\gamma^{-1}\sqrt{k}$ for each $1 \le j \le m$. Since each $S \in \mathcal{S}$ is uniquely identified by $\left\{x_{j,S}\right\}_j$ and there are $k+1$ choices for each $x_{j,S}$, it follows that $|\mathcal{S}| \le (k+1)^m$. Since $W_{ij} > \tau$ (for each $i$ and $j$),

$$\forall S \in \mathcal{S}, \quad \sum\nolimits_{i \in S \cap C_j} \ge 2\tau\gamma^{-1}\sqrt{k}.$$

Hence, each $S \in \mathcal{S}$ satisfies Equation (19) with $x = 2\tau\gamma^{-1}k^{-1/2}$. $\mathcal{S}$ is deterministically given $W$ and, hence, is independent of $G_1$. Consequently, Lemma F.4 is applicable for any $S \in \mathcal{S}$. Applying Lemma F.4 for each $S \in \mathcal{S}$ and taking the union bound, implies: For any $\delta \ge m \cdot (k+1)^m \cdot e^{-\tau^2\gamma^5\sqrt{k}}$, with probability at least $1-\delta$, for all $T \in \mathcal{S}$

$$\left|\widetilde{F}(T) - F(T)\right| \le \sqrt{\frac{1}{2\tau^2\gamma^5 k^{1/2}} \log \frac{m \cdot (k+1)^m}{\delta}} \cdot F(T).$$

In other words, for any $\delta \ge e^{-\Omega(\tau^2\gamma^5\sqrt{k})}$, it holds that with probability at least $1-\delta$, for all $T \in \mathcal{S}$

$$\left|\widetilde{F}(T) - F(T)\right| \le \sqrt{\frac{m}{\tau^2\gamma^5 k^{1/2}} \log \frac{mk}{\delta}} \cdot F(T).$$

Select the subset $T \in \mathcal{S}$ such that for each $1 \le j \le m$

$$\left|T \cap C_j\right| = x_{j,T} \ge \min\left\{\frac{n}{|G_1|} \cdot |\widetilde{S} \cap C_j|, \left|C_j\right|\right\}.$$

Such a subset exists because $\sum_{j=1}^{m} \frac{n}{|G_1|} \cdot |\widetilde{S} \cap C_j| = \frac{n}{|G_1|}|\widetilde{S}| = k$. Next, divide $1 \le j \le m$ into two cases.

**Case A ($|T \cap C_j| = |C_j|$):** It must be true that $T \cap C_j = C_j$ and, hence, $(\widetilde{S} \cap C_j) \subseteq (T \cap C_j)$. As $\widetilde{S} \subseteq G_1$, this implies that $(\widetilde{S} \cap C_j) \subseteq (T \cap C_j \cap G_1)$. Consequently

$$g_j\left(\frac{n}{|G_1|}\sum\nolimits_{i \in T \cap C_j \cap G_1} W_{ij}\right) \ge g_j\left(\frac{n}{|G_1|}\sum\nolimits_{i \in \widetilde{S} \cap C_j} W_{ij}\right). \tag{35}$$

**Case B ($|T \cap C_j| \ge \frac{n}{|G_1|} \cdot |\widetilde{S} \cap C_j|$):** In this case, we will $(T \cap C_j)$ contains at least $1 - O(k^{-1/4})$ fraction of the items in $(\widetilde{S} \cap C_j)$. To see this note that

$$\mathbb{E}_{G_1}\left[\left|T \cap C_j \cap G_1\right|\right] = \frac{|G_1|}{n}\left|T \cap C_j\right| \ge |\widetilde{S} \cap C_j|.$$

Applying the Chernoff bound to the following indicator random variables $\left\{\mathbb{I}[i \in G_1] : i \in T \cap C_j\right\}$, we get that for any $\delta > 0$

$$\Pr\left[\left|T \cap C_j \cap G_1\right| \ge |\widetilde{S} \cap C_j| - \sqrt{3 \cdot |\widetilde{S} \cap C_j| \cdot \log \frac{m}{\delta}}\right] \le \frac{\delta}{m}.$$

Because $|\widetilde{S} \cap C_j| \ge \sqrt{k}$, this implies that

$$\Pr\left[\left|T \cap C_j \cap G_1\right| \ge |\widetilde{S} \cap C_j|\left(1 - \sqrt{3 \cdot k^{-1/2} \cdot \log \frac{m}{\delta}}\right)\right] \le \frac{\delta}{m}. \tag{36}$$

Let the event in the above equation be $\mathscr{E}$. Let $\Delta := \left|C_j \cap G_1\right|$, $\phi := \left|\widetilde{S} \cap C_j\right|$, and let $i(1), i(2), \ldots, i(\Delta)$ be the elements of $C_j \cap G_1$ ordered in decreasing order of latent utility, i.e., $W_{i(1)j} \geq W_{i(2)j} \geq \cdots \geq W_{i(\Delta)j}$. Then we have, conditioned on $\mathscr{E}$

$$\begin{aligned}
\frac{\sum_{i\in T\cap C_j\cap G_1} W_{ij}}{\sum_{i\in \widetilde{S}\cap C_j} W_{ij}} &= \frac{\sum_{h=1}^{|T\cap C_j\cap G_1|} W_{i(h)j}}{\sum_{h=1}^{\phi} W_{i(h)j}} && \text{(Using that } \widetilde{S}, T \text{ are locally optimal and } \phi := |\widetilde{S}\cap C_j|\text{)} \\
&\geq \frac{\sum_{h=1}^{\phi\cdot\left(1-O\left(k^{-1/4}\sqrt{\log m/\delta}\right)\right)} W_{i(h)j}}{\sum_{h=1}^{\phi} W_{i(h)j}} && \text{(Using that } \mathscr{E} \text{ is the event in Equation (36))} \\
&= \left(1+\frac{\sum_{h=\phi\cdot\left(1-O\left(k^{-1/4}\sqrt{\log m/\delta}\right)\right)+1}^{\phi} W_{i(h)j}}{\sum_{h=1}^{\phi\cdot\left(1-O\left(k^{-1/4}\sqrt{\log m/\delta}\right)\right)} W_{i(h)j}}\right)^{-1} \\
&\geq \left(1+\frac{O\left(k^{-1/4}\sqrt{\log m/\delta}\right)}{1-O\left(k^{-1/4}\sqrt{\log m/\delta}\right)}\right)^{-1} && \text{(Using } W_{i(1)j} \geq W_{i(2)j} \geq \ldots \geq W_{i(\Delta)j} \text{ and } W_{ij} \geq 0\text{)} \\
&\geq 1 - O\left(k^{-1/4}\sqrt{\log m/\delta}\right).
\end{aligned}$$

Consequently, conditioned on $\mathscr{E}$,

$$\begin{aligned}
g_j\left(\frac{n}{|G_1|}\sum_{i\in T\cap C_j\cap G_1} W_{ij}\right) &\geq g_j\left(\frac{n}{|G_1|}\cdot\left(1-O\left(k^{-1/4}\sqrt{\log m/\delta}\right)\right)\cdot\sum_{i\in\widetilde{S}\cap C_j} W_{ij}\right) && (g_j \text{ is increasing}) \\
&\geq \left(1-O\left(k^{-1/4}\sqrt{\log m/\delta}\right)\right)\cdot g_j\left(\frac{n}{|G_1|}\cdot\sum_{i\in\widetilde{S}\cap C_j} W_{ij}\right). && \text{(Using that } g_j \text{ is concave and } g(0)\geq 0\text{)} \quad (37)
\end{aligned}$$

Taking the union bound over all $j$ where $\left|T\cap C_j\right| \neq \left|C_j\right|$, we get that with probability at least $1-\delta$, for all $1 \leq j \leq m$ Equation (37) holds. Combining this with Equation (35), we get that with probability at least $1-\delta$,

$$\sum_{j=1}^{m} g_j\left(\frac{n}{|G_1|}\sum_{i\in T\cap C_j\cap G_1} W_{ij}\right) \geq \left(1-O\left(k^{-1/4}\sqrt{\log m/\delta}\right)\right)\cdot\sum_{j=1}^{m} g_j\left(\frac{n}{|G_1|}\cdot\sum_{i\in\widetilde{S}\cap C_j} W_{ij}\right). \quad (38)$$

Combining, Cases A and B and using the definition of $\widetilde{F}$ (Equation (13)), it follows that with probability at least $1-\delta$,

$$\widetilde{F}(T) \geq \left(1-O\left(k^{-1/4}\sqrt{\log m/\delta}\right)\right)\cdot\widetilde{F}(\widetilde{S}).$$

Chaining the above inequality with the lower bound on $\widetilde{F}(\widetilde{S})$ from Lemma F.5, Lemma F.6 follows. □

Lemma F.7. *With probability at least $1-\delta$, it holds that for all $j \in [m]$, the subset $S_j$ computed in Algorithm 1 satisfies*

$$g_j(S_j) \geq \left(1-O\left(\gamma^{-1}k^{-1/2}\log\frac{1}{\delta}\right)\right)\cdot\max_{T\subseteq C_j:|T|\leq k_j} g_j\left(\sum_{i\in T} W_{ij}\right).$$

*The algorithm runs in time $O(\left|C_j\right|\log\left|C_j\right|)$.*

We present the proof of Lemma F.7 in Supplementary Material F.2.4.

