COVARIANCE-INDUCED FAIRNESS GAP PENALTY FOR SUBGROUP-FAIR CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fair clustering aims to make cluster assignment independent of sensitive attributes, but this goal becomes challenging when multiple sensitive attributes jointly define many subgroups. In such cases, directly extending existing fair clustering algorithms can be computationally expensive or numerically unstable, especially when the number of subgroups grows exponentially and some subgroups contain only a few instances. To address these challenges, we define a subgroup-fairness gap for clustering and derive a covariance-based surrogate that exactly matches this gap. We then introduce a continuous relaxation of the surrogate for efficient gradient-based optimization. The resulting algorithm, COVA-FC, has a first-order stationarity guarantee. We also show that a small subgroup-fairness gap does not imply marginal fairness, and extend our framework to capture a subgroup-marginal-fairness gap. Experiments on benchmark datasets show that COVA-FC achieves competitive cost-fairness trade-offs and improves computational efficiency over existing baselines, in both subgroup and higher-order marginal settings.

1 INTRODUCTION

Algorithmic fairness is a critical requirement for decision-making systems that affect access to opportunities, resources, or services, as systematic biases and disparities can harm protected and underrepresented populations. Accordingly, regulations and policies increasingly require trustworthy and non-discriminatory algorithms, making fair learning important both ethically and practically (Goodman & Flaxman, 2017; Mehrabi et al., 2021). Fairness in machine learning has been studied across various tasks, including classification (Goh et al., 2016; Agarwal et al., 2018; Donini et al., 2018; Kim et al., 2022; 2025a), regression (Agarwal et al., 2019; Chzhen et al., 2020), generation (Xu et al., 2018; Friedrich et al., 2023), and clustering (Chierichetti et al., 2017; Bera et al., 2019; Kleindessner et al., 2019).

Among these, *fair clustering* (Chierichetti et al., 2017) aims to construct clusters whose memberships are not strongly associated with sensitive attributes. The technical goal is to achieve a desirable trade-off between clustering cost and fairness, since improving fairness typically comes with an increase in cost (Bertsimas et al., 2011; Chhabra et al., 2021).

However, most existing fair clustering algorithms were developed mainly for a single sensitive attribute (Chierichetti et al., 2017; Backurs et al., 2019; Ziko et al., 2021; Esmaeili et al., 2021; Kim et al., 2025b). In many modern applications, individuals are associated with multiple sensitive attributes, whose intersections can reveal fairness violations even when a given algorithm is fair for each sensitive attribute. For example, a lending model may equalize approval rates across gender and race overall, yet the subgroup "female & minority race" may still receive markedly lower approval rates. This motivates the need for subgroup fairness. A *subgroup* is defined by a specific combination of sensitive attribute values. Although there is a large number of studies for subgroup-fair classification (Kearns et al., 2019; Blum & Lykouris, 2020; Molina & Loiseau, 2022; Hu et al., 2024; Lai & Guan, 2025; Kim et al., 2026), not much work has been done for subgroup-fair clustering.

Existing fair clustering algorithms could be extended to this setting by treating multiple sensitive attributes as a single sensitive attribute with multiple categories, but this naive extension creates substantial computational and optimization challenges. Given q binary sensitive attributes, the number of subgroups becomes 2^q, and many subgroups can contain very few instances, each representing a single data point. Algorithms using linear programming (LP), such as (Bera et al., 2019; Esmaeili

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

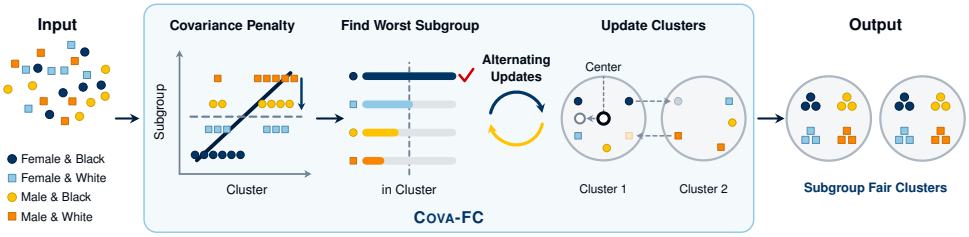


Figure 1: Overview of COVA-FC, where each marker is an instance and its color and shape indicate its subgroup. COVA-FC uses a covariance-based surrogate to penalize dependence between cluster assignment and subgroup membership (*Covariance Penalty*). It alternates between identifying the worst subgroup (red check) and updating the clusters, moving instances across clusters and shifting the cluster centers. These updates encourage each subgroup to approach the overall cluster proportions (right).

et al., 2021; Kim et al., 2025b), can offer approximation guarantees on clustering cost under their fairness constraints. However, these methods become computationally expensive as the number of subgroup-dependent constraints increases. Penalty-based methods such as VFC (Ziko et al., 2021) and FairKM (Abraham et al., 2020) could be extended to subgroup-fair clustering, but each has limitations. VFC becomes numerically unstable when very small subgroups exist because their fairness penalties depend on subgroup sizes. FairKM relies on hard cluster assignment, making it difficult to finely match the desired subgroup proportions across clusters when some subgroups contain only a few instances, but its modification for soft cluster assignment would not be easy. Thus, subgroup-fair clustering has received little attention, especially when the number of subgroups is large.

In this paper, we develop a learning algorithm for subgroup-fair clustering with multiple sensitive attributes. Formally, assume that we are given q binary sensitive attributes, and let the i-th instance have a sensitive attribute vector $\boldsymbol{s}_i \in \{0,1\}^q$, which induces 2^q number of subgroups.

Our approach. Inspired by Kearns et al. (2018), we first define a subgroup-fairness gap for clustering and derive a covariance-based surrogate for this gap. The surrogate leads to a gradient-based algorithm that optimizes the clustering cost with a fairness penalty and supports parallel per-instance assignment updates. This makes the method scalable to large data with many subgroups. We call the proposed algorithm COVA-FC (***Cova**riance-based **F**air **C**lustering*). We establish a first-order stationarity guarantee for the alternating updates of COVA-FC under suitable parameter schedules. Beyond subgroup fairness, we show that subgroup fairness alone may fail to control marginal fairness (fairness for each sensitive attribute), and extend COVA-FC to jointly control subgroup and marginal fairness with little additional computation.

The main contributions of this work are:

- ◇ We develop a subgroup-fair clustering algorithm under multiple sensitive attributes that is numerically stable and easily scalable to achieve competitive cost-fairness trade-offs.
- ◇ We propose a way to ensure subgroup and marginal fairness simultaneously without much additional computation burden.
- ◇ We conduct extensive numerical studies to illustrate the advantages of COVA-FC over existing fair clustering algorithms.

2 SUBGROUP-FAIRNESS GAP AND ITS RELAXATION

Let $\{(x_i, \boldsymbol{s}_i)\}_{i=1}^n$ be a given set of instances, where each $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector and $\boldsymbol{s}_i \in \{0,1\}^q$ is a vector of q binary sensitive attributes. For each subgroup vector $\boldsymbol{s} = (s_1, \ldots, s_q)^\top \in \{0,1\}^q$, s_j denotes the value of its j-th sensitive attribute. Let $n_{\boldsymbol{s}} := \sum_{i=1}^n \mathbb{I}(\boldsymbol{s}_i = \boldsymbol{s})$. For $N \in \mathbb{N}$, we write $[N] := \{1, \ldots, N\}$. We consider only binary sensitive attributes for notational simplicity, but our algorithm can be extended to multi-valued sensitive attributes.

Clustering objective. Following prior work on fair clustering (Bera et al., 2019; Li et al., 2020; Esmaeili et al., 2021; Ziko et al., 2021; Zeng et al., 2023; Kim et al., 2025b), we consider the K-means clustering problem. For a given number of clusters $K \in \mathbb{N}$, let $\mu = (\mu_1, \ldots, \mu_K)^\top$ with $\mu_k \in \mathbb{R}^d, k \in [K]$ denote the cluster centroids, and let $\mathbf{A} = (A_1, \ldots, A_n)$ with $A_i \in [K], i \in [n]$ denote the (hard) cluster assignment vector. We minimize the *clustering cost*, defined as the average squared distance to the assigned centroids: $\min_{\mu,\mathbf{A}} \frac{1}{n}\sum_{i=1}^n L_i(\mu, \mathbf{A})$, where $L_i(\mu, \mathbf{A}) := \sum_{k=1}^K \mathbb{I}(A_i = k)\|x_i - \mu_k\|^2$ is the cost for the i-th instance. For soft cluster assignment (see Section 2.3), we also use $\mathbf{A}$, with each $A_i \in \mathbb{S}^K$, where $\mathbb{S}^K$ is the probability simplex: $\mathbb{S}^K := \{a = (a_1, \ldots, a_K)^\top \in \mathbb{R}^K : a_k \geq 0,\ \forall k \in [K],\ \sum_{k=1}^K a_k = 1\}$.

2.1 Subgroup-Fairness gap

Our primary goal is to find the cluster centroids and cluster assignment $(\mu, \mathbf{A})$ that achieve (i) not only a low clustering cost $\frac{1}{n}\sum_i L_i(\mu, \mathbf{A})$, (ii) but also fairness for multiple sensitive attributes.

For each $k \in [K]$, define the overall assignment probability as $p_k(\mathbf{A}) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i = k)$. For each subgroup $\boldsymbol{s} \in \{0,1\}^q$ with $n_{\boldsymbol{s}} > 0$, define the conditional assignment probability in subgroup $\boldsymbol{s}$ as $p_{k|\boldsymbol{s}}(\mathbf{A}) := \frac{1}{n_{\boldsymbol{s}}}\sum_{i:\boldsymbol{s}_i=\boldsymbol{s}} \mathbb{I}(A_i = k)$. Let $\pi_{\boldsymbol{s}} := n_{\boldsymbol{s}}/n$ denote the subgroup proportion. Following Kearns et al. (2018), we define the following subgroup-fairness gap for clustering.

Definition 2.1 (Subgroup-fairness gap)**.** The *subgroup-fairness gap* of $\mathbf{A}$ is defined as

$$\Delta(\mathbf{A}) := \max_{\boldsymbol{s}\in\{0,1\}^q,\, n_{\boldsymbol{s}}>0} \pi_{\boldsymbol{s}} \sum_{k=1}^K \left| p_k(\mathbf{A}) - p_{k|\boldsymbol{s}}(\mathbf{A}) \right|. \tag{1}$$

We say $\mathbf{A}$ is *subgroup-fair* at level ε if $\Delta(\mathbf{A}) \leq \varepsilon$. The weight $\pi_{\boldsymbol{s}}$ reflects the size of subgroup $\boldsymbol{s}$, and prevents tiny subgroups from dominating the gap Δ. Without it, we may be unable to reduce the subgroup-fairness gap below a certain level depending on the tiny subgroups. See Appendix A.1 for an example. However, this weighting may also allow larger disparities within small subgroups. In Appendix E.3, we introduce a generalized fairness gap with weights $\pi_{\boldsymbol{s}}^\alpha$, $\alpha \in [0, 1]$, and examine the resulting trade-off between fairness for small and large subgroups.

Subgroup-fair clustering objective. Based on the subgroup-fairness gap defined in Definition 2.1, a natural goal is to minimize the clustering cost under the subgroup fairness constraint:

$$\min_{\mu,\mathbf{A}} \frac{1}{n}\sum_{i=1}^n L_i(\mu, \mathbf{A}) \quad \text{subject to} \quad \Delta(\mathbf{A}) \leq \varepsilon. \tag{2}$$

A common way to handle such a constrained optimization is to solve a penalized problem, i.e., minimize $\frac{1}{n}\sum_i L_i(\mu, \mathbf{A}) + \lambda\Delta(\mathbf{A})$ where $\lambda > 0$ is a Lagrangian multiplier.

2.2 CR($\mathbf{A}$): A Covariance-based surrogate for $\Delta(\mathbf{A})$

Motivation: challenges in using Δ as the fairness constraint. Directly using $\Delta(\mathbf{A})$ as the fairness penalty poses a critical computational challenge. This is because, unlike the standard K-means objective, the penalized objective is not separable in $\mathbf{A}$. In K-means, the clustering cost $\frac{1}{n}\sum_{i=1}^n L_i(\mu, \mathbf{A})$ decomposes over $i \in [n]$, so each assignment A_i can be updated in parallel given μ. In contrast, $\Delta(\mathbf{A})$ is not separable because the sums over instances appear inside the absolute value of the difference between the overall assignment proportion $p_k(\mathbf{A}) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i = k)$ and the subgroup-conditional assignment proportion $p_{k|\boldsymbol{s}}(\mathbf{A}) = \frac{1}{n_{\boldsymbol{s}}}\sum_{i:\boldsymbol{s}_i=\boldsymbol{s}} \mathbb{I}(A_i = k)$. This dependence complicates independent parallel assignment updates. To address the problem, we introduce a surrogate that rewrites $\Delta(\mathbf{A})$ in a more optimization-friendly form.

Proposed surrogate fairness gap. We now define the *covariance-based* surrogate. Intuitively, a clustering is unfair to a subgroup if subgroup membership is predictive of the cluster assignment. We therefore measure subgroup unfairness through the statistical dependence between membership and assignment.

For each subgroup membership $\boldsymbol{s} \in \{0,1\}^q$, define $c_{i\boldsymbol{s}} := 2\mathbb{I}(\boldsymbol{s}_i = \boldsymbol{s}) - 1$ and $\bar{c}_{\boldsymbol{s}} := \frac{1}{n}\sum_{i=1}^n c_{i\boldsymbol{s}} = 2\pi_{\boldsymbol{s}} - 1$. Let $\mathrm{C}(\mathbf{A}, \boldsymbol{s}, \beta) := \frac{1}{2n}\sum_{i=1}^n \left(\sum_{k=1}^K \beta_k \mathbb{I}(A_i = k)\right)(c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}})$, and define the ***Cova**Riance-based (CR) subgroup-fairness gap* as

$$\mathrm{CR}(\mathbf{A}) := \max_{\boldsymbol{s}\in\{0,1\}^q,\, n_{\boldsymbol{s}}>0} \max_{\beta\in\mathcal{B}_\infty} \mathrm{C}(\mathbf{A}, \boldsymbol{s}, \beta), \quad \text{where} \quad \mathcal{B}_\infty := \{\beta \in \mathbb{R}^K : \|\beta\|_\infty \le 1\}. \tag{3}$$

For a given subgroup membership $\boldsymbol{s} \in \{0,1\}^q$, the term $(c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}})$ is the centered subgroup sign, and $\sum_{k=1}^K \beta_k \mathbb{I}(A_i = k)$ is a signed score assigned to the cluster label A_i with $\beta_k \in [-1, 1]$. Hence, $\mathrm{C}(\mathbf{A}, \boldsymbol{s}, \beta)$ is the empirical covariance between the signed cluster assignment and the centered membership for subgroup s. Maximizing over $\beta \in \mathcal{B}_\infty$ chooses the sign pattern that aligns best with the imbalance in subgroups, and maximizing over s selects the most violated subgroup. Therefore, $\mathrm{CR}(\mathbf{A})$ measures the largest linear dependence between the cluster assignment $\mathbf{A}$ and the subgroup membership. See the middle panel of Figure 1 for an illustration.