### F.2.3 *Proof of Theorem 4.3 using Lemmas F.6 and F.7.*

Proof.

**(Consequence of Assumption 1).** Since Assumption 1 holds, we have that

$$F(S) = \sum_{j=1}^{m} F(S\cap C_j).$$

**Algorithm 4** Algorithm from Lemma F.7

**Input:** Observed utilities $\widehat{W} \in \mathbb{R}^{n\times m}_{\geq 0}$, a number $k$, and protected groups $G_1, \ldots, G_p$
**Output:** A subset $S$ with $|S| \leq k$

```
Initialize S = ∅
for groups t ∈ [p] do
    Initialize T := G_t
    for iterations r ∈ [k/p] do
        Let i be the item that maximizes Ŵ_{i1} among all items in T
        Set T = T \ {i}
        Set S = S ∪ {i}
    end for
end for
return S
```

Hence,

$$\begin{aligned}\max_{T\subseteq[n]:\forall j,\ |T\cap C_j|\leq k_j} F(T) &= \max_{T\subseteq[n]:\forall j,\ |T\cap C_j|\leq k_j} \sum_{j=1}^{m} F(T\cap C_j)\\ &= \max_{T_1\subseteq C_1,\ldots,T_m\subseteq C_m:\forall j,\ |T_j|\leq k_j} \sum_{j=1}^{m} F(T_j)\\ &= \sum_{j=1}^{m} \max_{T_j\subseteq C_j:|T_j|\leq k_j} F(T_j)\\ &= \sum_{j=1}^{m} \max_{T_j\subseteq C_j:|T_j|\leq k_j} g_j(T_j).\end{aligned}$$

(Using that under Assumption 1 for any $h, j \in [m]$, with $h \neq j$ and set $A \subseteq C_j$, $g_h(A) = 0$) (39)

**(Consequence of Lemma F.6).** Let $\mathscr{E}$ be the event that a subset $S_E$ satisfying Equation (34) exists. Conditioned on $\mathscr{E}$, it holds that

$$\max_{S\subseteq[n]:\forall j,\ |S\cap C_j|\leq k_j} F(S) \geq \mathrm{OPT}\cdot\left(1 - O\left(\sqrt{\frac{m^4}{\tau^2\gamma^5 k^{1/2}}\log\frac{m}{\delta}}\right)\right). \tag{40}$$

Combining Equations (39) and (40), we get that conditioned on $\mathscr{E}$, the following holds

$$\sum_{j=1}^{m} \max_{T_j\subseteq C_j:|T_j|\leq k_j} F(T_j) \geq \mathrm{OPT}\left(1 - O\left(\sqrt{\frac{m^4}{\tau^2\gamma^5 k^{1/2}}\log\frac{m}{\delta}}\right)\right). \tag{41}$$

Let $\mathscr{F}$ be the event that

$$\forall j \in [m],\quad g_j(S_j) \geq \left(1 - O\left(\sqrt{\frac{1}{\gamma^2 k}}\log\frac{1}{\delta}\right)\right)\cdot \max_{T_j\subseteq C_j:|T_j|\leq k_j} g_j(T_j). \tag{42}$$

Summing Equation (42) over all $j \in [m]$ and chaining it with Equation (41), we get the following: Conditioned on $\mathscr{E}$ and $\mathscr{F}$ it holds that

$$F(S) = \sum_{j=1}^{m} g_j(S_j) \geq \mathrm{OPT}\cdot\left(1 - O\left(\sqrt{\frac{m^4}{\tau^2\gamma^5 k^{1/2}}}\cdot\log\frac{m}{\delta}\right)\right). \tag{43}$$

Lemma F.6 implies that $\Pr[\mathscr{E}] \geq 1-\delta$ and Lemma F.3 implies that $\Pr[\mathscr{F}] \geq 1-\delta$. Hence, $\Pr[\mathscr{E},\mathscr{F}] \geq 1 - \Pr[\neg\mathscr{E}] - \Pr[\neg\mathscr{F}] \geq 1 - 2\delta$. Since, $k \geq k_0$ the result follows by choosing a small enough $k_0$ such that Equation (43) implies that $F(S) \geq (1-\varepsilon)\cdot\mathrm{OPT}$ and $2\delta \leq \varepsilon$. □

### *F.2.4 Proof of Lemma F.7.*

Proof. This proof only considers the $j$-th attribute and items in $C_j$. To simplify the notation, for each $i \in C_j$, let $w_i := W_{ij}$ and $\widehat{w}_i := \widehat{W}_{ij}$. Further, define

$$H(S) = g\left(\sum_{i\in S} w_i\right) \quad \text{and} \quad \widehat{H}(S) = g\left(\sum_{i\in S}\widehat{w}_i\right). \tag{44}$$

Let $S^\star \subseteq C_j$ be the subset that maximizes $H(S)$ subject to having size at most $k_j$. Let $S_j \subseteq C_j$ be the subset that maximizes $H(S)$ subject to having size at most $k_j$ and satisfying the proportional representation constraints (i.e., the constraints encoded by $\forall t \in [p], U_t = \left|G_t \cap C_j\right| \cdot \frac{k}{n}$). Note that this $S_j$ is the same as $S_j$ computed in Algorithm 1. We will prove that with high probability

$$\sum\nolimits_{i \in S_j} w_i \geq (1-\varepsilon) \cdot \sum\nolimits_{i \in S^\star} w_i.$$

Lemma F.7 follows from the above because $0 \leq \varepsilon < 1$ and $x > 0$,

$$g((1-\varepsilon)\cdot x) \geq (1-\varepsilon)\cdot g(x) + \varepsilon \cdot g(0) = (1-\varepsilon)\cdot g(x),$$

as $g$ is concave and $g(0) = 0$. In the remainder of the proof, we set

$$H(S) = \sum\nolimits_{i \in S} w_i \quad \text{and} \quad \widehat{H}(S) = \sum\nolimits_{i \in S} \widehat{w}_i. \tag{45}$$

Suppose that all values in $\left\{w_i : i \in C_j\right\}$ are unique. This can be ensured by perturbing the values by an infinitesimal amount. It only changes the value of $H(S)$ by an infinitesimal amount.

**Claim A.** We claim that for each $\ell \in [p]$, with probability at least $1-\delta$, it holds that

$$\frac{H(S_j \cap G_\ell)}{H(S^\star \cap G_\ell)} \geq 1-\varepsilon. \tag{46}$$

**Proof of Lemma F.7 assuming Claim A is true.** If this claim is true, then the result follows: Using the union bound over all $\ell \in [p]$, we get that with probability at least $1-\delta$, Equation (46) holds for all $\ell \in [p]$. Conditioned on the event that Equation (46) holds for all $\ell \in [p]$, we have

$$\begin{aligned}
\frac{H(S_j)}{H(S^\star)} &= \frac{\sum_{\ell \in [p]} H(S_j \cap G_\ell)}{\sum_{\ell \in [p]} H(S^\star \cap G_\ell)} \\
&\geq \frac{(1-\varepsilon)\cdot \sum_{\ell \in [p]} H(S^\star \cap G_\ell)}{\sum_{\ell \in [p]} H(S^\star \cap G_\ell)} && \text{(Using Equation (46))} \\
&= (1-\varepsilon).
\end{aligned}$$

**Proof of Claim A.** Fix any $\ell \subseteq [p]$. Let $i(h)$ have the $h$-th largest value of $w$ in $\left\{w_i : i \in G_\ell \cap C_j\right\}$. Because $\phi_\ell$ is increasing, it follows that $i(h)$ also has the $h$-th largest value of $\widehat{w}$ in $\left\{\widehat{w}_i : i \in G_\ell \cap C_j\right\} = \left\{\phi_\ell(w_i) : i \in G_\ell \cap C_j\right\}$. Define

$$r \coloneqq \left|S^\star \cap G_\ell\right| \quad \text{and} \quad \widetilde{r} \coloneqq \left|S_j \cap G_\ell\right|.$$

It holds that

$$S^\star \cap G_\ell = \{i(1), \ldots, i(r)\}.$$

(Otherwise, we can increase $H(S^\star)$ by swapping an element of $S^\star$ by an element in $\{i(1), \ldots, i(r)\} \setminus S^\star$.) Further, it also holds that

$$S_j \cap G_\ell = \{i(1), \ldots, i(\widetilde{r})\}.$$

(Otherwise, we can increase $\widehat{H}(S_j)$ by swapping an element $i$ of $S_j$ by an element $i'$ in $\{i(1), \ldots, i(\widetilde{r})\} \setminus S_j$; note that this does not violate the proportional representation constraint because both $i, i' \in G_\ell$.)

*(Step 1: Lower bound on $\frac{H(S_j \cap G_\ell)}{H(S^\star \cap G_\ell)}$)* Using the above observations we have the following expression

$$\frac{H(S_j \cap G_\ell)}{H(S^\star \cap G_\ell)} = \frac{w_{i(1)} + w_{i(2)} + \cdots + w_{i(\widetilde{r})}}{w_{i(1)} + w_{i(2)} + \cdots + w_{i(r)}}.$$

Consequently, if $\widetilde{r} \geq r$, then $\frac{H(S_j \cap G_\ell)}{H(S^\star \cap G_\ell)} \geq 1$. Otherwise, we have the following lower bound

$$\begin{aligned}
\frac{H(S_j \cap G_\ell)}{H(S^\star \cap G_\ell)} &= \frac{w_{i(1)} + w_{i(2)} + \cdots + w_{i(\widetilde{r})}}{w_{i(1)} + w_{i(2)} + \cdots + w_{i(r)}} \\
&\geq \frac{w_{i(1)} + w_{i(2)} + \cdots + w_{i(\widetilde{r})}}{w_{i(1)} + w_{i(2)} + \cdots + w_{i(\widetilde{r})} \cdot (r - \widetilde{r} + 1)} && \text{(Using that } w_{i(1)} \geq w_{i(2)} \geq \cdots \geq w_{i(r)}\text{)} \\
&\geq 1 - \frac{w_{i(\widetilde{r})} \cdot (r - \widetilde{r})}{w_{i(1)} + w_{i(2)} + \cdots + w_{i(\widetilde{r})} \cdot (r - \widetilde{r} + 1)} && \text{(Using that } w_{i(1)} \geq w_{i(2)} \geq \cdots \geq w_{i(r)}\text{)} \\
&\geq \frac{\widetilde{r}}{r}. && \text{(Using that } w_{i(\widetilde{r})} \leq w_{i(i)} \text{ for all } 1 \leq i \leq \widetilde{r}\text{)}
\end{aligned}$$

Hence, in either case

$$\frac{H(S_j \cap G_\ell)}{H(S^\star \cap G_\ell)} \geq \frac{\widetilde{r}}{r}. \tag{47}$$

*(Step 2: Lower bound on $\frac{\widetilde{r}}{r}$).* First,

$$\widetilde{r} = k \cdot \frac{\left|G_\ell \cap C_j\right|}{n}. \tag{48}$$

(Otherwise, adding $i(\widetilde{r}+1)$ to $S_j$ increases its utility and does not violate the proportional representation constraint). Next, because $S^\star \subseteq C_j$ is independent of the protected groups $G_1, \ldots, G_p$ and because $G_\ell$ is constructed by drawing $|G_\ell|$ elements uniformly without replacement from $C_j$, it follows that

$$r \coloneqq \left|S^\star \cap G_\ell\right|,$$

is a random variable distributed according to the hyper-geometric distribution with parameters $n = |G_\ell|$, $K = \left|S^\star\right|$, and $N = \left|C_j\right|$. (Recall that for a hyper-geometric random variable $X$, $\Pr[X = r]$ denotes the probability of obtaining $r$ red balls in $n$ draws, without replacement, from an urn with $K$ red balls and $N - K$ blue balls.) From standard properties of the hyper-geometric distribution [29, Theorem 1], it follows that

$$\mathbb{E}[r] = \left|G_\ell \cap C_j\right| \cdot \frac{\left|S^\star\right|}{n} = \left|G_\ell \cap C_j\right| \cdot \frac{k}{n} \quad \text{and} \quad \Pr\left[\left|r - \left|G_\ell \cap C_j\right| \cdot \frac{k}{n}\right| \leq \ln\left(\frac{1}{\delta}\right)\sqrt{k}\right] \leq \delta. \tag{49}$$

Consequently, with probability at least $1 - \delta$

$$r \leq \left|G_\ell \cap C_j\right| \cdot \frac{k}{n} + \ln\left(\frac{1}{\delta}\right)\sqrt{k} = \widetilde{r} + \ln\left(\frac{1}{\delta}\right)\sqrt{k}. \tag{50}$$

*(Step 3: Proof of Claim A).*. It follows that, with probability at least $1 - \delta$,

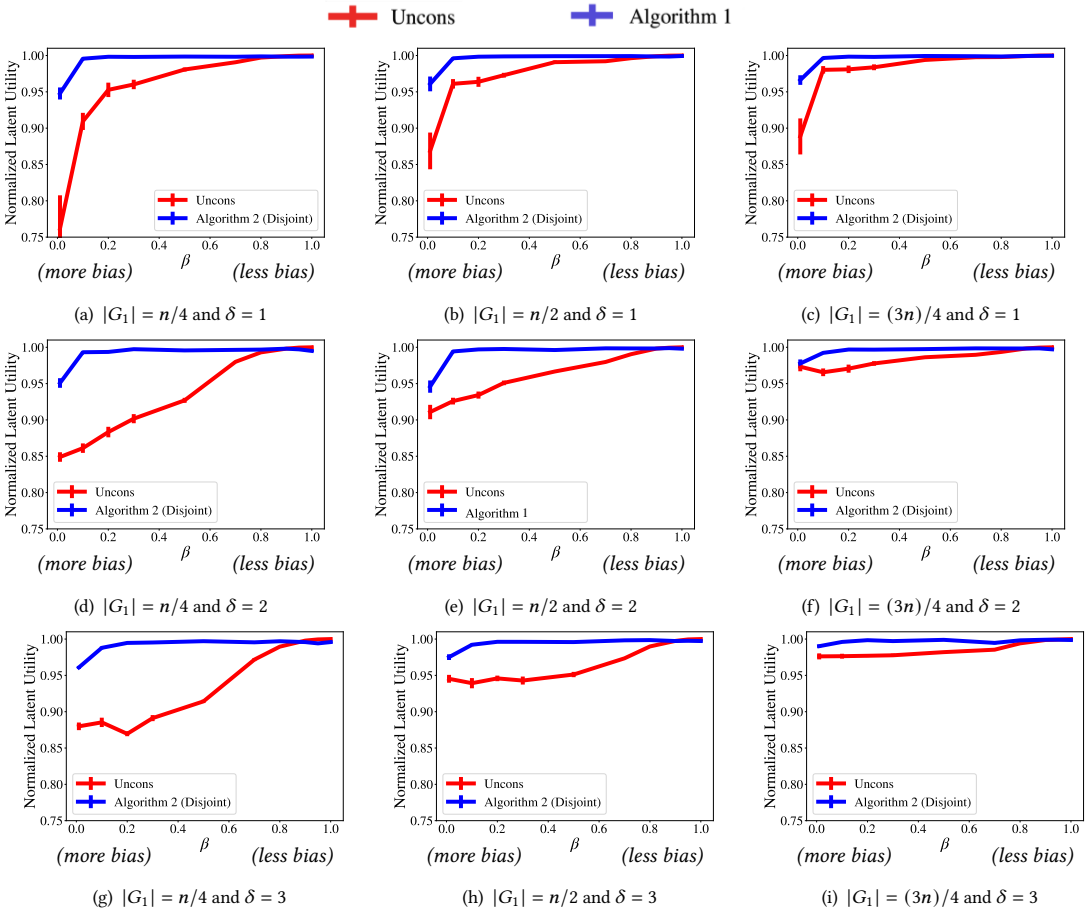
$$\begin{aligned}
\frac{H(S_j \cap G_\ell)}{H(S^\star \cap G_\ell)} &\geq \frac{\widetilde{r}}{r} && \text{(Using Equation (47))} \\
&\geq \frac{1}{1 + \frac{\ln\left(\frac{1}{\delta}\right)\sqrt{k}}{\widetilde{r}}} && \text{(Using Equation (50))} \\
&\geq 1 - \frac{\ln\left(\frac{1}{\delta}\right)\sqrt{k}}{\widetilde{r}} && \text{(Using } 1 - x \leq \tfrac{1}{1+x} \text{ for all } x \in \mathbb{R}\text{)} \\
&= 1 - \frac{\ln\left(\frac{1}{\delta}\right) \cdot n}{\sqrt{k} \cdot |G_\ell|} && \text{(Using Equation (48))} \\
&\geq 1 - \frac{\ln\left(\frac{1}{\delta}\right)}{\gamma\sqrt{k}}.
\end{aligned}$$