Theorem 2.1 (Equivalence of Δ and CR)**.** *For any hard cluster assignment* $\mathbf{A}$*,* $\Delta(\mathbf{A}) = \mathrm{CR}(\mathbf{A})$.

The proof is deferred to Appendix B.1.

2.3 Optimization challenges of $\mathrm{CR}(\mathbf{A})$ and a continuous relaxation

Although $\mathrm{CR}(\mathbf{A})$ is exactly equivalent to $\Delta(\mathbf{A})$, it remains difficult to minimize directly in practice. The difficulty comes from two discrete components: the maximization over subgroups $\boldsymbol{s} \in \{0,1\}^q$ and the hard cluster assignment $\mathbf{A} \in [K]^n$. Even a small change in $\mathbf{A}$ can switch the most violated subgroup abruptly. To improve optimization stability, we relax both the hard cluster assignment and the discrete maximization over subgroups into continuous ones.

(1) Relaxation of the cluster assignment A. We first replace the hard cluster assignment vector with a soft cluster assignment vector. Specifically, we let $\mathbf{A} = (A_1, \ldots, A_n)^\top \in [0,1]^{n\times K}$, where each row $A_i = (A_{i1}, \ldots, A_{iK})^\top$ lies in the probability simplex $\mathbb{S}^K$. Here, A_{ik} represents the assignment probability of x_i to cluster k. Hard assignment is included as a special case in which each A_i is one-hot. Under this relaxation, we replace $\mathbb{I}(A_i = k)$ by A_{ik} in the covariance term, and rewrite the clustering cost as $L_i(\mu, \mathbf{A}) = \sum_{k=1}^K A_{ik}\|x_i - \mu_k\|^2$. This makes both the clustering loss and the covariance term differentiable with respect to $\mathbf{A}$, which enables gradient-based updates.

(2) Relaxation of the worst-case subgroup. We next relax the discrete maximization over subgroups. Let $\mathcal{S}_+ := \{\boldsymbol{s} \in \{0,1\}^q : n_{\boldsymbol{s}} > 0\}$ be the set of observed subgroups and let $M_+ := |\mathcal{S}_+|$. Instead of selecting a single observed subgroup, we relax this discrete choice by allowing a convex combination of subgroup indicators through a weight vector $\mathbf{v} \in \mathbb{S}^{M_+}$. Since the objective is linear in $\mathbf{v}$, its maximum over the simplex is attained at a vertex, preserving the original worst-subgroup maximization. This yields the continuous surrogate

$$\overline{\mathrm{CR}}(\mathbf{A}) := \max_{\mathbf{v}\in\mathbb{S}^{M_+}} \max_{\beta\in\mathcal{B}_\infty} \frac{1}{2n}\sum_{i=1}^n \left(\sum_{k=1}^K \beta_k A_{ik}\right)\mathbf{v}^\top(c_i - \bar{c}), \tag{4}$$

where $c_i := (c_{i\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{S}_+}$ and $\bar{c} := \frac{1}{n}\sum_{i=1}^n c_i$. The following theorem shows that $\overline{\mathrm{CR}}(\mathbf{A})$ is a continuous relaxation of $\Delta(\mathbf{A})$.

Theorem 2.2 (Relationship: Δ and $\overline{\mathrm{CR}}$)**.** *For any hard cluster assignment* $\mathbf{A}$*,* $\Delta(\mathbf{A}) = \overline{\mathrm{CR}}(\mathbf{A})$.

The equivalence holds for both hard and soft assignment, with $\mathbb{I}(A_i = k)$ replaced by A_{ik}. Since the original fairness gap Δ is defined for hard assignment, we convert soft assignment to hard assignment using argmax rounding. Proposition B.1 bounds the difference between the fairness gaps before and after rounding. Proofs of the equivalence and the rounding bound are provided in Appendix B.1 and Appendix B.2, respectively.

3 Considering marginal fairness

The subgroup-fairness surrogate introduced in the previous section provides a tractable objective for handling multiple sensitive attributes. However, it evaluates fairness only at the finest subgroup

level and may therefore provide loose control over marginal fairness. This can be problematic when marginal fairness is also required, even when the subgroup-fairness gap is sufficiently small. See Appendix A.2 for a counterexample and Section 5.3 for empirical evidence. To address this limitation, we modify the subgroup-fairness gap to enforce subgroup and marginal fairness simultaneously.

3.1 Generalization of the subgroup-fairness gap

To account for marginal fairness, we consider groups formed by fixing some of the sensitive attributes. A *subgroup-subset* $W \subseteq \{0,1\}^q$ specifies which full subgroups are included in a group. Let $n_W := \sum_{i=1}^n \mathbb{I}(\boldsymbol{s}_i \in W)$ and $\pi_W := n_W/n$. For a hard assignment $\mathbf{A}$ and $n_W > 0$, define the conditional cluster proportion $p_{k|W}(\mathbf{A}) := \frac{1}{n_W} \sum_{i:\, \boldsymbol{s}_i \in W} \mathbb{I}(A_i = k)$ and the weighted fairness gap as,

$$g(\mathbf{A}; W) := \pi_W \sum_{k=1}^{K} \big| p_k(\mathbf{A}) - p_{k|W}(\mathbf{A}) \big|.$$

Let $\mathcal{W}$ be a collection of subgroup-subsets, with $n_W > 0$ for at least one $W \in \mathcal{W}$. Define the fairness gap over this collection as

$$\Delta(\mathbf{A}; \mathcal{W}) := \max_{W \in \mathcal{W}:\, n_W > 0} g(\mathbf{A}; W). \tag{5}$$

3.2 Incorporating marginal and higher-order fairness

The *marginal subgroup* $W_{j,a} := \{\boldsymbol{s} \in \{0,1\}^q : s_j = a\}$, for $j \in [q]$ and $a \in \{0,1\}$, contains all subgroups with $s_j = a$. We collect these groups into $\mathcal{W}^{(1)} := \{W_{j,a} : j \in [q],\ a \in \{0,1\}\}$. The assignment $\mathbf{A}$ is *marginally fair* if $\Delta(\mathbf{A}; \mathcal{W}^{(1)})$ is small.

We also extend this notion of fairness to *higher-order marginal subgroups*, which are defined by jointly fixing several sensitive attributes. For $r \in [q]$ and $\theta \subseteq [q]$ with $|\theta| = r$, let $\boldsymbol{s}_\theta$ denote the vector of coordinates of $\boldsymbol{s}$ indexed by θ, listed in increasing index order. For $a \in \{0,1\}^r$, define the groups and their collection by

$$W_{\theta,a} := \{\boldsymbol{s} \in \{0,1\}^q : \boldsymbol{s}_\theta = a\}, \qquad \mathcal{W}^{(r)} := \{W_{\theta,a} : \theta \subseteq [q],\ |\theta| = r,\ a \in \{0,1\}^r\}.$$

The assignment $\mathbf{A}$ is *r-th order marginally fair* if $\Delta(\mathbf{A}; \mathcal{W}^{(r)})$ is small.

The choices $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(q)}$ correspond to marginal and subgroup fairness, respectively. Indeed, $\Delta(\mathbf{A}; \mathcal{W}^{(q)}) = \Delta(\mathbf{A})$. To control subgroup fairness together with marginal fairness up to order r, choose $\mathcal{W} = \mathcal{W}^{(q)} \cup \bigcup_{\ell=1}^{r} \mathcal{W}^{(\ell)}$. With this choice of $\mathcal{W}$, we refer to $\Delta(\mathbf{A}; \mathcal{W})$ as the *subgroup-marginal-fairness gap* up to order r.

3.3 Surrogate for subgroup-marginal-fairness gap

We now extend the surrogate in Section 2.3 to the fairness gap $\Delta(\mathbf{A}; \mathcal{W})$ for a given collection $\mathcal{W}$. The choice of $\mathcal{W}$ determines which groups are controlled, while the surrogate remains the same. Write $\mathcal{W} = \{W_1, \ldots, W_M\}$. Let $\mathcal{M}_+ := \{m \in [M] : n_{W_m} > 0\}$ index the groups containing observed instances, and let $M_+ := |\mathcal{M}_+|$. For each group W_m, define its membership encoding and empirical mean by $c_{im} := 2\mathbb{I}(\boldsymbol{s}_i \in W_m) - 1$, $\bar{c}_m := \frac{1}{n} \sum_{i=1}^n c_{im}$. Collect the coordinates for observed groups into $c_i := (c_{im})_{m \in \mathcal{M}_+}$ and $\bar{c} := (\bar{c}_m)_{m \in \mathcal{M}_+}$.

Following Section 2.2, define for hard assignment

$$\mathrm{CR}(\mathbf{A}; \mathcal{W}) := \max_{m \in \mathcal{M}_+} \max_{\beta \in \mathcal{B}_\infty} \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \beta_k \mathbb{I}(A_i = k) \right) (c_{im} - \bar{c}_m). \tag{6}$$

The continuous relaxation replaces $\mathbb{I}(A_i = k)$ with A_{ik} and uses simplex weights for group selection:

$$\overline{\mathrm{CR}}(\mathbf{A}; \mathcal{W}) := \max_{\mathbf{v} \in \mathbb{S}^{M_+}} \max_{\beta \in \mathcal{B}_\infty} \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \beta_k A_{ik} \right) \mathbf{v}^\top (c_i - \bar{c}). \tag{7}$$

Theorem 3.1 (Equivalence for fairness over $\mathcal{W}$)**.** *For any hard cluster assignment* $\mathbf{A}$*,*

$$\Delta(\mathbf{A}; \mathcal{W}) = \mathrm{CR}(\mathbf{A}; \mathcal{W}) = \overline{\mathrm{CR}}(\mathbf{A}; \mathcal{W}). \tag{8}$$

Theorem 3.1 includes the subgroup-only case as a special case when $\mathcal{W}$ is chosen as $\mathcal{W}^{(q)}$, the collection of all subgroups. The equivalence also extends to soft assignment by replacing $\mathbb{I}(A_i = k)$ with A_{ik} in the assignment probabilities and covariance terms. Proposition B.2 bounds the difference between the soft surrogate and the fairness gap after argmax rounding for a general collection $\mathcal{W}$. Proofs of the equivalence and the rounding bound are given in Appendix B.1 and Appendix B.2, respectively.

4 Proposed Algorithm: Cova-FC

Given a collection $\mathcal{W}$ of subgroup-subsets, we learn fair clusters by minimizing

$$\frac{1}{n}\sum_{i=1}^{n} L_i(\mu, \mathbf{A}) + \lambda\, \overline{\mathrm{CR}}(\mathbf{A}; \mathcal{W}) \tag{9}$$

over the centroids μ and soft assignment $\mathbf{A}$, with $A_i \in \mathbb{S}^K$. $\lambda \geq 0$ controls the trade-off between clustering cost and fairness. The covariance term in Eq. (7) depends on $(\mathbf{v}, \beta)$ only through the outer product $\mathbf{v}\beta^\top$. Let $\mathbf{C} \in \mathbb{R}^{n \times M_+}$ have rows $(c_i - \bar{c})^\top$, and define $\mathcal{U} := \mathrm{conv}\{\mathbf{v}\beta^\top : \mathbf{v} \in \mathbb{S}^{M_+},\ \beta \in \mathcal{B}_\infty\}$, where conv denotes the convex hull. By Lemma B.3, minimizing Eq. (9) is equivalent to

$$\min_{\mathbf{A},\mu} \max_{U \in \mathcal{U}} \widetilde{F}(\mathbf{A}, \mu, U), \quad \text{where} \quad \widetilde{F}(\mathbf{A}, \mu, U) := \frac{1}{n}\sum_{i=1}^{n} L_i(\mu, \mathbf{A}) + \frac{\lambda}{2n}\langle \mathbf{C}^\top \mathbf{A}, U\rangle_F. \tag{10}$$

Optimization and theoretical guarantees. We update $\mathbf{A}$ and then μ by projected gradient descent, holding the other variables fixed at each step. We then update U by projected gradient ascent using the updated assignment and centroids and the regularized objective

$$\widetilde{F}_t(\mathbf{A}, \mu, U) = \widetilde{F}(\mathbf{A}, \mu, U) - \frac{c_t}{2}\|U\|_F^2, \quad c_t \downarrow 0. \tag{11}$$

For $c_t > 0$, the quadratic term makes $\widetilde{F}_t$ strongly concave in U. These updates follow the block alternating proximal gradient (BAPG) method of Xu et al. (2023). Once U and the centroids are fixed, the covariance term decomposes over instances. Thus, each assignment row can be updated and projected onto $\mathbb{S}^K$ independently, enabling parallel assignment updates. The simplex projection is detailed in Appendix C.2, while the update formulas and full pseudocode are provided in Appendix C.1.

Under the conditions of Theorem B.5, Cova-FC obtains a first-order ε-stationary point of $\widetilde{F}$ in Eq. (10). For any tolerance $\varepsilon \in (0, 1)$, this requires at most $\mathcal{O}(\varepsilon^{-4})$ iterations. The stationarity measure, assumptions, parameter schedules, and proof are provided in Appendix B.3.

Computational efficiency. In the subgroup-only setting, recall that M_+ is the number of observed subgroups. Each iteration of Cova-FC, including the gradient computations and projections, has complexity $\mathcal{O}(nKd + nK\log K + KM_+\log(KM_+))$. Since $M_+ \leq 2^q$, this is upper bounded by $\mathcal{O}(nKd + nK\log K + K2^q\log(K2^q))$.

Table 1: Comparison of fair clustering methods in terms of computational complexity.

Method	Complexity
Cova-FC ✓	$nKd + nK\log K + K2^q\log(K2^q)$
VFC (Ziko et al., 2021)	$nK(d + 2^q)$
FairKM (Abraham et al., 2020)	$nK(nd + 2^q)$
FCA (Kim et al., 2025b)	$(2^q - 1)n^2(n + Kd)$
FCBC (Esmaeili et al., 2021)	$(nK + 2^qK)^3$
FRA (Bera et al., 2019)	$(nK + 2^q)^3$

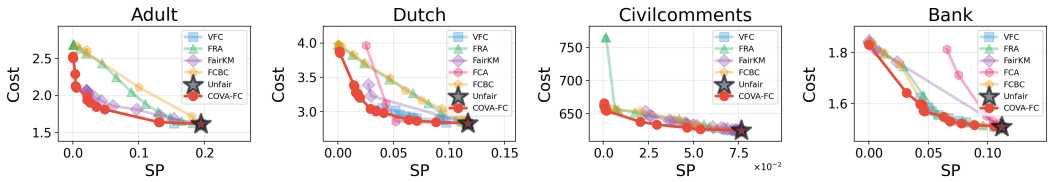


Figure 2: Comparison of trade-offs between subgroup-fairness gap (SP) and cost on (left to right) `Adult`, `Dutch`, `Civilcomments`, and `Bank` datasets. Each point shows the mean over five random seeds.