□

# G ADDITIONAL PLOTS FOR SIMULATIONS FROM SECTION 5

In this section, we give additional plots for the simulations in Section 5 on synthetic data (Supplementary Materials G.1 and G.2) and on MovieLens 20M (Supplementary Material G.3).

## G.1 Additional plots for the simulation with synthetic data 1



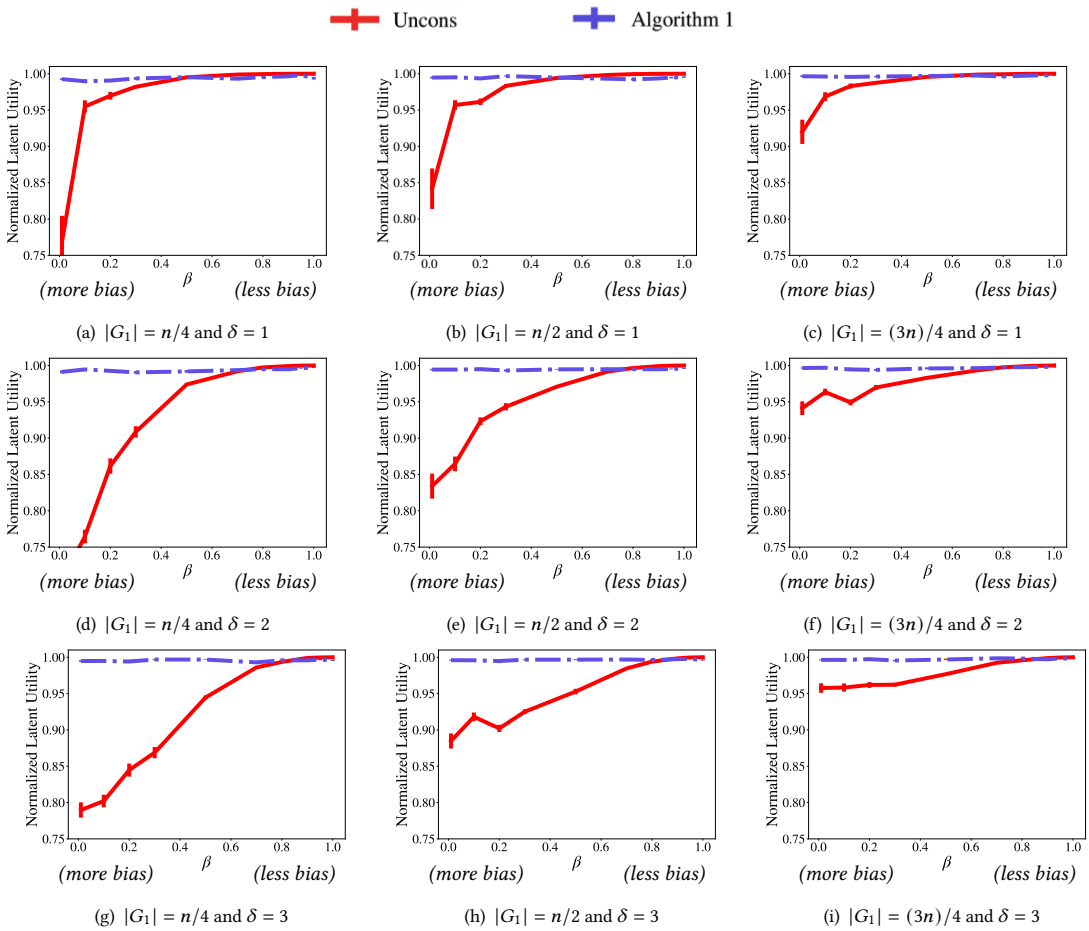

(a) $|G_1| = n/4$ and $\delta = 1$ (b) $|G_1| = n/2$ and $\delta = 1$ (c) $|G_1| = (3n)/4$ and $\delta = 1$

(d) $|G_1| = n/4$ and $\delta = 2$ (e) $|G_1| = n/2$ and $\delta = 2$ (f) $|G_1| = (3n)/4$ and $\delta = 2$

(g) $|G_1| = n/4$ and $\delta = 3$ (h) $|G_1| = n/2$ and $\delta = 3$ (i) $|G_1| = (3n)/4$ and $\delta = 3$

**Figure 3: *Simulation on synthetic data 1:* We vary the bias parameter $\beta \in [0, 1]$, run Algorithm 1 and Uncons on synthetic data 1, and report their normalized latent utility (error bars denote standard error of the mean). The $x$-axis shows $\beta \in [0, 1]$ and the $y$-axis plots the normalized latent utility. The results show that Uncons can lose significant fraction of the optimal latent utility in the presence of bias (up to 15% for $\beta < 0.1$, $\frac{|G_1|}{n} = \frac{1}{4}$, and $\delta = 3$). While Algorithm 1 loses less than 5% of the optimal latent utility across all choices of parameters.**

## G.2 Additional plots for the simulation with synthetic data 2




(a) $|G_1| = n/4$ and $\delta = 1$ (b) $|G_1| = n/2$ and $\delta = 1$ (c) $|G_1| = (3n)/4$ and $\delta = 1$

(d) $|G_1| = n/4$ and $\delta = 2$ (e) $|G_1| = n/2$ and $\delta = 2$ (f) $|G_1| = (3n)/4$ and $\delta = 2$

(g) $|G_1| = n/4$ and $\delta = 3$ (h) $|G_1| = n/2$ and $\delta = 3$ (i) $|G_1| = (3n)/4$ and $\delta = 3$

**Figure 4: *Simulation on synthetic data 2:* We vary the bias parameter $\beta \in [0, 1]$, run Algorithm 1 and Uncons on synthetic data 2 and report their normalized latent utility (error bars denote standard error of the mean). The $x$-axis shows $\beta \in [0, 1]$ and the $y$-axis plots the normalized latent utility. The results show that Uncons can lose significant fraction of the optimal latent utility in the presence of bias (up to 25% for $\beta < 0.1$, $\frac{|G_1|}{n} = \frac{1}{4}$, and $\delta = 2$). While Algorithm 1 loses less than 1% of the optimal latent utility across all choices of parameters.**

## G.3 Additional plots for simulation with MovieLens 20M data

| Genre | Male led | Non-Male led | Ratio $\left(\frac{\text{Non-Male led}}{\text{Male led}}\right)$ |
|---|---|---|---|
| Action | 0.0604 | 0.0213 | 0.352 |
| Adventure | 0.0334 | 0.0146 | 0.437 |
| Animation | 0.0187 | 0.0138 | 0.741 |
| Children | 0.0168 | 0.0164 | 0.9779 |
| Comedy | 0.0929 | 0.0646 | 0.6951 |
| Crime | 0.0311 | 0.0135 | 0.4351 |
| Documentary | 0.0157 | 0.0141 | 0.903 |
| Drama | 0.1035 | 0.1238 | 1.1961 |
| Fantasy | 0.0165 | 0.0145 | 0.8798 |
| Horror | 0.0293 | 0.0489 | 1.6675 |
| Musical | 0.0105 | 0.0162 | 1.5458 |
| Mystery | 0.0128 | 0.0123 | 0.9612 |
| Romance | 0.0306 | 0.0689 | 2.2492 |
| Sci-fi | 0.0287 | 0.0169 | 0.5878 |
| Thriller | 0.0378 | 0.0316 | 0.8368 |
| War | 0.014 | 0.0061 | 0.4371 |
| Western | 0.0093 | 0.0027 | 0.2923 |

**Table 1: *Average relevance scores for movies which are led by Male and Non-Male actors respectively, across different genres.* Let $S_{\mathrm{M}}$ be the set of movies led by male actors and $S_{\mathrm{NM}}$ be the set of movies led by non-male actors (see Section 5 for details about computing $S_{\mathrm{M}}$ and $S_{\mathrm{NM}}$). The MovieLens 20M data specifies sets the $S_g$ of movies in genre $g$ and for each movie $i$ and genre $g$, it specifies a predicted relevance score $r_{ig} \in [0, 1]$ indicating "how strongly [movie $i$] exhibits particular properties represented by [genre $g$]." For each genre $g$, we report the average relevance scores of movies in $S_{\mathrm{M}} \cap S_g$ and $S_{\mathrm{NM}} \cap S_g$ We observe that in genres that are stereotypically associated with men (e.g., "Action" or "War") movies led by male actors have a disproportionately higher relevance-scores on compared to movies led by non-male actors (differing by up to 300%).**

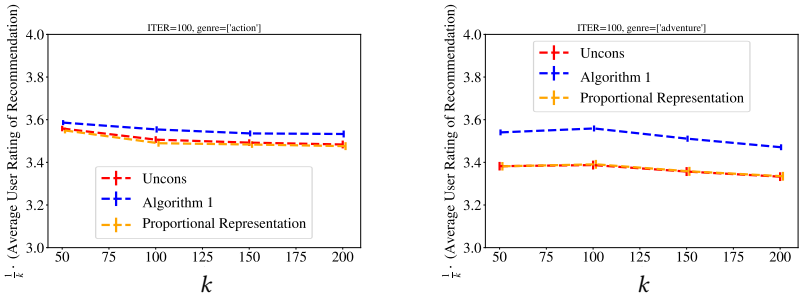
| Genre | Male led | Non-Male led | Ratio $\left(\frac{\text{Non-Male led}}{\text{Male led}}\right)$ |
|---|---|---|---|
| Action | 3.11 (0.53) | 3.04 (0.54) | 0.98 |
| Adventure | 3.22 (0.51) | 3.07 (0.60) | 0.95 |
| Animation | 3.32 (0.45) | 3.43 (0.49) | 1.03 |
| Children | 3.03 (0.54) | 3.18 (0.53) | 1.05 |
| Comedy | 3.16 (0.53) | 3.15 (0.48) | 1.00 |
| Crime | 3.36 (0.45) | 3.24 (0.50) | 0.96 |
| Documentary | 3.61 (0.41) | 3.59 (0.34) | 1.00 |
| Drama | 3.44 (0.40) | 3.40 (0.40) | 0.99 |
| Fantasy | 3.22 (0.52) | 3.24 (0.48) | 1.01 |
| Horror | 2.88 (0.57) | 2.88 (0.53) | 1.00 |
| Musical | 3.36 (0.43) | 3.30 (0.46) | 0.98 |
| Mystery | 3.41 (0.44) | 3.21 (0.50) | 0.94 |
| Romance | 3.37 (0.43) | 3.31 (0.45) | 0.98 |
| Sci-fi | 3.11 (0.57) | 3.07 (0.57) | 0.99 |
| Thriller | 3.25 (0.47) | 3.12 (0.49) | 0.96 |
| War | 3.51 (0.45) | 3.61 (0.31) | 1.03 |
| Western | 3.38 (0.41) | 3.36 (0.34) | 0.99 |

**Table 2: *Average user-ratings for movies which are led by Male and Non-Male actors respectively, across different genres.* Let $S_{\text{M}}$ be the set of movies led by male actors and $S_{\text{NM}}$ be the set of movies led by non-male actors (see Section 5 for details about computing $S_{\text{M}}$ and $S_{\text{NM}}$). The MovieLens 20M data specifies sets the $S_g$ of movies in genre $g$ and for each movie $i$, it specifies the average user rating $\text{rat}_i \in [0, 1]$ for movie $i$. For each genre $g$, we report the average user rating of movies in $S_{\text{M}} \cap S_g$ and $S_{\text{NM}} \cap S_g$ We observe that in genres that are stereotypically associated with men (e.g., "Action" or "War") movies led by male actors have a disproportionately higher relevance-scores on compared to movies led by non-male actors (differing by up to 300%).**

| Genre | Male led | Non-Male led | Ratio $\left(\frac{\text{Non-Male led}}{\text{Male led}}\right)$ |
|---|---|---|---|
| Action | 0.061 | 0.0274 | 0.4486 |
| Adventure | 0.0331 | 0.0168 | 0.5079 |
| Animation | 0.0185 | 0.0161 | 0.8704 |
| Children | 0.0169 | 0.0171 | 1.0114 |
| Comedy | 0.0937 | 0.0719 | 0.7679 |
| Crime | 0.0312 | 0.0148 | 0.475 |
| Documentary | 0.0157 | 0.0126 | 0.8029 |
| Drama | 0.1042 | 0.1181 | 1.1333 |
| Fantasy | 0.0162 | 0.0154 | 0.945 |
| Horror | 0.0287 | 0.0492 | 1.7129 |
| Musical | 0.0107 | 0.0148 | 1.3819 |
| Mystery | 0.0126 | 0.0118 | 0.9346 |
| Romance | 0.0306 | 0.0653 | 2.1365 |
| Sci-fi | 0.0285 | 0.0179 | 0.6298 |
| Thriller | 0.0372 | 0.0334 | 0.8988 |
| War | 0.0141 | 0.0063 | 0.4467 |
| Western | 0.0091 | 0.0027 | 0.3024 |

**Table 3: *Average user-ratings for movies which are led by Male and Non-Male actors respectively, across different genres.* For each movie $i$, we predict the (probable) gender of its lead actor using the Genderize API (gender-api.com). For the main simulation, we remove all movies for which this prediction has confidence less than 0.9. Here, we consider the set of all movies where Genderize outputs a prediction which is not NA, and for which we have both user ratings and a relevance scores. This results in 7382 movies $S'_{\text{M}}$ led by male actors and 2572 movies $S'_{\text{NM}}$ led by non-male actors. The MovieLens 20M data specifies sets the $S_g$ of movies in genre $g$ and for each movie $i$, it specifies the average user rating $\text{rat}_i \in [0, 1]$ for movie $i$. For each genre $g$, we report the average user rating of movies in $S_{\text{M}} \cap S_g$ and $S_{\text{NM}} \cap S_g$ We observe that in genres that are stereotypically associated with men (e.g., "Action" or "War") movies led by male actors have a disproportionately higher relevance-scores on compared to movies led by non-male actors (differing by up to 300%).**