Table 1 compares the computational complexity of COVA-FC with other baselines. It shows that COVA-FC avoids the cubic dependence on n in LP-based methods. Compared with VFC, whose subgroup-dependent term $nK2^q$ grows with both the number of instances and subgroups, COVA-FC has the corresponding term $K2^q \log(K2^q)$, independent of n, while its $nK \log K$ projection cost is independent of the number of subgroups. This computational advantage is also observed empirically on the large-scale `ACSIncome` dataset (Section 5.4), with further computational details in Appendix C.3.

5 EXPERIMENTS

This section presents the results of our experimental analyses. Section 5.2 reports cost-fairness trade-offs on four benchmark datasets (`Adult`, `Dutch`, `Bank`, `Civilcomments`), confirming favorable trade-offs and computational efficiency. Section 5.3 evaluates the improvement in marginal fairness from the extension on `Communities` ($q = 18$), where the number of subgroups is large and many subgroups are sparse. Section 5.4 studies scalability to large n on `ACSIncome` ($n \approx 1.66$M) and to high-dimensional features on an image dataset `CelebA`. Finally, Section 5.5 analyzes the practical effects of the two continuous relaxations introduced in Section 2.3.

5.1 SETTINGS

Datasets and baselines. We use seven benchmark datasets, including five tabular datasets (`Adult`, `Dutch`, `Bank`, `Communities`, and `ACSIncome`), one text dataset (`Civilcomments`), and one image dataset (`CelebA`). See Table 5 in Appendix D.1 for detailed descriptions of these datasets.

We compare COVA-FC with well-known existing methods: FRA (Bera et al., 2019), FairKM (Abraham et al., 2020), FCBC (Esmaeili et al., 2021), VFC (Ziko et al., 2021) and FCA (Kim et al., 2025b). Due to runtime constraints in our experimental environment, we evaluate FCA only on `Dutch` and `Bank`. Details are provided in Appendix D.1.

Performance metrics. We use the hard assignment $\hat{A}_i = \arg\max_k A_{ik}$ for performance evaluation (abbreviated A_i in this section). Clustering performance is measured by the K-means clustering cost $\frac{1}{n}\sum_{i=1}^{n} L_i(\mu, \mathbf{A})$. For fairness, we assess the fairness gap $\Delta(\mathbf{A}; \mathcal{W})$ defined in Eq. (5). We denote the subgroup-fairness gap as SP $:= \Delta(\mathbf{A})$, and the r-th order marginal-fairness gap as $\text{MP}^{(r)} := \Delta(\mathbf{A}; \mathcal{W}^{(r)})$ defined in Section 3.2. Note that Sections 2 and 3 define Δ for binary sensitive attributes ($a \in \{0, 1\}$), but they can be extended easily for a multi-valued sensitive attribute with m categories by replacing $a \in \{0, 1\}$ with $a \in \{0, 1, \ldots, m-1\}$.

We use the gap-based metrics as our primary fairness measures because the Balance metric can collapse to zero when one subgroup has no instances in a cluster while another compared subgroup is present. We discuss this issue further and report additional Balance-based results in Appendix D.2.

Implementation details. For the convergence criterion, we terminate the algorithm when the objective in Eq. (9) changes by less than 2×10^{-5} for 15 iterations, with a maximum of 10,000 iterations. In the main subgroup-fairness experiments, we construct $\mathcal{W}$ using all full subgroups only, and fix $K = 10$ for all experiments following prior work (Kim et al., 2025b; Ziko et al., 2021). For the cost-fairness trade-off experiments in Figure 2, we run five random seeds. Additional details are provided in Appendix D.1.

5.2 PERFORMANCE COMPARISON

Cost-fairness trade-off. By examining the cost-fairness trade-off, we quantify the level of subgroup fairness achieved for a given clustering cost. Figure 2 shows the Pareto-front trade-offs between clustering cost and subgroup-fairness gap (SP). Since the variability across seeds is small, we report only the mean results in Figure 2 and provide the mean $\pm$ standard deviation at the lowest-SP points in Table 6 of Appendix D.2. COVA-FC attains lower SP than the baselines in most cases. FCBC (Esmaeili et al., 2021) and VFC (Ziko et al., 2021) often fail to attain high fairness levels due to numerical overflow, whereas COVA-FC works over a wide range of fairness levels. Additionally, we report the trade-off between $\text{MP}^{(1)}$ and cost in Appendix D.2, where COVA-FC attains lower $\text{MP}^{(1)}$ than the baselines at comparable clustering cost on these four benchmarks, despite optimizing only subgroup fairness. However, subgroup fairness alone may not adequately control marginal fairness when subgroups contain few instances, motivating our marginal-fairness extension evaluated on `Communities` in Section 5.3.

Table 2: Comparison between COVA-FC and VFC in terms of the lowest achievable SP.

	SP (Cost)	
Dataset	VFC (Ziko et al., 2021)	COVA-FC
`Adult`	0.0158 (0.47)	**0.0002** (0.50)
`Dutch`	0.0146 (0.40)	**0.0026** (0.41)
`Civilcomments`	0.0063 (0.86)	**0.0009** (0.90)
`Bank`	0.0354 (0.24)	**0.0000** (0.27)

Comparison with VFC. With L_2-normalized data (as used in Ziko et al. (2021)), COVA-FC achieves lower SP than VFC on all four datasets, with modest increases in clustering cost (Table 2).

5.3 MARGINAL FAIRNESS EXTENSION

We evaluate the subgroup-marginal-fairness gap (Section 3) on the `Communities` dataset (Redmond & Baveja, 2002) with $q = 18$ sensitive attributes (1,180 subgroups), where many subgroups contain only a few instances. We compare the subgroup-only and subgroup-marginal variants of COVA-FC. The latter includes first- and second-order marginal subgroups with full subgroups. The subgroup-marginal variant achieves better $\text{MP}^{(1)}$ and $\text{MP}^{(2)}$ than the subgroup-only variant, with a mild degradation in SP. Figure 3 reports results averaged over five random splits.

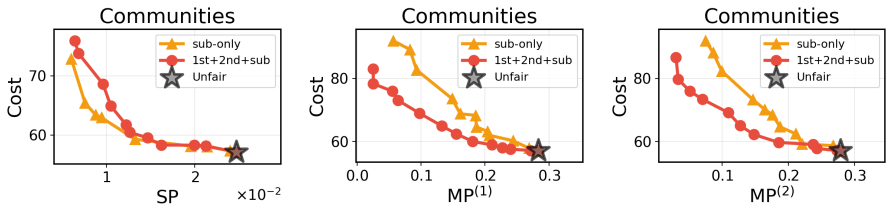


Figure 3: Comparison on `Communities` between subgroup-only COVA-FC and subgroup-marginal COVA-FC (1st+2nd-order marginal + subgroup): (left) SP, (middle) $\text{MP}^{(1)}$, (right) $\text{MP}^{(2)}$.

5.4 COMPUTATIONAL EFFICIENCY AND SCALABILITY

We evaluate the computational efficiency and scalability of COVA-FC, along with its applicability to image-feature data. We also examine the benefit of GPU parallelization through a runtime comparison of loop and matrix implementations of the assignment update in Appendix D.4.

Computation time. Motivated by the computational complexity analysis in Table 1, the results in Table 3 show that COVA-FC achieves the lowest total runtime among the compared methods on the `Dutch` dataset.

Table 3: Comparison of computation time on `Dutch` dataset.

Method	Total time (sec)
COVA-FC ✓	**39.07**
VFC (Ziko et al., 2021)	48.25
FairKM (Abraham et al., 2020)	292.26
FCA (Kim et al., 2025b)	226.42
FCBC (Esmaeili et al., 2021)	179.26
FRA (Bera et al., 2019)	64.29

Scalability to large-scale and image data. On `ACSIncome` (Ding et al., 2021) ($n = 1{,}664{,}500$, 16 subgroups), COVA-FC runs $3.22\times$ faster than VFC, the fastest baseline in Table 3. A detailed comparison is reported in Table 9 in Appendix D.4. On `CelebA` (Liu et al., 2015), we use 512-dimensional features extracted from a pretrained ResNet-18 (He et al., 2016). As shown in Figure 8, COVA-FC improves subgroup fairness as the penalty increases, supporting its applicability across various data modalities.

5.5 ABLATION STUDY

Sensitivity and optimization. We examine sensitivity to λ, K, and initialization, as well as the effects of step-size decay. On `Adult`, larger λ generally lowers SP at higher clustering cost. Decaying step sizes improve optimization progress in the tested settings. This trade-off persists across the tested values of K, while final SP is robust to initialization across the tested values of λ. Detailed experimental settings and results are provided in Appendix E.4.

Effects of the continuous relaxations. To examine the effect of each continuous relaxation, we consider four variants of the surrogate depending on which discrete component is continuously relaxed: CR (both hard), $\mathrm{CR}_{\mathbf{v}}$ (soft $\mathbf{v}$, hard $\mathbf{A}$), $\mathrm{CR}_{\mathbf{A}}$ (hard $\mathbf{v}$, soft $\mathbf{A}$), and $\overline{\mathrm{CR}}$ (both soft).

In Appendix E.1, we observe that relaxing the assignment $\mathbf{A}$ is crucial for stability: hard-assignment variants often fail to reduce the fairness gap, whereas $\mathrm{CR}_{\mathbf{A}}$ achieves near-zero gaps by relaxing the hard cluster assignment CR (see Table 10). This is also reflected in the loss curves, where $\mathrm{CR}_{\mathbf{v}}$ oscillates but $\overline{\mathrm{CR}}$ decreases steadily (see Figures 9 and 10). Relaxing the worst-subgroup selection improves efficiency by replacing $\mathrm{CR}_{\mathbf{A}}$ with $\overline{\mathrm{CR}}$, reducing runtime while preserving comparable cost and SP (see Table 11).

6 CONCLUDING REMARKS

In this paper, we proposed COVA-FC, a novel algorithm for subgroup-fair clustering with multiple sensitive attributes. Our covariance-based surrogate exactly matches the subgroup-fairness gap and admits efficient gradient-based optimization with parallel assignment updates. We further establish a first-order stationarity guarantee for the resulting alternating updates. The framework additionally extends to marginal and higher-order subgroups via a user-specified collection $\mathcal{W}$. Across benchmark datasets, COVA-FC achieves favorable cost-fairness trade-offs with improved computational efficiency and scalability, when compared with existing baselines.

It would be interesting to extend our proposed framework in several ways. First, the framework would be naturally extendable for clustering paradigms other than K-means such as hierarchical clustering and mixture models, where the cluster assignment matrix $\mathbf{A}$ admits a similar soft relaxation. Second, the covariance-based surrogate can be applied to constraints other than fairness: any constraint expressible as "low statistical dependence between $\mathbf{A}$ and a side-information vector" fits the same template. An example would be clustering with size constraints (Höppner & Klawonn, 2008; Zhu et al., 2010). We leave a systematic study of these extensions as future work.

AI USE STATEMENT

We used generative AI tools to assist with debugging experiment code, checking mathematical proofs, and editing the manuscript for clarity. All AI-assisted work was reviewed and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies statistical fairness in clustering under multiple sensitive attributes. While COVA-FC can reduce disparities in cluster assignment across specified groups, statistical fairness at the clustering level does not by itself guarantee fair outcomes in downstream applications. The choice of sensitive attributes and subgroup definitions may also affect the resulting fairness assessment. A small size-weighted fairness gap does not rule out disparities in rare subgroups. To give greater emphasis to small subgroups, we provide a tunable weighting parameter α and examine its effects in Appendix E.3. Therefore, the proposed method should be used with domain-specific evaluation and auditing.

REPRODUCIBILITY STATEMENT

Section 5.1 describes the experimental setup, evaluation metrics, and stopping criterion. Appendix D.1 provides details on the datasets, preprocessing, baselines, and implementation settings. Appendix C.1 presents the optimization updates and full pseudocode. Appendix B.3 states the convergence assumptions and parameter schedule and provides the proof.

REFERENCES

Savitha Sam Abraham, Deepak P, and Sowmya S. Sundaram. Fairness in clustering with multiple sensitive attributes. In *Proceedings of the 23rd International Conference on Extending Database Technology (EDBT)*, pp. 287–298, 2020. doi: 10.5441/002/edbt.2020.26. URL https://openproceedings.org/2020/conf/edbt/paper_148.pdf.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 120–129. PMLR, 2019. URL https://proceedings.mlr.press/v97/agarwal19d.html.

Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In *International Conference on Machine Learning*, volume 97, pp. 405–413. PMLR, 2019. URL https://proceedings.mlr.press/v97/backurs19a.html.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. URL https://doi.org/10.24432/C5XW20.

Suman Kalyan Bera, Deeparnab Chakrabarty, Nicolas J. Flores, and Maryam Negahbani. Fair algorithms for clustering. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fc192b0c0d270dbf41870a63a8c76c2f-Abstract.html.

Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. The price of fairness. *Operations Research*, 59(1):17–31, jan 2011. ISSN 0030-364X. doi: 10.1287/opre.1100.0865. URL https://doi.org/10.1287/opre.1100.0865.

Avrim Blum and Thodoris Lykouris. Advancing Subgroup Fairness via Sleeping Experts. In Thomas Vidick (ed.), *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 55:1–55:24, Dagstuhl,

Germany, 2020. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-134-4. doi: 10.4230/LIPIcs.ITCS.2020.55. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2020.55.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. An overview of fairness in clustering. *IEEE Access*, 9:130698–130720, 2021. doi: 10.1109/ACCESS.2021.3114099.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5029–5037, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/978fce5bcc4eccc88ad48ce3914124a2-Abstract.html.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7321–7331. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/51cdbd2611e844ece5d80878eb770436-Paper.pdf.

Frances Ding, Moritz Hardt, John P. Miller, and Ludwig Schmidt. Retiring Adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6478–6490, Red Hook, NY, USA, 2021. Curran Associates Inc. URL https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2791–2801, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/83cdcec08fbf90370fcf53bdd56604ff-Abstract.html.

Seyed Esmaeili, Brian Brubach, Aravind Srinivasan, and John Dickerson. Fair clustering under a bounded cost. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14345–14357. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/781877bda0783aac5f1cf765c128b437-Paper.pdf.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair Diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. URL https://arxiv.org/abs/2302.10893.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2415–2423, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/dc4c44f624d600aa568390f1f1104aa0-Abstract.html.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017. doi: 10.1609/aimag.v38i3.2741.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Frank Höppner and Frank Klawonn. Clustering with size constraints. In *Computational Intelligence Paradigms*, volume 137 of *Studies in Computational Intelligence*, pp. 167–180. Springer, Berlin, Heidelberg, 2008. ISBN 978-3-540-79473-8. doi: 10.1007/978-3-540-79474-5_8. URL https://doi.org/10.1007/978-3-540-79474-5_8.

Francois Hu, Philipp Ratz, and Arthur Charpentier. A sequentially fair mechanism for multiple sensitive attributes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12502–12510, mar 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i11.29143. URL http://dx.doi.org/10.1609/aaai.v38i11.29143.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR, 2018. URL https://proceedings.mlr.press/v80/kearns18a.html.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 100–109, 2019.

Kunwoong Kim, Ilsang Ohn, Sara Kim, and Yongdai Kim. Slide: A surrogate fairness constraint to ensure fairness consistency. *Neural Networks*, 154:441–454, 2022. ISSN 0893-6080. doi: 10.1016/j.neunet.2022.07.027. URL https://www.sciencedirect.com/science/article/pii/S0893608022002891.

Kunwoong Kim, Insung Kong, Jongjin Lee, Minwoo Chae, Sangchul Park, and Yongdai Kim. Fairness through matching. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=dHljjaNHh1.

Kunwoong Kim, Jihu Lee, Sangchul Park, and Yongdai Kim. Fair clustering via alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30363–30399. PMLR, 2025b. URL https://proceedings.mlr.press/v267/kim25t.html.

Kunwoong Kim, Kyungseon Lee, Jihu Lee, Dongyoon Yang, and Yongdai Kim. Doubly-regressing approach for subgroup fairness. In *International Conference on Learning Representations*, volume 2026, pp. 137532–137558, 2026.

Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. Guarantees for spectral clustering with fairness constraints. In *International conference on machine learning*, volume 97, pp. 3458–3467. PMLR, 2019. URL https://proceedings.mlr.press/v97/kleindessner19b.html.

Yuheng Lai and Leying Guan. FairICP: Encouraging equalized odds via inverse conditional permutation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32228–32245. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/lai25b.html.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, 12(3):e1452, 2022. doi: 10.1002/widm.1452. URL https://doi.org/10.1002/widm.1452.

Peizhao Li, Han Zhao, and Hongfu Liu. Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9070–9079, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Deep_Fair_Clustering_for_Visual_Learning_CVPR_2020_paper.pdf.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, December 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489. URL https://doi.org/10.1109/TIT.1982.1056489.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021. doi: 10.1145/3457607.

Mathieu Molina and Patrick Loiseau. Bounding and approximating intersectional fairness through marginal fairness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16796–16807. Curran Associates, Inc., 2022. URL https://arxiv.org/abs/2206.05828.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014. ISSN 0167-9236. doi: 10.1016/j.dss.2014.03.001. URL https://www.sciencedirect.com/science/article/pii/S016792361400061X.

Ariadna Quattoni, Xavier Carreras, Michael Collins, and Trevor Darrell. An efficient projection for l1, ∞ regularization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 857–864, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553484. URL https://doi.org/10.1145/1553374.1553484.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002. doi: 10.1016/S0377-2217(01)00264-8. URL https://doi.org/10.1016/S0377-2217(01)00264-8.

Weiran Wang and Miguel Á. Carreira-Perpiñán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application, 2013. URL https://arxiv.org/abs/1309.1541.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. FairGAN: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, 2018. doi: 10.1109/BigData.2018.8622525. URL https://ieeexplore.ieee.org/document/8622525/.

Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, 201:635–706, 2023. doi: 10.1007/s10107-022-01919-z. URL https://doi.org/10.1007/s10107-022-01919-z.

Pengxin Zeng, Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, and Xi Peng. Deep fair clustering via maximizing and minimizing mutual information: Theory, algorithm and metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23986–23995, 2023.

Shunzhi Zhu, Dingding Wang, and Tao Li. Data clustering with size constraints. *Knowledge-Based Systems*, 23(8):883–889, 2010. ISSN 0950-7051. doi: 10.1016/j.knosys.2010.06.003. URL https://www.sciencedirect.com/science/article/pii/S095070511000095X.

Imtiaz Masud Ziko, Jing Yuan, Eric Granger, and Ismail Ben Ayed. Variational fair clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11202–11209, 2021. doi: 10.1609/aaai.v35i12.17336. URL https://ojs.aaai.org/index.php/AAAI/article/view/17336.

A MOTIVATING EXAMPLES

A.1 ROLE OF SUBGROUP-SIZE WEIGHTING $\pi_{\boldsymbol{s}}$

We illustrate why the subgroup-size weight $\pi_{\boldsymbol{s}}$ is necessary in the subgroup-fairness gap. The issue is that, without this weight, a tiny subgroup can dominate the worst-case gap regardless of the overall clustering quality.

Consider two clusters and suppose that one subgroup $\boldsymbol{s}^\star$ contains only one instance. Assume that this instance is assigned to cluster 1. Then the cluster distribution within $\boldsymbol{s}^\star$ is $(1, 0)$.

Now suppose that the overall cluster distribution is balanced, namely $(0.5, 0.5)$. The unweighted fairness gap for this subgroup is the total absolute difference between these two distributions:

$$\sum_{k=1}^{2} \left| p_k(\mathbf{A}) - p_{k|\boldsymbol{s}^\star}(\mathbf{A}) \right| = |0.5 - 1| + |0.5 - 0| = 1.$$

Reducing this unweighted fairness gap below 1 requires unbalanced clusters and may increase clustering cost. With the weighted fairness gap used in our definition, the same discrepancy is multiplied by the subgroup proportion $\pi_{\boldsymbol{s}^\star} = 1/n$. Hence the contribution of this subgroup becomes

$$\pi_{\boldsymbol{s}^\star} \sum_{k=1}^{2} \left| p_k(\mathbf{A}) - p_{k|\boldsymbol{s}^\star}(\mathbf{A}) \right| = \frac{1}{n} \times 1 = \frac{1}{n}.$$

Therefore, the weighted fairness gap prevents a tiny subgroup from determining the whole fairness objective.

A.2 SMALL SUBGROUP GAPS AND MARGINAL DISPARITIES

We give a simple example showing that a small subgroup-fairness gap can coexist with a much larger marginal-fairness gap.

Let $q = 8$, $K = 2$, and $n = 2^8 = 256$. Assume that there is exactly one instance in each full subgroup $\boldsymbol{s} \in \{0, 1\}^8$, so $\pi_{\boldsymbol{s}} = 1/256$ for every s. Let $W = \{\boldsymbol{s} \in \{0, 1\}^8 : s_1 = 0\}$, which is the marginal subgroup defined by fixing the first sensitive attribute to 0. Then W contains $2^7 = 128$ subgroups and has proportion $\pi_W = 1/2$. Now assign every instance in W to cluster 1, and every instance outside W to cluster 2. The overall cluster proportions are therefore $(p_1(\mathbf{A}), p_2(\mathbf{A})) = (1/2, 1/2)$.

For each full subgroup s, the conditional cluster distribution is either $(1, 0)$ or $(0, 1)$. Hence the subgroup-fairness gap is

$$\Delta(\mathbf{A}) = \frac{1}{256} \left(\left| \frac{1}{2} - 1 \right| + \left| \frac{1}{2} - 0 \right| \right) = \frac{1}{256}. \tag{12}$$

By contrast, within the marginal subgroup W, all instances are assigned to cluster 1, so $(p_{1|W}(\mathbf{A}), p_{2|W}(\mathbf{A})) = (1, 0)$. Therefore the marginal-fairness gap on W is

$$\pi_W \sum_{k=1}^{2} |p_k(\mathbf{A}) - p_{k|W}(\mathbf{A})| = \frac{1}{2} \left(\left| \frac{1}{2} - 1 \right| + \left| \frac{1}{2} - 0 \right| \right) = \frac{1}{2}. \tag{13}$$

Thus the subgroup-fairness gap is only $1/256$, while the marginal-fairness gap is $1/2$. More generally, under the same construction with arbitrary q, the subgroup-fairness gap is $1/2^q$, which goes to 0 as q increases, whereas the marginal-fairness gap remains $1/2$. This shows that a small subgroup-fairness gap does not imply marginal fairness.

B THEORETICAL STUDIES

B.1 PROOFS

Proof of Theorem 2.1. Fix a subgroup $\boldsymbol{s} \in \{0,1\}^q$ with $n_{\boldsymbol{s}} > 0$. Since $c_{i\boldsymbol{s}} = 2\mathbb{I}(\boldsymbol{s}_i = \boldsymbol{s}) - 1$ and $\bar{c}_{\boldsymbol{s}} = 2\pi_{\boldsymbol{s}} - 1$, we have $\mathbb{I}(\boldsymbol{s}_i = \boldsymbol{s}) - \pi_{\boldsymbol{s}} = \frac{1}{2}(c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}})$.

For each $k \in [K]$, define $a_k(\boldsymbol{s}) := \frac{1}{2n}\sum_{i=1}^n \mathbb{I}(A_i = k)(c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}})$. Then

$$a_k(\boldsymbol{s}) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i = k)\big(\mathbb{I}(\boldsymbol{s}_i = \boldsymbol{s}) - \pi_{\boldsymbol{s}}\big) = \pi_{\boldsymbol{s}}\big(p_{k|\boldsymbol{s}}(\mathbf{A}) - p_k(\mathbf{A})\big).$$

Therefore,

$$\pi_{\boldsymbol{s}}\sum_{k=1}^K \big|p_k(\mathbf{A}) - p_{k|\boldsymbol{s}}(\mathbf{A})\big| = \sum_{k=1}^K |a_k(\boldsymbol{s})| = \max_{\beta\in\mathcal{B}_\infty}\sum_{k=1}^K \beta_k a_k(\boldsymbol{s}),$$

where the maximum is attained by taking $\beta_k = \mathrm{sign}(a_k(\boldsymbol{s}))$. By the definition of $\mathrm{C}(\mathbf{A}, \boldsymbol{s}, \beta)$,

$$\pi_{\boldsymbol{s}}\sum_{k=1}^K \big|p_k(\mathbf{A}) - p_{k|\boldsymbol{s}}(\mathbf{A})\big| = \max_{\beta\in\mathcal{B}_\infty} \mathrm{C}(\mathbf{A}, \boldsymbol{s}, \beta).$$

Taking the maximum over all s with $n_{\boldsymbol{s}} > 0$ gives $\Delta(\mathbf{A}) = \mathrm{CR}(\mathbf{A})$. □

Proof of Theorem 2.2. For a hard assignment $\mathbf{A}$, we identify each assignment with its one-hot representation, so that $A_{ik} = \mathbb{I}(A_i = k)$. For fixed $\beta \in \mathcal{B}_\infty$, define

$$G_{\boldsymbol{s}}(\beta) := \frac{1}{2n}\sum_{i=1}^n \left(\sum_{k=1}^K \beta_k A_{ik}\right)(c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}}), \qquad \boldsymbol{s} \in \mathcal{S}_+,$$

and let $G(\beta) := (G_{\boldsymbol{s}}(\beta))_{\boldsymbol{s}\in\mathcal{S}_+}$.

Since $\mathbf{v} \mapsto \mathbf{v}^\top G(\beta)$ is linear on the probability simplex $\mathbb{S}^{M_+}$, its maximum is attained at a vertex. Hence,

$$\max_{\mathbf{v}\in\mathbb{S}^{M_+}} \mathbf{v}^\top G(\beta) = \max_{\boldsymbol{s}\in\mathcal{S}_+} G_{\boldsymbol{s}}(\beta).$$

Therefore,

$$\begin{aligned}\overline{\mathrm{CR}}(\mathbf{A}) &= \max_{\mathbf{v}\in\mathbb{S}^{M_+}}\max_{\beta\in\mathcal{B}_\infty} \mathbf{v}^\top G(\beta)\\ &= \max_{\beta\in\mathcal{B}_\infty}\max_{\boldsymbol{s}\in\mathcal{S}_+} G_{\boldsymbol{s}}(\beta)\\ &= \max_{\boldsymbol{s}\in\mathcal{S}_+}\max_{\beta\in\mathcal{B}_\infty} G_{\boldsymbol{s}}(\beta) = \mathrm{CR}(\mathbf{A}).\end{aligned}$$

By Theorem 2.1, $\Delta(\mathbf{A}) = \mathrm{CR}(\mathbf{A})$. Combining the two identities gives $\Delta(\mathbf{A}) = \overline{\mathrm{CR}}(\mathbf{A})$, which proves the result. □

Proof of Theorem 3.1. For each $m \in \mathcal{M}_+$, let $r_m(i) := \mathbb{I}(\boldsymbol{s}_i \in W_m)$ and recall that $\pi_{W_m} = n_{W_m}/n$. From the definitions of c_{im} and $\bar{c}_m$, we have $r_m(i) - \pi_{W_m} = \frac{1}{2}(c_{im} - \bar{c}_m)$.

For each $k \in [K]$, define $a_k(m) := \frac{1}{2n}\sum_{i=1}^n \mathbb{I}(A_i = k)(c_{im} - \bar{c}_m)$. Then

$$a_k(m) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(A_i = k)\big(r_m(i) - \pi_{W_m}\big) = \pi_{W_m}\big(p_{k|W_m}(\mathbf{A}) - p_k(\mathbf{A})\big).$$

Therefore,

$$\pi_{W_m}\sum_{k=1}^K \big|p_k(\mathbf{A}) - p_{k|W_m}(\mathbf{A})\big| = \sum_{k=1}^K |a_k(m)|.$$

Since $\mathcal{B}_\infty = \{\beta \in \mathbb{R}^K : \|\beta\|_\infty \le 1\}$,

$$\sum_{k=1}^K |a_k(m)| = \max_{\beta\in\mathcal{B}_\infty}\sum_{k=1}^K \beta_k a_k(m),$$

where the maximum is attained by taking $\beta_k = \mathrm{sign}(a_k(m))$. Substituting the definition of $a_k(m)$ gives

$$\pi_{W_m} \sum_{k=1}^{K} \big|p_k(\mathbf{A}) - p_{k|W_m}(\mathbf{A})\big| = \max_{\beta \in \mathcal{B}_\infty} \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \beta_k \mathbb{I}(A_i = k) \right) (c_{im} - \bar{c}_m).$$

Taking the maximum over $m \in \mathcal{M}_+$ yields $\Delta(\mathbf{A}; \mathcal{W}) = \mathrm{CR}(\mathbf{A}; \mathcal{W})$.

We next show that $\mathrm{CR}(\mathbf{A}; \mathcal{W})$ is also equal to $\overline{\mathrm{CR}}(\mathbf{A}; \mathcal{W})$. For fixed $\beta \in \mathcal{B}_\infty$, define

$$g_m(\beta) := \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \beta_k \mathbb{I}(A_i = k) \right) (c_{im} - \bar{c}_m), \qquad m \in \mathcal{M}_+,$$

and let $g(\beta) := (g_m(\beta))_{m \in \mathcal{M}_+}$. Then $\mathrm{CR}(\mathbf{A}; \mathcal{W}) = \max_{m \in \mathcal{M}_+} \max_{\beta \in \mathcal{B}_\infty} g_m(\beta)$.