### G.3.1 *Plots with one genre ($|T| = 1$).*



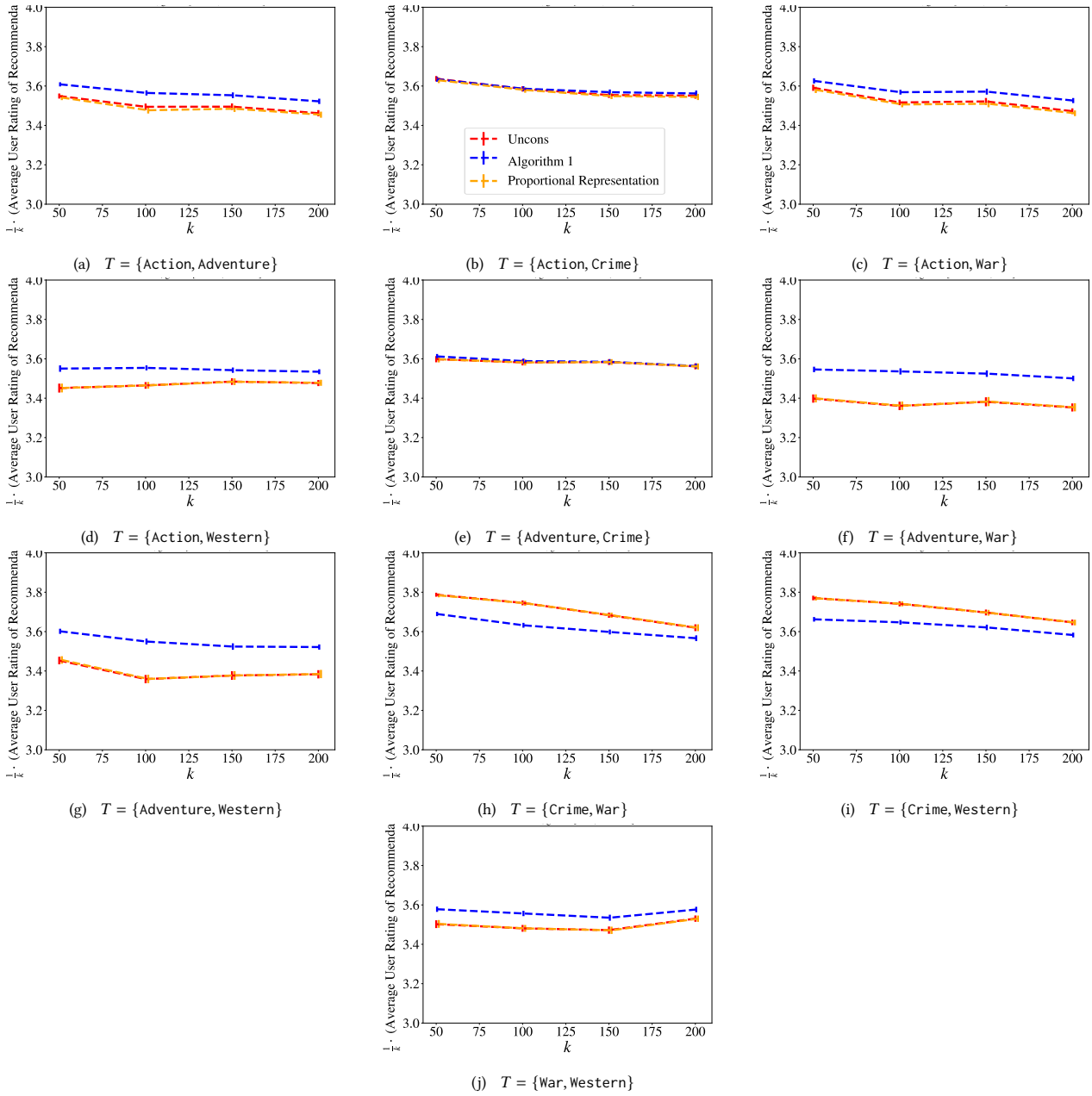

(a) $T = \{\texttt{Action}\}$ (b) $T = \{\texttt{Adventure}\}$

(c) $T = \{\texttt{Crime}\}$ (d) $T = \{\texttt{War}\}$

(e) $T = \{\texttt{Western}\}$

**Figure 5: *Simulation with MovieLens data with movie recommendations from a single men-stereotypical genre:* We observe that the relevance scores in the MovieLens data are disproportionately higher (by up to 3 times) for movies led male actors compared to movies led by non-male actors in genres stereotypically associated with men. In contrast, user ratings for these sets of movies are within 6% of each other in all genres. We chose genres where the ratio of average relevance scores of men-led movies is at least twice that of non-men-led movies, they are: $B = \{\texttt{action}, \texttt{adventure}, \texttt{crime}, \texttt{western}, \text{and } \texttt{war}\}$. We use relevance scores to recommend $k \in \{50,100,150,200\}$ movies from different subsets $T$ of $B$. This figure presents the results for all subsets $T \subseteq B$ of size 1. We observe that in 4 out of 5 subfigures Algorithm 1 has a higher normalized latent utility than Uncons for all $k$. (Results for other subsets of genres appear in Figures 5 to 9.)**

### G.3.2 *Plots with two genres ($|T| = 2$).*




(a) $T = \{\texttt{Action}, \texttt{Adventure}\}$ (b) $T = \{\texttt{Action}, \texttt{Crime}\}$ (c) $T = \{\texttt{Action}, \texttt{War}\}$

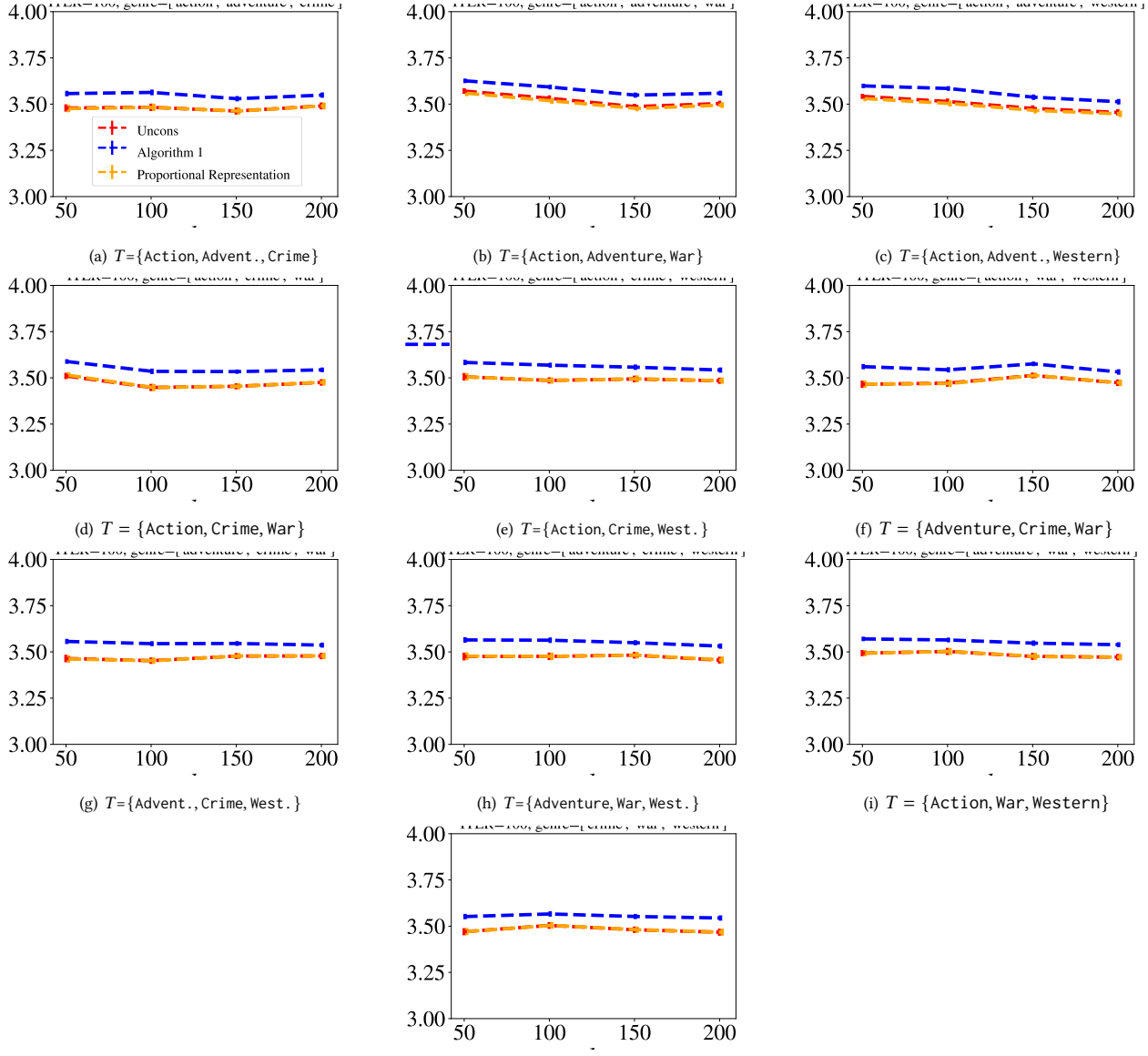
(d) $T = \{\texttt{Action}, \texttt{Western}\}$ (e) $T = \{\texttt{Adventure}, \texttt{Crime}\}$ (f) $T = \{\texttt{Adventure}, \texttt{War}\}$

(g) $T = \{\texttt{Adventure}, \texttt{Western}\}$ (h) $T = \{\texttt{Crime}, \texttt{War}\}$ (i) $T = \{\texttt{Crime}, \texttt{Western}\}$

(j) $T = \{\texttt{War}, \texttt{Western}\}$

**Figure 6: *Simulation with MovieLens data with movie recommendations from two men-stereotypical genres:* We observe that the relevance scores in the MovieLens data are disproportionately higher (by up to 3 times) for movies led male actors compared to movies led by non-male actors in genres stereotypically associated with men. In contrast, user ratings for these sets of movies are within 6% of each other in all genres. We chose genres where the ratio of average relevance scores of men-led movies is at least twice that of non-men-led movies, they are: $B = \{\texttt{action}, \texttt{adventure}, \texttt{crime}, \texttt{western}, \text{and } \texttt{war}\}$. We use relevance scores to recommend $k \in \{50,100,150,200\}$ movies from different subsets $T$ of $B$. This figure presents the results for all subsets $T \subseteq B$ of size 2. We observe that in 8 out of 10 subfigures Algorithm 1 has a similar or higher normalized latent utility than Uncons for all $k$. (Results for other subsets of genres appear in Figures 5 to 9.)**