For any fixed β, the function $\mathbf{v} \mapsto \mathbf{v}^\top g(\beta)$ is linear on the probability simplex $\mathbb{S}^{M_+}$, and hence its maximum is attained at a vertex. Thus,

$$\max_{\mathbf{v} \in \mathbb{S}^{M_+}} \mathbf{v}^\top g(\beta) = \max_{m \in \mathcal{M}_+} g_m(\beta).$$

Since the maximizations are over a product set,

$$\begin{aligned} \overline{\mathrm{CR}}(\mathbf{A}; \mathcal{W}) &= \max_{\beta \in \mathcal{B}_\infty} \max_{\mathbf{v} \in \mathbb{S}^{M_+}} \mathbf{v}^\top g(\beta) \\ &= \max_{\beta \in \mathcal{B}_\infty} \max_{m \in \mathcal{M}_+} g_m(\beta) \\ &= \max_{m \in \mathcal{M}_+} \max_{\beta \in \mathcal{B}_\infty} g_m(\beta) \\ &= \mathrm{CR}(\mathbf{A}; \mathcal{W}). \end{aligned}$$

Combining the two identities gives $\Delta(\mathbf{A}; \mathcal{W}) = \mathrm{CR}(\mathbf{A}; \mathcal{W}) = \overline{\mathrm{CR}}(\mathbf{A}; \mathcal{W})$. □

Soft-assignment extension. The equivalence arguments for hard assignment extend to soft assignment by replacing $\mathbb{I}(A_i = k)$ with A_{ik}. We first establish the extension for the subgroup-fairness gap and then extend it to a general collection $\mathcal{W}$.

Let $\mathbf{A} \in (\mathbb{S}^K)^n$ be a soft assignment, with $p_k(\mathbf{A}) := \frac{1}{n} \sum_{i=1}^{n} A_{ik}$ and $p_{k|\boldsymbol{s}}(\mathbf{A}) := \frac{1}{n_{\boldsymbol{s}}} \sum_{i:\boldsymbol{s}_i = \boldsymbol{s}} A_{ik}$. The subgroup-fairness gap is defined by $\Delta(\mathbf{A}) := \max_{\boldsymbol{s} \in \mathcal{S}_+} \pi_{\boldsymbol{s}} \sum_{k=1}^{K} |p_k(\mathbf{A}) - p_{k|\boldsymbol{s}}(\mathbf{A})|$.

For each $\boldsymbol{s} \in \mathcal{S}_+$ and $k \in [K]$, let $a_k(\boldsymbol{s}) := \frac{1}{2n} \sum_{i=1}^{n} A_{ik}(c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}})$. Since $\mathbb{I}(\boldsymbol{s}_i = \boldsymbol{s}) - \pi_{\boldsymbol{s}} = \frac{1}{2}(c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}})$,

$$a_k(\boldsymbol{s}) = \frac{1}{n} \sum_{i=1}^{n} A_{ik} \big(\mathbb{I}(\boldsymbol{s}_i = \boldsymbol{s}) - \pi_{\boldsymbol{s}}\big) = \pi_{\boldsymbol{s}} \big(p_{k|\boldsymbol{s}}(\mathbf{A}) - p_k(\mathbf{A})\big).$$

Hence,

$$\pi_{\boldsymbol{s}} \sum_{k=1}^{K} \big|p_k(\mathbf{A}) - p_{k|\boldsymbol{s}}(\mathbf{A})\big| = \max_{\beta \in \mathcal{B}_\infty} \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \beta_k A_{ik} \right) (c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}}).$$

Maximizing over $\boldsymbol{s} \in \mathcal{S}_+$ gives the subgroup-fairness gap above. Since the objective is linear in $\mathbf{v}$, maximizing over $\mathbf{v} \in \mathbb{S}^{M_+}$ is attained at a vertex and gives the same worst-subgroup maximization. Therefore, $\Delta(\mathbf{A}) = \overline{\mathrm{CR}}(\mathbf{A})$.

For a general collection $\mathcal{W} = \{W_1, \ldots, W_M\}$, let $p_{k|W_m}(\mathbf{A}) := \frac{1}{n_{W_m}} \sum_{i:\boldsymbol{s}_i \in W_m} A_{ik}$. Then $\Delta(\mathbf{A}; \mathcal{W}) := \max_{m \in \mathcal{M}_+} \pi_{W_m} \sum_{k=1}^{K} |p_k(\mathbf{A}) - p_{k|W_m}(\mathbf{A})|$. For each $m \in \mathcal{M}_+$ and $k \in [K]$, let $a_k(m) := \frac{1}{2n} \sum_{i=1}^{n} A_{ik}(c_{im} - \bar{c}_m)$. Using $r_m(i) - \pi_{W_m} = \frac{1}{2}(c_{im} - \bar{c}_m)$,

$$a_k(m) = \frac{1}{n} \sum_{i=1}^{n} A_{ik} \big(r_m(i) - \pi_{W_m}\big) = \pi_{W_m} \big(p_{k|W_m}(\mathbf{A}) - p_k(\mathbf{A})\big).$$

Thus,

$$\pi_{W_m} \sum_{k=1}^{K} \big|p_k(\mathbf{A}) - p_{k|W_m}(\mathbf{A})\big| = \max_{\beta \in \mathcal{B}_\infty} \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \beta_k A_{ik} \right) (c_{im} - \bar{c}_m).$$

Maximizing over $m \in \mathcal{M}_+$ and, equivalently, over $\mathbf{v} \in \mathbb{S}^{M_+}$ gives $\Delta(\mathbf{A}; \mathcal{W}) = \overline{\mathrm{CR}}(\mathbf{A}; \mathcal{W})$.

B.2 Fairness guarantees under rounding

The difference between the fairness gap after argmax rounding and the soft-assignment surrogate is bounded by the change from the soft assignment to the rounded hard assignment. We first establish the bound for subgroup fairness and then extend it to a general collection of subgroups.

For an assignment matrix $\mathbf{A}$, write $A_i := (A_{i1}, \dots, A_{iK})^\top$ for the assignment vector of instance i. Let $\widetilde{\mathbf{A}} \in (\mathbb{S}^K)^n$ be a soft assignment. For each $i \in [n]$, choose $k_i^\star \in \arg\max_{k\in[K]} \widetilde{A}_{ik}$ and define $\widehat{A}_{ik} := \mathbb{I}(k = k_i^\star)$. The resulting hard assignment matrix is $\widehat{\mathbf{A}} := (\widehat{A}_{ik})_{i\in[n],\, k\in[K]}$. We measure its deviation from the soft assignment by $\gamma(\widetilde{\mathbf{A}}) := \frac{1}{n}\sum_{i=1}^n \left(1 - \max_{k\in[K]} \widetilde{A}_{ik}\right)$.

Proposition B.1 (Rounding bound)**.** *For the argmax rounding $\widehat{\mathbf{A}}$ defined above,*

$$\left|\Delta(\widehat{\mathbf{A}}) - \overline{\mathrm{CR}}(\widetilde{\mathbf{A}})\right| \le 2\gamma(\widetilde{\mathbf{A}}). \tag{14}$$

Proof of Proposition B.1. By argmax rounding, $\|\widetilde{A}_i - \widehat{A}_i\|_1 = 2(1 - \max_{k\in[K]} \widetilde{A}_{ik})$ for each $i \in [n]$. Hence,

$$\frac{1}{n}\sum_{i=1}^n \|\widetilde{A}_i - \widehat{A}_i\|_1 = 2\gamma(\widetilde{\mathbf{A}}). \tag{15}$$

Recall that $\mathcal{S}_+$ is the set of observed subgroups, $c_{i\boldsymbol{s}} = 2\mathbb{I}(\boldsymbol{s}_i = \boldsymbol{s}) - 1$, and $\bar{c}_{\boldsymbol{s}} = 2\pi_{\boldsymbol{s}} - 1$. For each $\boldsymbol{s} \in \mathcal{S}_+$ and $\beta \in \mathcal{B}_\infty$, define

$$G_{\boldsymbol{s},\beta}(\mathbf{A}) := \frac{1}{2n}\sum_{i=1}^n \left(\sum_{k=1}^K \beta_k A_{ik}\right)(c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}}).$$

Since the objective is linear in the subgroup weights, maximization over their probability simplex is attained at a vertex. Thus, for every soft assignment $\mathbf{A}$,

$$\overline{\mathrm{CR}}(\mathbf{A}) = \max_{\boldsymbol{s}\in\mathcal{S}_+} \max_{\beta\in\mathcal{B}_\infty} G_{s,\beta}(\mathbf{A}).$$

Since $\|\beta\|_\infty \le 1$ and $|c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}}| \le 2$, we obtain

$$\begin{aligned}\left|G_{s,\beta}(\widetilde{\mathbf{A}}) - G_{s,\beta}(\widehat{\mathbf{A}})\right| &\le \frac{1}{n}\sum_{i=1}^n \|\widetilde{A}_i - \widehat{A}_i\|_1 \\ &= 2\gamma(\widetilde{\mathbf{A}}),\end{aligned}$$

where the equality follows from Eq. (15). This bound holds uniformly over (s, β), so

$$\left|\overline{\mathrm{CR}}(\widetilde{\mathbf{A}}) - \overline{\mathrm{CR}}(\widehat{\mathbf{A}})\right| \le 2\gamma(\widetilde{\mathbf{A}}).$$

Finally, since $\widehat{\mathbf{A}}$ is a hard assignment, Theorem 2.2 gives $\Delta(\widehat{\mathbf{A}}) = \overline{\mathrm{CR}}(\widehat{\mathbf{A}})$. Therefore, Eq. (14) follows. □

Proposition B.2 (Rounding bound over $\mathcal{W}$)**.** *For the argmax rounding $\widehat{\mathbf{A}}$ defined above,*

$$\left|\Delta(\widehat{\mathbf{A}}; \mathcal{W}) - \overline{\mathrm{CR}}(\widetilde{\mathbf{A}}; \mathcal{W})\right| \le 2\gamma(\widetilde{\mathbf{A}}). \tag{16}$$

Proof of Proposition B.2. By argmax rounding, $\|\widetilde{A}_i - \widehat{A}_i\|_1 = 2(1 - \max_{k\in[K]} \widetilde{A}_{ik})$ for each $i \in [n]$. Hence,

$$\frac{1}{n}\sum_{i=1}^n \|\widetilde{A}_i - \widehat{A}_i\|_1 = 2\gamma(\widetilde{\mathbf{A}}). \tag{17}$$

For any $\mathbf{v} \in \mathbb{S}^{M_+}$ and $\beta \in \mathcal{B}_\infty$, define

$$F_{\mathbf{v},\beta}(\mathbf{A}) := \frac{1}{2n}\sum_{i=1}^n \left(\sum_{k=1}^K \beta_k A_{ik}\right)\mathbf{v}^\top(c_i - \bar{c}).$$

Then

$$\overline{\mathrm{CR}}(\mathbf{A};\mathcal{W}) = \max_{\mathbf{v}\in\mathbb{S}^{M}_{+}} \max_{\beta\in\mathcal{B}_{\infty}} F_{\mathbf{v},\beta}(\mathbf{A}).$$

Since $\|\beta\|_\infty \le 1$ and $|\mathbf{v}^\top(c_i - \bar{c})| \le 2$, we obtain

$$\begin{aligned}\left|F_{\mathbf{v},\beta}(\widetilde{\mathbf{A}}) - F_{\mathbf{v},\beta}(\widehat{\mathbf{A}})\right| &\le \frac{1}{n}\sum_{i=1}^{n} \|\widetilde{A}_i - \widehat{A}_i\|_1 \\ &= 2\gamma(\widetilde{\mathbf{A}}),\end{aligned}$$

where the last equality follows from Eq. (17).

Since this bound holds uniformly over $(\mathbf{v}, \beta)$,

$$\left|\overline{\mathrm{CR}}(\widetilde{\mathbf{A}};\mathcal{W}) - \overline{\mathrm{CR}}(\widehat{\mathbf{A}};\mathcal{W})\right| \le 2\gamma(\widetilde{\mathbf{A}}).$$

Finally, since $\widehat{\mathbf{A}}$ is a hard assignment, Theorem 3.1 gives $\Delta(\widehat{\mathbf{A}};\mathcal{W}) = \overline{\mathrm{CR}}(\widehat{\mathbf{A}};\mathcal{W})$. Therefore, Eq. (16) follows. □

B.3 Convergence analysis

We establish a first-order convergence guarantee for the matrix formulation in Eq. (10). We first derive an explicit form of the adversarial feasible set and show that the matrix formulation is equivalent to the factorized formulation after maximization. Then we verify the regularity conditions required by the block alternating proximal gradient (BAPG) analysis of Xu et al. (2023) and apply their result to the updates defined below. Throughout this subsection, projections are Euclidean. For a matrix M, we write $\|M\|_F := (\sum_{i,j} M_{ij}^2)^{1/2}$ and $\|M\|_{\mathrm{op}} := \sup_{\|x\|_2=1} \|Mx\|_2$ for its Frobenius and operator norms, respectively.

Lemma B.3 (Exact adversarial reformulation). *Let* $\mathcal{U} := \mathrm{conv}\{\mathbf{v}\beta^\top : \mathbf{v} \in \mathbb{S}^{M_+},\ \beta \in \mathcal{B}_\infty\}$. *Then*

$$\mathcal{U} = \left\{ U \in \mathbb{R}^{M_+ \times K} : \sum_{m=1}^{M_+} \|U_{m,:}\|_\infty \le 1 \right\}. \tag{18}$$

In particular, $\mathcal{U}$ *is nonempty, compact, and convex, and* $\|U\|_F \le \sqrt{K}$ *for every* $U \in \mathcal{U}$.

Moreover, for every soft assignment $\mathbf{A} \in (\mathbb{S}^K)^n$,

$$\overline{\mathrm{CR}}(\mathbf{A}; \mathcal{W}) = \max_{U \in \mathcal{U}} \frac{1}{2n} \left\langle \mathbf{C}^\top \mathbf{A}, U \right\rangle_F. \tag{19}$$

Hence, maximizing $\widetilde{F}(\mathbf{A}, \mu, U)$ *over* $U \in \mathcal{U}$ *gives exactly the objective in Eq. (9).*

Proof. Let $\mathcal{S} := \{\mathbf{v}\beta^\top : \mathbf{v} \in \mathbb{S}^{M_+},\ \beta \in \mathcal{B}_\infty\}$, and denote the set on the right-hand side of Eq. (18) by $\mathcal{B}_U$.

If $V = \mathbf{v}\beta^\top \in \mathcal{S}$, then $V_{m,:} = v_m \beta^\top$ and $\sum_m \|V_{m,:}\|_\infty = \|\beta\|_\infty \sum_m v_m \le 1$. Hence $\mathcal{S} \subseteq \mathcal{B}_U$, and convexity of $\mathcal{B}_U$ gives $\mathcal{U} = \mathrm{conv}(\mathcal{S}) \subseteq \mathcal{B}_U$.

For the reverse inclusion, let $U \in \mathcal{B}_U$ and set $r_m := \|U_{m,:}\|_\infty$. For $r_m > 0$, define $\beta^{(m)} := U_{m,:}^\top / r_m$. Then $\beta^{(m)} \in \mathcal{B}_\infty$ and $e_m(\beta^{(m)})^\top \in \mathcal{S}$, where e_m is the mth standard basis vector. Since $\sum_m r_m \le 1$,

$$U = \sum_{m: r_m > 0} r_m e_m (\beta^{(m)})^\top + \left(1 - \sum_{m=1}^{M_+} r_m\right) 0.$$

Because $0 \in \mathcal{S}$, this is a convex combination of elements of $\mathcal{S}$. Thus $U \in \mathcal{U}$ and Eq. (18) follows. The representation in Eq. (18) also shows that $\mathcal{U}$ is nonempty, compact, and convex.

For any $U \in \mathcal{U}$,

$$\|U\|_F^2 \le K \sum_{m=1}^{M_+} \|U_{m,:}\|_\infty^2 \le K \left(\sum_{m=1}^{M_+} \|U_{m,:}\|_\infty \right)^2 \le K,$$

which gives $\|U\|_F \le \sqrt{K}$.

For $V = \mathbf{v}\beta^\top \in \mathcal{S}$, $\langle \mathbf{C}^\top \mathbf{A}, V \rangle_F = \beta^\top \mathbf{A}^\top \mathbf{C} \mathbf{v}$. Using Eq. (7) and the linearity of the objective,

$$\overline{\mathrm{CR}}(\mathbf{A}; \mathcal{W}) = \frac{1}{2n} \max_{V \in \mathcal{S}} \langle \mathbf{C}^\top \mathbf{A}, V \rangle_F = \frac{1}{2n} \max_{U \in \mathrm{conv}(\mathcal{S})} \langle \mathbf{C}^\top \mathbf{A}, U \rangle_F,$$

which proves Eq. (19). □

Feasible sets and regularity. Let $R_x := \max_{i \in [n]} \|x_i\|_2$ and fix any $R > R_x$. We restrict the centroids to $\mathcal{C}_\mu := \{\mu = (\mu_1, \ldots, \mu_K) : \|\mu_k\|_2 \le R,\ k \in [K]\}$, and define

$$\mathcal{A} := (\mathbb{S}^K)^n, \quad \mathcal{Z} := \mathcal{A} \times \mathcal{C}_\mu. \tag{20}$$

The sets $\mathcal{A}$, $\mathcal{C}_\mu$, and $\mathcal{U}$ are nonempty, compact, and convex.

Restricting the centroids to $\mathcal{C}_\mu$ does not change the minimum of Eq. (9). Indeed, since every x_i lies in the radius-R ball, projecting a centroid onto this ball cannot increase its distance to any observation, while the fairness penalty is independent of μ.

For $z = (\mathbf{A}, \mu) \in \mathcal{Z}$, we use the product norm $\|z\|^2 = \|\mathbf{A}\|_F^2 + \|\mu\|_F^2$.

Lemma B.4 (Regularity of the matrix formulation). *The objective $\widetilde{F}$ in Eq. (10) is continuously differentiable on $\mathcal{Z} \times \mathcal{U}$ and affine, hence concave, in U. Its partial gradients are*

$$\nabla_{\mathbf{A}} \widetilde{F} = \frac{1}{n} D(\mu) + \frac{\lambda}{2n} \mathbf{C} U, \tag{21}$$

$$\nabla_{\mu_k} \widetilde{F} = \frac{2}{n} \sum_{i=1}^{n} A_{ik}(\mu_k - x_i), \quad k \in [K], \tag{22}$$

$$\nabla_U \widetilde{F} = \frac{\lambda}{2n} \mathbf{C}^\top \mathbf{A}. \tag{23}$$

Here, $D(\mu)_{ik} := \|x_i - \mu_k\|_2^2$.

Moreover, the blockwise Lipschitz conditions in Assumption 2.1 of Xu et al. (2023) hold on $\mathcal{Z} \times \mathcal{U}$. In particular, all four Lipschitz constants in that assumption may be chosen as the common positive upper bound

$$L_{\mathrm{B}} := \max\left\{1,\ 2 + \frac{4R}{\sqrt{n}},\ \frac{\lambda}{2n}\|\mathbf{C}\|_{\mathrm{op}}\right\}. \tag{24}$$

Proof. The gradient formulas follow directly from Eq. (10). For fixed U, the $\mathbf{A}$–$\mathbf{A}$ Hessian is zero. For a perturbation $h = (h_1, \ldots, h_K)$ in μ,

$$\left\| D_\mu(\nabla_{\mathbf{A}} \widetilde{F})[h] \right\|_F^2 = \frac{4}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{K} \langle \mu_k - x_i, h_k \rangle^2 \le \frac{16R^2}{n} \|h\|_F^2,$$

where $\|x_i - \mu_k\|_2 \le 2R$ on $\mathcal{C}_\mu$. Thus, the mixed derivative has operator norm at most $4R/\sqrt{n}$. Also, since $\sum_i A_{ik} \le n$, the μ_k–μ_k Hessian $\frac{2}{n}(\sum_i A_{ik}) I_d$ has operator norm at most 2. Hence, the gradient with respect to $z = (\mathbf{A}, \mu)$ is Lipschitz in z with constant at most $2 + 4R/\sqrt{n}$.

The dependence between z and U comes only from the bilinear term $\frac{\lambda}{2n}\langle \mathbf{C}^\top \mathbf{A}, U \rangle_F$, so both cross-block Lipschitz constants are bounded by $\frac{\lambda}{2n}\|\mathbf{C}\|_{\mathrm{op}}$. Since $\nabla_U \widetilde{F}$ does not depend on U, its Lipschitz constant in U is zero. Therefore, all four constants in Assumption 2.1 of Xu et al. (2023) are bounded by L_{B} in Eq. (24).

□

BAPG parameters. For $t \ge 1$, let

$$\xi = \frac{1}{10 L_{\mathrm{B}}}, \quad c_t = \frac{19}{20 \xi t^{1/4}}, \quad \eta_t = \left(\xi L_{\mathrm{B}}^2 + \frac{64 L_{\mathrm{B}}^2}{\xi c_t^2} - L_{\mathrm{B}} \right)^{-1}. \tag{25}$$

The two minimization blocks, $\mathbf{A}$ and μ, use the step size η_t, while the maximization block U uses the fixed step size ξ. In particular,

$$\eta_t = \Theta(t^{-1/2}), \quad c_t = \Theta(t^{-1/4}).$$

Thus, the quadratic term in Eq. (11) is used during the updates but vanishes asymptotically.

BAPG updates. Using the step sizes above, we sequentially update the three blocks as

$$\mathbf{A}^{t+1} = \Pi_{\mathcal{A}}\left[\mathbf{A}^t - \eta_t \nabla_{\mathbf{A}} \widetilde{F}(\mathbf{A}^t, \mu^t, U^t) \right], \tag{26}$$

$$\mu^{t+1} = \Pi_{\mathcal{C}_\mu}\left[\mu^t - \eta_t \nabla_{\mu} \widetilde{F}(\mathbf{A}^{t+1}, \mu^t, U^t) \right], \tag{27}$$

$$U^{t+1} = \Pi_{\mathcal{U}}\left[U^t + \xi \left\{ \nabla_U \widetilde{F}(\mathbf{A}^{t+1}, \mu^{t+1}, U^t) - c_t U^t \right\} \right]. \tag{28}$$

The corresponding explicit update formulas are provided in Appendix C.1.

Stationarity measure. Let $w^t := (\mathbf{A}^t, \mu^t, U^t)$. We define the projected gradient mapping of the unregularized objective $\widetilde{F}$ at w^t by

$$\begin{aligned} G_{A,t} &:= \eta_t^{-1}\left[\mathbf{A}^t - \Pi_{\mathcal{A}}\left\{\mathbf{A}^t - \eta_t \nabla_{\mathbf{A}}\widetilde{F}(w^t)\right\}\right], \\ G_{\mu,t} &:= \eta_t^{-1}\left[\mu^t - \Pi_{\mathcal{C}_\mu}\left\{\mu^t - \eta_t \nabla_{\mu}\widetilde{F}(w^t)\right\}\right], \\ G_{U,t} &:= \xi^{-1}\left[U^t - \Pi_{\mathcal{U}}\left\{U^t + \xi \nabla_{U}\widetilde{F}(w^t)\right\}\right], \\ \|\mathcal{G}_t\|^2 &:= \|G_{A,t}\|_F^2 + \|G_{\mu,t}\|_F^2 + \|G_{U,t}\|_F^2. \end{aligned} \tag{29}$$

We call w^t ε-stationary if $\|\mathcal{G}_t\| \le \varepsilon$.

Theorem B.5 (First-order stationarity)**.** *Suppose $M_+ \ge 1$. Let the iterates be generated by Eqs. (26) to (28) from any feasible initialization, using exact projections and the schedule in Eq. (25).*

For $T(\varepsilon) := \min\{t \ge 1 : \|\mathcal{G}_t\| \le \varepsilon\}$, we have

$$T(\varepsilon) = \mathcal{O}(\varepsilon^{-4}), \quad \varepsilon \in (0, 1). \tag{30}$$

Thus, the BAPG updates obtain a first-order ε-stationary iterate of the original matrix objective $\widetilde{F}$ within $\mathcal{O}(\varepsilon^{-4})$ iterations.

Proof. We apply the nonconvex-concave BAPG result of Xu et al. (2023, Section 5.1.2, Theorem 5.2) with two minimization blocks $x^{(1)} = \mathbf{A}$ and $x^{(2)} = \mu$ and one maximization block $y^{(1)} = U$. The updates in Eqs. (26) to (28) are the corresponding BAPG updates.

The feasible sets $\mathcal{A}$, $\mathcal{C}_\mu$, and $\mathcal{U}$ are compact and convex by definition and Lemma B.3. By Lemma B.4, $\widetilde{F}$ is smooth, possibly nonconvex in $(\mathbf{A}, \mu)$, concave in U, and satisfies the required blockwise Lipschitz conditions with constant L_{B}.

It remains to verify the parameter conditions in Xu et al. (2023, Theorem 5.2). In their notation, $\widehat{b}_t$ and $\widehat{c}_t$ are the quadratic regularization parameters for the minimization and maximization blocks, respectively. Since our minimization blocks are unregularized, we set $\widehat{b}_t = 0$, while the regularization of the U-block gives $\widehat{c}_t = c_t$. For the remaining parameters, we set

$$\iota = L_{\mathrm{B}}, \quad \tau = 4, \quad \widehat{b}_t = 0, \quad \widehat{c}_t = c_t, \quad \vartheta_t = \xi L_{\mathrm{B}}^2 + \frac{64 L_{\mathrm{B}}^2}{\xi c_t^2} - 2L_{\mathrm{B}}.$$

Then $\eta_t = (\iota + \vartheta_t)^{-1}$, which agrees with the step size in Eq. (25).

Since c_t decreases with t, ϑ_t increases. Moreover, $\vartheta_1 = (2560/361 - 19/10)L_{\mathrm{B}} > L_{\mathrm{B}}$, so $\vartheta_t > L_{\mathrm{B}}$ for all $t \ge 1$. The choice $\xi = 1/(10L_{\mathrm{B}})$ also satisfies the dual step-size condition of Theorem 5.2.

With $c_1 = 19/(20\xi) = 19L_{\mathrm{B}}/2$, the remaining threshold condition is

$$\frac{\xi c_1^2 \left(L_{\mathrm{B}} + 2L_{\mathrm{B}} - \xi L_{\mathrm{B}}^2\right)}{16 L_{\mathrm{B}}^2} = \frac{10469}{6400} < 2 < \tau.$$

Thus, all parameter conditions of Theorem 5.2 are satisfied.

With these block definitions and parameter settings, the stationarity measure in Definition 5.1 of Xu et al. (2023) coincides with the projected gradient mapping $\mathcal{G}_t$ in Eq. (29). Theorem 5.2 therefore yields $T(\varepsilon) = \mathcal{O}(\varepsilon^{-4})$ for first-order stationarity of the original objective $\widetilde{F}$.

The U-update uses the regularized objective $\widetilde{F}_t = \widetilde{F} - \frac{c_t}{2}\|U\|_F^2$. Let $\overline{\mathcal{G}}_t$ be the corresponding projected gradient mapping. Its $\mathbf{A}$- and μ-components agree with those of $\mathcal{G}_t$, while the U-gradients differ by $c_t U^t$. By nonexpansiveness of the projection,

$$\|\mathcal{G}_t - \overline{\mathcal{G}}_t\| \le c_t \|U^t\|_F \le c_t\sqrt{K} \longrightarrow 0,$$

where the second inequality follows from Lemma B.3. Thus, the vanishing regularization does not affect the limiting first-order stationarity condition for $\widetilde{F}$. □

C Algorithmic Details

C.1 Block Updates and Pseudocode of Cova-FC

We optimize the objective in Eq. (10) using sequential block updates in the order $\mathbf{A} \to \mu \to U$, where U replaces the product $\mathbf{v}\beta^\top$ in the maximization. The step sizes η_t and ξ and the regularization parameter c_t are defined in Eq. (25).

Update of A. With μ^t and U^t fixed, the soft assignment matrix $\mathbf{A}$ is updated by projected gradient descent:

$$\mathbf{A}^{t+1} = \Pi_{\mathcal{A}}\left[\mathbf{A}^t - \eta_t\left\{\frac{1}{n}D(\mu^t) + \frac{\lambda}{2n}\mathbf{C}U^t\right\}\right], \tag{31}$$

where $D(\mu^t)_{ik} := \|x_i - \mu_k^t\|_2^2$. The projection onto $\mathcal{A}$ is performed row-wise onto $\mathbb{S}^K$ as described in Appendix C.2.

Update of μ. Given $\mathbf{A}^{t+1}$, each centroid is updated as

$$\mu_k^{t+1} = \Pi_{\{u:\|u\|_2\le R\}}\left[\mu_k^t - \frac{2\eta_t}{n}\sum_{i=1}^n A_{ik}^{t+1}(\mu_k^t - x_i)\right], \quad k \in [K]. \tag{32}$$

The fixed radius $R > \max_{i\in[n]}\|x_i\|_2$ is the centroid norm bound used in Appendix B.3.

Update of U. Given $\mathbf{A}^{t+1}$, the U-block is updated by projected gradient ascent on the regularized objective in Eq. (11):

$$U^{t+1} = \Pi_{\mathcal{U}}\left[U^t + \xi\left\{\frac{\lambda}{2n}\mathbf{C}^\top\mathbf{A}^{t+1} - c_t U^t\right\}\right], \tag{33}$$

where $\Pi_{\mathcal{U}}$ denotes the Euclidean projection onto $\mathcal{U}$ in Eq. (18).

Each iteration of Cova-FC applies these three updates in sequence, as summarized in Algorithm 1.

C.2 Simplex projection of the assignment vectors

For each $i \in [n]$, let $\mathbf{A}_i^{\boldsymbol{s}} \in \mathbb{R}^K$ denote the assignment vector obtained after a gradient step. This assignment vector may have negative entries or entries that do not sum to one. We therefore project it onto $\mathbb{S}^K$ by solving the strictly convex problem

$$\Pi_{\mathbb{S}^K}(\mathbf{A}_i^{\boldsymbol{s}}) = \arg\min_{v\in\mathbb{S}^K}\frac{1}{2}\|v - \mathbf{A}_i^{\boldsymbol{s}}\|^2. \tag{34}$$

The projected vector $v_i^\star$ is obtained by shifting all entries by the same threshold τ_i and clipping the result at zero. This follows from the KKT conditions (Wang & Carreira-Perpiñán, 2013):

$$v_{ik}^\star = \max\{A_{ik}^{\boldsymbol{s}} - \tau_i,\, 0\}, \quad k \in [K], \tag{35}$$

where τ_i is chosen uniquely so that $\sum_{k=1}^K v_{ik}^\star = 1$.