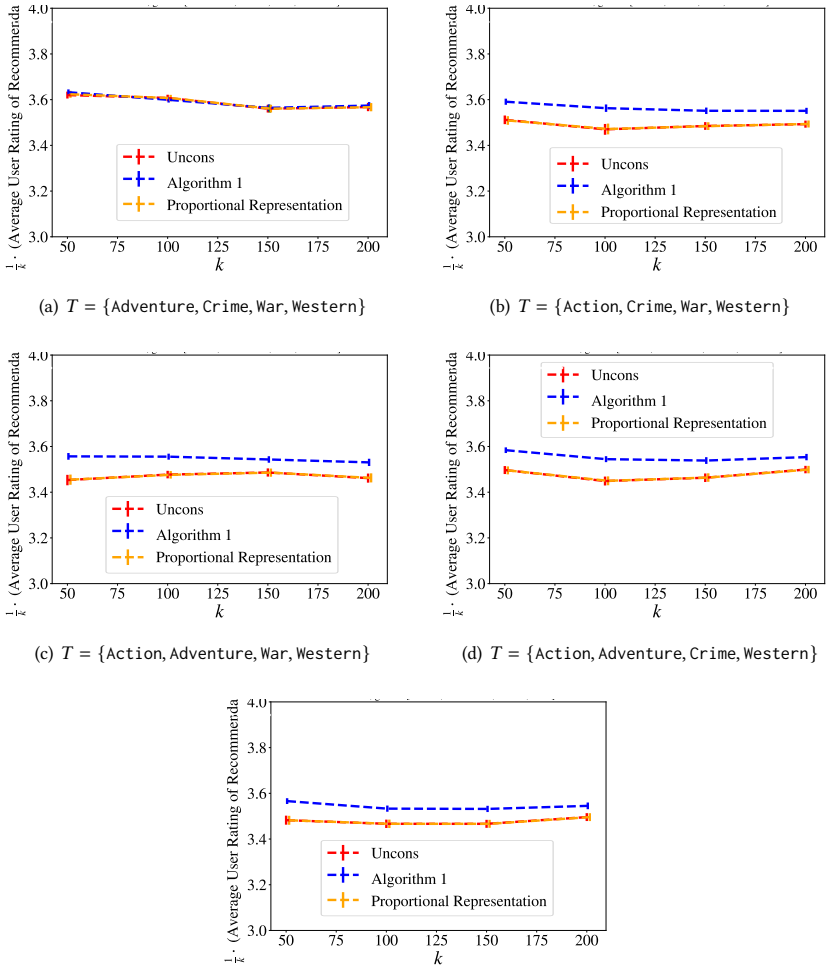
### G.3.3 *Plots with three genres ($|T| = 3$).*




(a) $T$={Action, Advent., Crime}
(b) $T$={Action, Adventure, War}
(c) $T$={Action, Advent., Western}
(d) $T$ = {Action, Crime, War}
(e) $T$={Action, Crime, West.}
(f) $T$ = {Adventure, Crime, War}
(g) $T$={Advent., Crime, West.}
(h) $T$={Adventure, War, West.}
(i) $T$ = {Action, War, Western}
(j) $T$ = {Crime, War, Western}

**Figure 7: *Simulation with MovieLens data with movie recommendations from three men-stereotypical genres:* We observe that the relevance scores in the MovieLens data are disproportionately higher (by up to 3 times) for movies led male actors compared to movies led by non-male actors in genres stereotypically associated with men. In contrast, user ratings for these sets of movies are within 6% of each other in all genres. We chose genres where the ratio of average relevance scores of men-led movies is at least twice that of non-men-led movies, they are: $B$ = {action, adventure, crime, western, and war}. We use relevance scores to recommend $k \in \{50,100,150,200\}$ movies from different subsets $T$ of $B$. This figure presents the results for all subsets $T \subseteq B$ of size 3. We observe that in 10 out of 10 subfigures Algorithm 1 has a similar or higher normalized latent utility than Uncons for all $k$. (Results for other subsets of genres appear in Figures 5 to 9.)**

### G.3.4 *Plots with four genres ($|T| = 4$).*




(a) $T = \{\texttt{Adventure}, \texttt{Crime}, \texttt{War}, \texttt{Western}\}$

(b) $T = \{\texttt{Action}, \texttt{Crime}, \texttt{War}, \texttt{Western}\}$

(c) $T = \{\texttt{Action}, \texttt{Adventure}, \texttt{War}, \texttt{Western}\}$

(d) $T = \{\texttt{Action}, \texttt{Adventure}, \texttt{Crime}, \texttt{Western}\}$

(e) $T = \{\texttt{Action}, \texttt{Adventure}, \texttt{Crime}, \texttt{War}\}$

**Figure 8: *Simulation with MovieLens data with movie recommendations from four men-stereotypical genres:* We observe that the relevance scores in the MovieLens data are disproportionately higher (by up to 3 times) for movies led male actors compared to movies led by non-male actors in genres stereotypically associated with men. In contrast, user ratings for these sets of movies are within 6% of each other in all genres. We chose genres where the ratio of average relevance scores of men-led movies is at least twice that of non-men-led movies, they are: $B = \{\texttt{action}, \texttt{adventure}, \texttt{crime}, \texttt{western}, \text{ and } \texttt{war}\}$. We use relevance scores to recommend $k \in \{50{,}100{,}150{,}200\}$ movies from different subsets $T$ of $B$. This figure presents the results for all subsets $T \subseteq B$ of size 4. We observe that in 4 out of 5 subfigures Algorithm 1 has a higher normalized latent utility than Uncons for all $k$. (Results for other subsets of genres appear in Figures 5 to 9.)**

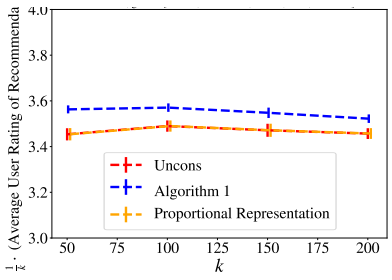



**Figure 9: *Simulation with MovieLens data with movie recommendations from five men-stereotypical genres:* We observe that the relevance scores in the MovieLens data are disproportionately higher (by up to 3 times) for movies led male actors compared to movies led by non-male actors in genres stereotypically associated with men. In contrast, user ratings for these sets of movies are within 6% of each other in all genres. We chose genres where the ratio of average relevance scores of men-led movies is at least twice that of non-men-led movies, they are: $B = \{$`action`, `adventure`, `crime`, `western`, and `war`$\}$. We use relevance scores to recommend $k \in \{50,100,150,200\}$ movies from different subsets $T$ of $B$. This figure presents the results for all subsets $T \subseteq B$ of size 5. We observe that the figure Algorithm 1 has a similar normalized latent utility than Uncons for all $k$. (Results for other subsets of genres appear in Figures 5 to 9.)**