To compute τ_i efficiently, let $u_1 \ge u_2 \ge \cdots \ge u_K$ be the sorted entries of $\mathbf{A}_i^{\boldsymbol{s}}$, and define the cumulative sums $\sigma_j := \sum_{\ell=1}^j u_\ell$ for $j \in [K]$. The number ρ of positive entries after projection and the corresponding threshold are given by

$$\rho := \max\left\{j \in [K] : u_j - \frac{1}{j}(\sigma_j - 1) > 0\right\}, \quad \tau_i := \frac{1}{\rho}(\sigma_\rho - 1). \tag{36}$$

Substituting this threshold into Eq. (35) gives the projected vector $v_i^\star = \Pi_{\mathbb{S}^K}(\mathbf{A}_i^s)$. The computational complexity of this projection is $O(K\log K)$ for each $i \in [n]$ due to sorting.

Algorithm 1 COVA-FC

Require: $\{(x_i, \boldsymbol{s}_i)\}_{i=1}^n$, $\mathcal{W} = \{W_m\}_{m=1}^M$, K, λ

Initialize $\mathbf{A}$ and μ using standard K-means; set $U = 0$

Compute $C_{im} = 2\big(\mathbb{I}(\boldsymbol{s}_i \in W_m) - \pi_{W_m}\big)$ for $i \in [n]$, $m \in \mathcal{M}_+$

Set ξ and choose $R > \max_{i\in[n]} \|x_i\|_2$

for $t = 1, 2, \ldots$ **do**

  Compute c_t and η_t

  Update the soft assignment:

$$\mathbf{A} \leftarrow \Pi_{\mathcal{A}}\left[\mathbf{A} - \eta_t\left\{\frac{1}{n}D(\mu) + \frac{\lambda}{2n}\mathbf{C}U\right\}\right]$$

  Update the centroids for each $k \in [K]$:

$$\mu_k \leftarrow \Pi_{\{u:\|u\|_2\le R\}}\left[\mu_k - \frac{2\eta_t}{n}\sum_{i=1}^n A_{ik}(\mu_k - x_i)\right]$$

  Update U by regularized projected gradient ascent:

$$U \leftarrow \Pi_{\mathcal{U}}\left[U + \xi\left\{\frac{\lambda}{2n}\mathbf{C}^\top\mathbf{A} - c_t U\right\}\right]$$

  if the stopping criterion is satisfied **then**

    break

  end if

end for

Return $(\mathbf{A}, \mu)$

C.3 COMPUTATIONAL COMPLEXITY

The fairness penalty in our objective can be written as

$$\overline{\mathrm{CR}}(\mathbf{A}; \mathcal{W}) = \max_{U\in\mathcal{U}} \frac{1}{2n}\langle \mathbf{C}^\top\mathbf{A}, U\rangle_F,$$

where $\mathcal{U} := \mathrm{conv}\{\mathbf{v}\beta^\top : \mathbf{v} \in \mathbb{S}^{M_+}, \beta \in \mathcal{B}_\infty\}$. The fairness terms in the $\mathbf{A}$- and U-gradients of the matrix objective involve $\mathbf{C}U$ and $\mathbf{C}^\top\mathbf{A}$, respectively. We compute these products using group-wise sums.

Let $H \in \{0,1\}^{n\times M_+}$ be the membership matrix with entries $H_{im} := \mathbb{I}(\boldsymbol{s}_i \in W_m)$, and let $\pi := (\pi_{W_m})_{m\in\mathcal{M}_+}$. The centered membership matrix satisfies $\mathbf{C} = 2(H - \mathbf{1}\pi^\top)$. Therefore,

$$\mathbf{C}U = 2\big[HU - \mathbf{1}(\pi^\top U)\big], \quad \mathbf{C}^\top\mathbf{A} = 2\big[H^\top\mathbf{A} - \pi(\mathbf{1}^\top\mathbf{A})\big].$$

where $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector.

Thus, exploiting the sparsity of H, the products $\mathbf{C}U$ and $\mathbf{C}^\top\mathbf{A}$ can be computed efficiently. Let Z denote the number of nonzero entries in H. Computing HU and $H^\top\mathbf{A}$ costs $\mathcal{O}(ZK)$, and the terms $\mathbf{1}(\pi^\top U)$ and $\pi(\mathbf{1}^\top\mathbf{A})$ cost $\mathcal{O}(nK + KM_+)$. Including the $\mathcal{O}(nKd)$ cost of the clustering gradients, the gradient computation cost per iteration is $\mathcal{O}(nKd + ZK + KM_+)$. In addition, projecting the n assignment rows onto $\mathbb{S}^K$ costs $\mathcal{O}(nK\log K)$ per iteration, as detailed in Appendix C.2. Projection onto the centroid balls requires only norm computations and rescaling, costing $\mathcal{O}(Kd)$ in total, which is dominated by the clustering-gradient cost.

By Lemma B.3, $\mathcal{U}$ is the unit $\ell_{1,\infty}$-norm ball, $\{U : \sum_{m=1}^{M_+}\|U_{m,:}\|_\infty \le 1\}$. The algorithm of Quattoni et al. (2009) computes Euclidean projection onto this set in $\mathcal{O}(KM_+\log(KM_+))$ time. With this projection routine, a full iteration can be implemented in $\mathcal{O}(nKd + ZK + nK\log K + KM_+\log(KM_+))$ time.

In the subgroup-only setting, each instance belongs to exactly one observed subgroup, so $Z = n$. The overall per-iteration complexity is therefore $\mathcal{O}(nKd + nK \log K + KM_+ \log(KM_+))$. Using $M_+ \leq 2^q$, this is upper bounded by $\mathcal{O}(nKd + nK \log K + K2^q \log(K2^q))$.

Table 4: Comparison of fair clustering methods in terms of computational complexity.

Method	Complexity
COVA-FC ✓	$nKd + nK \log K + K2^q \log(K2^q)$
VFC (Ziko et al., 2021)	$nK(d + 2^q)$
FairKM (Abraham et al., 2020)	$nK(nd + 2^q)$
FCA (Kim et al., 2025b)	$(2^q - 1)n^2(n + Kd)$
FCBC (Esmaeili et al., 2021)	$(nK + 2^q K)^3$
FRA (Bera et al., 2019)	$(nK + 2^q)^3$

In contrast, most of the fair clustering algorithms have higher computational complexity:

(i) VFC (Ziko et al., 2021) has complexity linear in n, similarly to COVA-FC, but its subgroup-dependent term is $nK2^q$, whereas the corresponding term in COVA-FC is $K2^q \log(K2^q)$.

(ii) FairKM (Abraham et al., 2020) includes a quadratic term $\mathcal{O}(n^2Kd)$, which is substantially larger than first-order updates for large-scale datasets.

(iii) FCA (Kim et al., 2025b), FCBC (Esmaeili et al., 2021), and FRA (Bera et al., 2019) are dominated by solving LPs, yielding cubic complexity with respect to n.

Overall, COVA-FC achieves a favorable computational complexity, making it more suitable for large datasets with multiple sensitive attributes. We also observe that COVA-FC runs faster than the baselines, whose results are provided in Table 3 of Section 5.4. Furthermore, COVA-FC runs $3.2\times$ faster than VFC on a large-scale dataset ($n \approx$ millions, see Appendix D.4 for details).

D Additional Experimental Results

D.1 Settings

Datasets. Table 5 summarizes the seven datasets. Feature selection, sensitive attributes, and preprocessing are described below. For tabular datasets, we use numerical (continuous) features only. We standardize the features of all the datasets to zero mean and unit variance.

- `Adult` (U.S. census; Becker & Kohavi, 1996): The features are `age`, `fnlwgt`, `education-num`, `capital-gain`, and `hours-per-week`. We use four binary sensitive attributes: gender (male/female), race (white/non-white), age (≥ 40 / < 40), and marital status (married/unmarried), so $q = 4$ and there are $2^q = 16$ subgroups. The number of instances is $n = 32{,}561$.
- `Dutch` (Dutch census; Le Quy et al. (2022)): We use the remaining census attributes as features. We use two binary sensitive attributes: gender (male/female) and age (above median / at or below median),
- `Bank` (bank marketing; Moro et al., 2014): The features are `age`, `call-duration`, `3-month-Euribor-rate`, `number-of-employees`, `consumer-price-index`, and `number-of-contacts`. We treat marital status as a sensitive attribute with three categories (single/married/divorced) and exclude `unknown` entries.
- `Civilcomments` (online comment toxicity; Borkan et al., 2019): We use the provided feature embeddings. We use three multi-valued sensitive attributes: gender (male/female/other), religion (christian/other), and race (black/white/asian/other).
- `Communities` (community demographics and crime; Redmond & Baveja, 2002): We remove entries with missing values. We use $q = 18$ binary sensitive attributes derived from race-related demographic ratios, each binarized at its median. The Cartesian product yields up to 2^{18} combinations, of which $1{,}180$ are non-empty in the data.
- `ACSIncome` (American Community Survey; Ding et al., 2021): We use three sensitive attributes: gender (male/female), marital status (married/unmarried), and race (white/black/asian/other).
- `CelebA` (face images; Liu et al., 2015): We extract image features from a pre-trained ResNet-18 (He et al., 2016) and apply CovA-FC on the fixed features without modification. We use four binary sensitive attributes (Male, Young, Smiling, Attractive).

Table 5: Summary statistics of datasets. n is the number of instances, d is the number of features, and q is the number of sensitive attributes. "# subgroups per attr" lists the number of categories for each sensitive attribute and "# Subgroups" is the number of subgroups formed by the Cartesian product of sensitive attributes. For `Communities`, the subgroup count includes only observed combinations.

Dataset	n	d	q	# subgroups per attr	# Subgroups
`Adult`	32,561	5	4	2/2/2/2	16
`Dutch`	60,420	9	2	2/2	4
`Bank`	41,108	6	1	3	3
`Civilcomments`	3,365	768	3	3/2/4	24
`Communities`	1,994	99	18	2 (each)	1,180
`ACSIncome`	1,664,500	7	3	2/2/4	16
`CelebA`	202,599	512	4	2/2/2/2	16

Baselines. FCA faces a bottleneck: its core step estimates joint distributions over subgroups by solving a linear program, which incurs high computational cost as the number of subgroups and instances increases. On `Adult` and `Civilcomments`, which have 16 and 24 subgroups, an experiment did not complete within 6 hours, so we omit FCA from these comparisons and report it only on `Dutch` and `Bank`.

Implementation details. To control the cost-fairness trade-off, we sweep λ on a grid: $\lambda \in [10^{-1}, 10^{4}]$ for `Adult`, `Dutch`, and `Bank`, and $\lambda \in [10^{-1}, 10^{6}]$ for `Civilcomments`. We initialize the centroids μ and one-hot assignment $\mathbf{A}$ using K-means (Lloyd, 1982), and set $U = 0$. Algorithm 1 uses the parameter schedule specified in Eq. (25). All experiments are conducted on a Linux server with four NVIDIA RTX 3090 GPUs (24 GB VRAM each) and a 48-core Intel Xeon Silver 4310 CPU.

D.2 Additional Cost–Fairness Comparisons

Variability across random seeds. For each method we select the point with the lowest mean SP in Figure 2 and report its mean $\pm$ standard deviation over five random seeds. Due to runtime constraints in our experimental environment, we evaluate FCA only on `Dutch` and `Bank`. Details are provided in Appendix D.1.

Table 6: Mean $\pm$ standard deviation over five random seeds at the lowest-SP point of each method.

Dataset	Method	Cost	SP
`Adult`	Unfair	1.61 ± 0.0000	0.1953 ± 0.0004
	COVA-FC	2.52 ± 0.0020	$\mathbf{0.0003 \pm 0.0000}$
	FCBC	2.67 ± 0.0010	0.0061 ± 0.0000
	FairKM	2.07 ± 0.0000	0.0205 ± 0.0000
	FRA	2.89 ± 0.1100	$\mathbf{0.0003 \pm 0.0000}$
	VFC	1.62 ± 0.0030	0.1867 ± 0.0185
`Bank`	Unfair	1.51 ± 0.0000	0.1119 ± 0.0001
	COVA-FC	1.83 ± 0.0000	$\mathbf{0.0001 \pm 0.0000}$
	FCBC	1.84 ± 0.0000	0.0002 ± 0.0000
	FairKM	1.85 ± 0.0020	0.0003 ± 0.0000
	FRA	1.88 ± 0.0230	0.0002 ± 0.0000
	VFC	1.57 ± 0.0000	0.0529 ± 0.0000
	FCA	2.76 ± 0.0027	0.0196 ± 0.0008
`Dutch`	Unfair	2.85 ± 0.0300	0.1074 ± 0.0089
	COVA-FC	3.95 ± 0.0200	$\mathbf{0.0001 \pm 0.0001}$
	FCBC	3.98 ± 0.0200	0.0003 ± 0.0001
	FairKM	3.87 ± 0.0200	0.0265 ± 0.0007
	FRA	4.19 ± 0.1200	0.0002 ± 0.0000
	VFC	3.05 ± 0.0000	0.0398 ± 0.0002
	FCA	6.25 ± 0.0020	0.0041 ± 0.0007
`Civilcomments`	Unfair	624.51 ± 0.1500	0.0752 ± 0.0011
	COVA-FC	661.79 ± 1.0900	$\mathbf{0.0011 \pm 0.0001}$
	FCBC	658.11 ± 0.9100	0.0028 ± 0.0003
	FairKM	654.32 ± 0.9500	0.0246 ± 0.0010
	FRA	664.92 ± 1.4400	0.0025 ± 0.0001
	VFC	624.71 ± 0.3200	0.0743 ± 0.0017

Most methods show little variation across random seeds. COVA-FC achieves the lowest mean SP in most cases, with SP standard deviations of at most 10^{-4} across all four datasets.

Cost-fairness trade-off: marginal-fairness gap. In the main experiments, the subgroup collection $\mathcal{W}$ contains only the full subgroups, so COVA-FC is optimized for subgroup fairness. To examine marginal fairness under this subgroup-only setting, we additionally report the trade-off between $\text{MP}^{(1)}$ and cost on the four benchmark datasets in Figure 4. Although marginal fairness is not separately targeted, COVA-FC attains lower $\text{MP}^{(1)}$ than the baselines in most cases. When subgroups contain few instances, however, subgroup fairness alone may be insufficient. Our marginal-fairness extension addresses this limitation, as evaluated on `Communities` in Section 5.3.

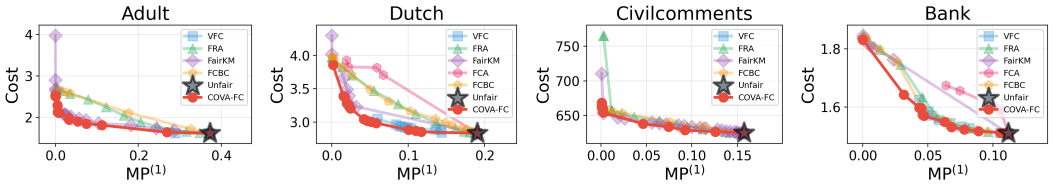


Figure 4: Trade-offs between $\text{MP}^{(1)}$ and cost on (left to right) `Adult`, `Dutch`, `Civilcomments`, and `Bank` datasets.