*G.3.5 Plots with five genres ($|T| = 5$).*

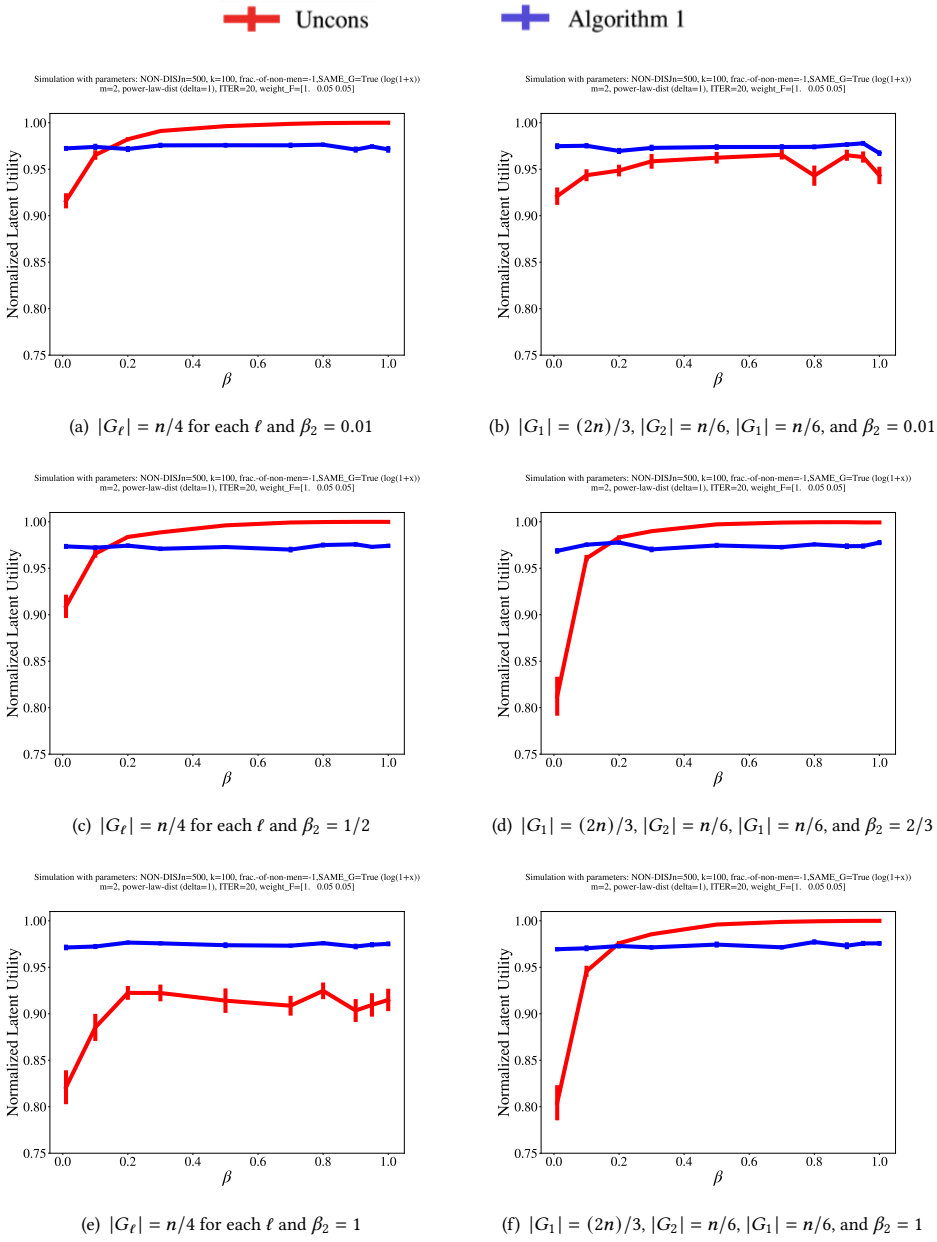
# H SIMULATIONS ON SYNTHETIC DATA WITH MORE THAN TWO GROUPS

In this section, we evaluate the robustness of the simulations from Section 5 with synthetic data with respect to the number of groups. We present simulations with three groups $G_1$, $G_2$, and $G_3$.

*Setup.* We follow the set up introduced in Supplementary Material C. For simulations in this section, we fix $n := 500$, $k := 50$, and $\delta := 1$. For the bias parameters, we fix $\phi_\ell(x) = \beta_\ell \cdot x$ for all $x$ and $\beta_1 := 1$. We vary $\beta_2$ and $\beta_3$ in the range $[0, 1]$. We consider two group structures: (1) In the first, all three groups have an equal size, $|G_1| = |G_2| = |G_3|$, and (2) in the second, the group which does not face any bias also forms the majority, $|G_1| = \frac{2n}{3}$, $|G_2| = \frac{n}{6}$, and $|G_3| = \frac{n}{6}$.

## H.1 Simulations with more than two protected groups with synthetic data 1

In results with synthetic data 1, we observe that Algorithm 1 achieves NLU higher than 97% for variations of the parameters. Moreover, it outperforms Uncons in 5/9 of the plots–by up to 18%. In the remaining four plots, Algorithm 1's NLU is within 3% of Uncons's NLU.




(a) $|G_\ell| = n/4$ for each $\ell$ and $\beta_2 = 0.01$

(b) $|G_1| = (2n)/3$, $|G_2| = n/6$, $|G_1| = n/6$, and $\beta_2 = 0.01$

(c) $|G_\ell| = n/4$ for each $\ell$ and $\beta_2 = 1/2$

(d) $|G_1| = (2n)/3$, $|G_2| = n/6$, $|G_1| = n/6$, and $\beta_2 = 2/3$

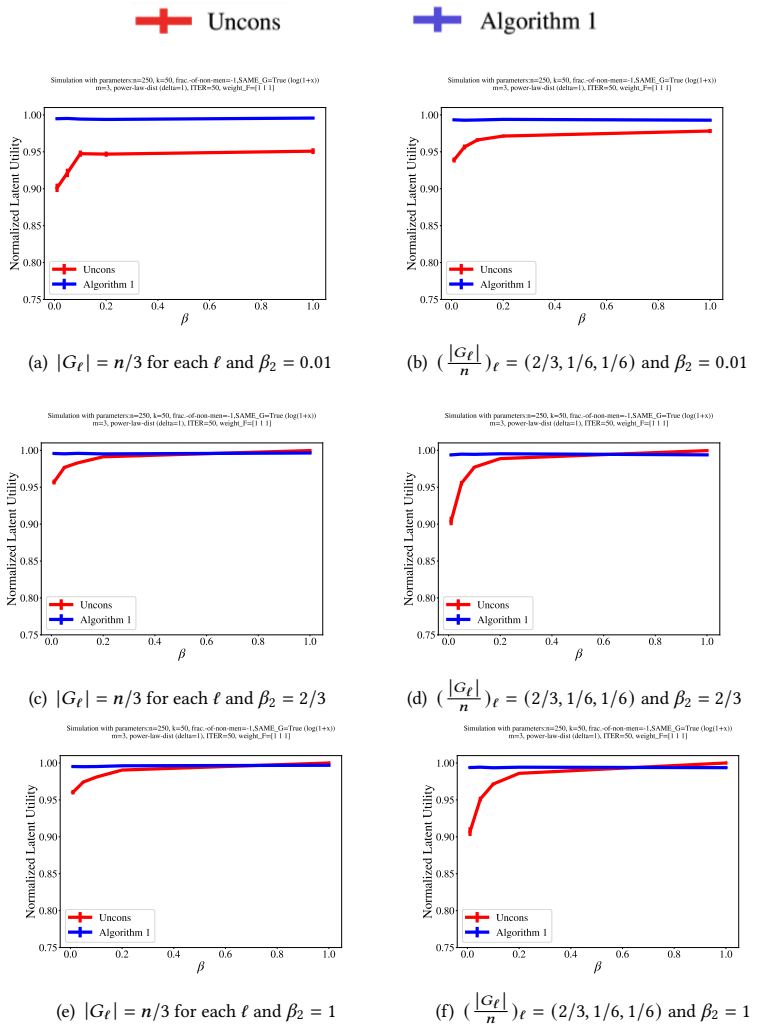
(e) $|G_\ell| = n/4$ for each $\ell$ and $\beta_2 = 1$

(f) $|G_1| = (2n)/3$, $|G_2| = n/6$, $|G_1| = n/6$, and $\beta_2 = 1$

**Figure 10: Simulation on synthetic data 1 from Supplementary Material H.**

## H.2 Simulations with more than two protected groups with synthetic data 2

In results with synthetic data 2, we observe that both Algorithm 1 achieve NLU higher than 97% for variations of the parameters. Moreover, both algorithms outperforms Uncons in all simulations–by up to 10%.




(a) $|G_\ell| = n/3$ for each $\ell$ and $\beta_2 = 0.01$

(b) $(\frac{|G_\ell|}{n})_\ell = (2/3, 1/6, 1/6)$ and $\beta_2 = 0.01$

(c) $|G_\ell| = n/3$ for each $\ell$ and $\beta_2 = 2/3$

(d) $(\frac{|G_\ell|}{n})_\ell = (2/3, 1/6, 1/6)$ and $\beta_2 = 2/3$

(e) $|G_\ell| = n/3$ for each $\ell$ and $\beta_2 = 1$

(f) $(\frac{|G_\ell|}{n})_\ell = (2/3, 1/6, 1/6)$ and $\beta_2 = 1$

**Figure 11: Simulation on synthetic data 2 from Supplementary Material H.**