Cost-fairness trade-off: sum-aggregated gap. For a more comprehensive analysis, we further examine whether the same conclusions hold under an alternative aggregation rule. While the gap Δ in Eq. (5) takes a worst-case (maximum) over subgroups, we also report sum-aggregated variants SP_{sum} and $\text{MP}^{(r)}_{\text{sum}}$ defined in Table 7.

Table 7: Sum-aggregated variants of the gap. SP and MP are special cases of $\Delta(\mathbf{A}; \mathcal{W})$ in Eq. (5) with appropriate $\mathcal{W}$. We follow the notation of Sections 2 and 3: $\theta \subset [q]$ with $|\theta| = r$ indexes an r-th order marginal subgroup, and $W_{\theta,a} = \{\boldsymbol{s} \in \{0,1\}^q : \boldsymbol{s}_\theta = a\}$ for $a \in \{0,1\}^r$.

Metric	Base	Aggregation
SP_{sum}	$\delta_{k,\boldsymbol{s}} := \lvert p_k(\mathbf{A}) - p_{k\lvert\boldsymbol{s}}(\mathbf{A})\rvert$	$\sum_{\boldsymbol{s}\in\mathcal{S}_+} \pi_{\boldsymbol{s}} \sum_{k\in[K]} \delta_{k,\boldsymbol{s}}$
$\text{MP}^{(r)}_{\text{sum}}$	$\delta_{k,\theta,a} := \lvert p_k(\mathbf{A}) - p_{k\lvert W_{\theta,a}}(\mathbf{A})\rvert$	$\sum_{\substack{\theta\subseteq[q]\\\lvert\theta\rvert=r}} \sum_{\substack{a\in\{0,1\}^r\\ n_{W_{\theta,a}}>0}} \pi_{W_{\theta,a}} \sum_{k\in[K]} \delta_{k,\theta,a}$

As shown in Figure 5, evaluating fairness using sum aggregation does not change the main messages of our experiments. That is, COVA-FC still attains competitive cost-fairness trade-offs under the sum-aggregated fairness metrics.

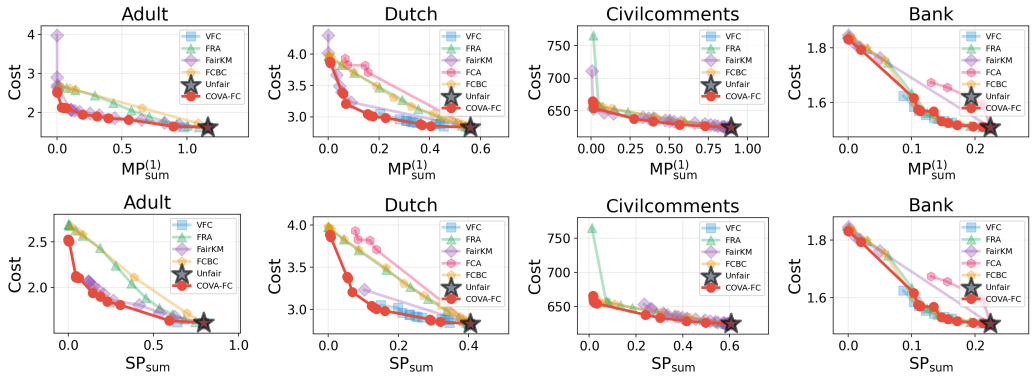


Figure 5: Comparison of trade-offs between 'sum-aggregated' gap (top: $\text{MP}^{(1)}_{\text{sum}}$, bottom: SP_{sum}) and cost on (left to right) `Adult`, `Dutch`, `Civilcomments`, and `Bank` datasets. Note that SP_{sum} = $\text{MP}^{(1)}_{\text{sum}}$ for `Bank` dataset since $q = 1$.

Cost-fairness trade-off: subgroup Balance.

Remark D.1 (Why we use the fairness gap rather than Balance). Balance (Chierichetti et al., 2017) is a commonly used ratio-based fairness measure, but we do not use it as a primary metric for two reasons: *(i)* it can collapse to zero when one subgroup is absent from a cluster while another compared subgroup is present; a subgroup with fewer than K instances is necessarily absent from some clusters (e.g., on `Civilcomments`); and *(ii)* it can be unstable when subgroup sizes are highly imbalanced. In contrast, the gap remains well-defined under zero cluster counts and reflects subgroup size through π_W rather than raw count ratios.

We use a normalized Balance that compares conditional cluster proportions rather than raw subgroup counts. Using the notation of Sections 2 and 3, for subgroup-subsets W, W' with $n_W, n_{W'} > 0$, define

$$B_{k,W,W'} := \frac{\min\{p_{k|W}(\mathbf{A}), p_{k|W'}(\mathbf{A})\}}{\max\{p_{k|W}(\mathbf{A}), p_{k|W'}(\mathbf{A})\}}, \quad \text{if } p_{k|W}(\mathbf{A}) + p_{k|W'}(\mathbf{A}) > 0,$$

$$\text{SB} := \min_{\boldsymbol{s}\neq\boldsymbol{s}'} \min_{k\in[K]} B_{k,\boldsymbol{s},\boldsymbol{s}'}, \qquad \text{MB}^{(r)} := \min_{\substack{\theta\subseteq[q]\\|\theta|=r}} \min_{a\neq a'} \min_{k\in[K]} B_{k,W_{\theta,a},W_{\theta,a'}}.$$

Here, $B_{k,\boldsymbol{s},\boldsymbol{s}'} := B_{k,\{\boldsymbol{s}\},\{\boldsymbol{s}'\}}$, and $\text{MB}^{(r)}$ minimizes over pairs of r-th order marginal subgroups. We consider only subgroup pairs with positive sample counts. If both conditional cluster proportions are zero, the comparison is treated as N/A and excluded from the minimum.

On `Adult`, `Dutch`, and `Bank`, Figure 6 shows that COVA-FC attains higher SB and $\text{MB}^{(1)}$ at comparable cost than the baselines in most cases, consistent with the gap-based results.

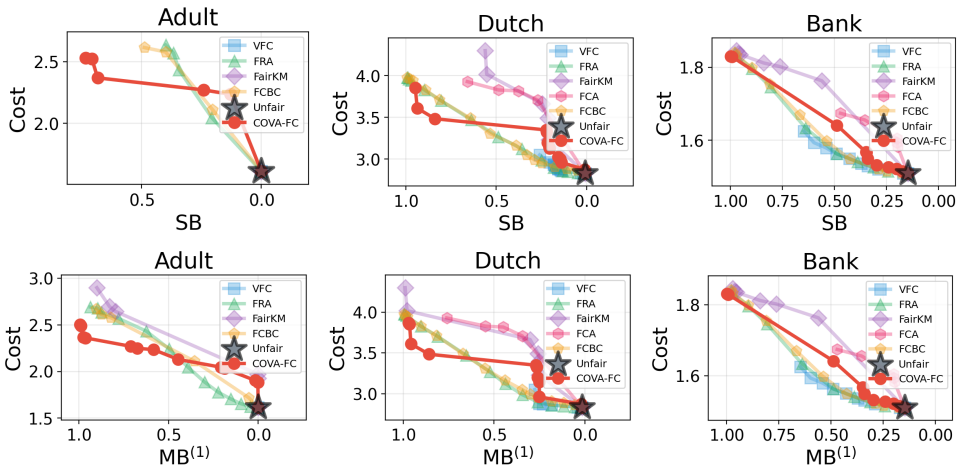


Figure 6: Comparison of trade-offs between Balance (top: SB , bottom: $\text{MB}^{(1)}$) and cost on (left to right) `Adult`, `Dutch`, and `Bank` datasets. `Civilcomments` is excluded because some subgroups contain fewer than K instances, so they are absent from some clusters, making the corresponding Balance comparisons zero or N/A.

On `Adult`, both VFC and FairKM stay at SB $= 0$, while COVA-FC reaches strictly positive SB. Empirically, we observe that FairKM's hard assignment often leave at least one small subgroup empty in some cluster ($p_{k|\boldsymbol{s}} = 0$), and VFC can collapse to the K-means solution under tiny subgroups due to its inverse subgroup-cluster mass term (Appendix D.3). For COVA-FC, the induced hard assignment avoid empty subgroup–cluster cells in the reported settings, yielding strictly positive SB.

D.3 DEEPER COMPARISON WITH VFC & FAIRKM

Why VFC is the closest baseline. Among the baselines, VFC is closest to COVA-FC in computational profile: both use soft assignment and per-instance parallel updates. We therefore compare the two methods more closely below.

Let $\mathbf{A}^{(t)} \in [0,1]^{n\times K}$ denote the soft assignment matrix at iteration t, and let $A_k^{(t)}$ be its k-th column. Let $V_j \in \{0,1\}^n$ be the indicator vector of subgroup j, and let $v_{j,p}$ be its p-th entry. Writing μ_j for the target proportion of subgroup j, $a_{p,k}$ for the clustering term, and L for the bound-update scaling parameter, VFC assignment update is

$$A_{p,k}^{(t+1)} \propto A_{p,k}^{(t)} \exp\left(\frac{-a_{p,k} - \lambda b_{p,k}(\mathbf{A}^{(t)})}{L}\right), \tag{37}$$

followed by row normalization over k, where

$$b_{p,k}^{(t)} = \sum_j \left(\frac{\mu_j}{\mathbf{1}^\top A_k^{(t)}} - \frac{\mu_j v_{j,p}}{V_j^\top A_k^{(t)}}\right). \tag{38}$$

High-λ breakdown in VFC. The term $V_j^\top A_k^{(t)}$ is the soft mass of subgroup j in cluster k. When some subgroup-cluster masses are small, the inverse-mass term in $b_{p,k}^{(t)}$ can become large. Since this term is multiplied by λ inside an exponential update, increasing λ can make the assignment update numerically extreme. In Table 8, VFC initially improves SP on `Communities`. However, at slightly larger λ, the soft update becomes non-finite and the implementation falls back to the K-means solution. COVA-FC avoids this inverse subgroup-cluster mass term.

Table 8: VFC behavior on Communities with $K = 10$, $L = 2$, as λ increases.

λ	Cost	SP	$\max_{p,k} \lvert\lambda b_{p,k}\rvert$
0	**56.99**	**0.0251**	–
1×10^{-11}	57.06	0.0240	2.56×10^{2}
2×10^{-11}	57.34	0.0216	5.12×10^{2}
4×10^{-11}	60.57	0.0141	1.02×10^{3}
6×10^{-11}	**56.99**	**0.0251**	1.54×10^{3}

Why FairKM is hard to extend to soft assignment. FairKM (Abraham et al., 2020) updates cluster assignment by hard reassignment A_i for each instance i, which can be viewed as solving

$$\arg\min_{A_i\in[K]} \left(\|x_i - \mu_{A_i}\|^2 + \lambda \cdot \textsc{FairnessPenalty}(A_i; \text{others})\right)$$

The fairness penalty is evaluated from hard membership counts, so this combinatorial update (i) requires evaluating a per-cluster fairness penalty for each candidate A_i, whose cost increases with both K and M, where $M = 2^q$; (ii) is difficult to parallelize across instances because each update changes the subgroup proportions within the affected clusters; and (iii) is non-differentiable in A_i, ruling out direct gradient-based scaling. A soft-assignment relaxation, replacing $\mathbb{I}(A_i = k)$ with A_{ik}, does not preserve the original hard-count-based update structure of FairKM and would require re-deriving the optimization. In contrast, our continuous soft-assignment formulation directly optimizes a differentiable subgroup-fairness surrogate.

D.4 COMPUTATIONAL EFFICIENCY AND SCALABILITY

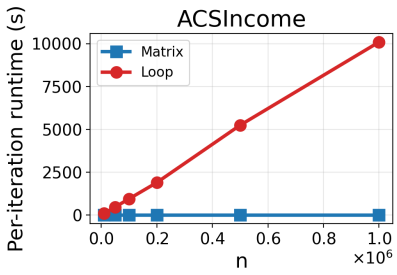


Figure 7: Per-iteration runtime of the loop and matrix implementations of the assignment update on `ACSIncome`.

GPU parallelization. We compare a per-instance loop implementation of the assignment update against our matrix implementation (and so parallel update) on `ACSIncome` with varying $n \in \{10\text{k}, 100\text{k}, 500\text{k}, 1\text{M}\}$. While the runtime of this loop version grows almost linearly with n, the matrix version grows much more slowly over the same range, confirming that the per-instance assignment updates are effectively parallelized on GPU (Figure 7).

In contrast, most baselines do not admit the same instance-wise decomposable assignment update: FRA, FCBC, and FCA rely on LP-based optimization, and FairKM uses hard reassignment steps depending on all cluster centroids and sizes of subgroups. Among the baselines, VFC is the only parallelizable baseline.

Scalability. To examine the scalability of COVA-FC, we compare its runtime on a large-scale dataset with that of VFC (Ziko et al., 2021), the fastest baseline in Table 3. We use the `ACSIncome` dataset with $n = 1{,}664{,}500$ instances, feature dimension $d = 7$, and $q = 3$ sensitive attributes (gender, marital status, and race with categories White/Black/Asian/Other), which yields 16 subgroups. We run each method to reach a comparable SP level. The results in Table 9 show that COVA-FC runs $3.22\times$ faster than VFC.

Table 9: Comparison between COVA-FC and VFC on `ACSIncome` in terms of fairness level (SP) and runtime.

Dataset	Method	SP	Runtime (sec)
`ACSIncome`	**COVA-FC✓**	0.19	179.72
	VFC (Ziko et al., 2021)	0.18	578.75

D.5 APPLICATION TO IMAGE FEATURES ON CELEBA

We additionally evaluate COVA-FC on the `CelebA` dataset to test its applicability in an image-feature setting. We use four binary sensitive attributes, which induce 16 subgroups, and extract feature vectors using a pre-trained ResNet-18 model. We then apply COVA-FC to these extracted features in the same way as for the other datasets. The results in Figure 8 show that COVA-FC can be applied without modification to image-derived feature representations and still improves subgroup fairness as the fairness penalty increases.

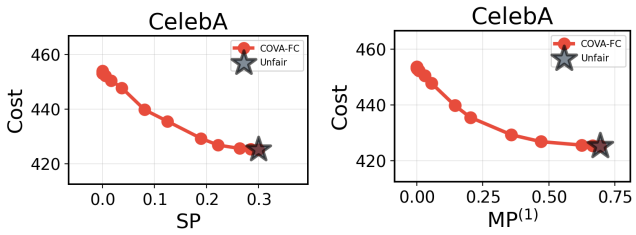


Figure 8: Cost-fairness trade-off on the `CelebA` dataset with four binary sensitive attributes (16 subgroups), using image features extracted by a pre-trained ResNet-18 model.

E ABLATION STUDIES

E.1 EFFECTS OF THE CONTINUOUS RELAXATION

Four variants of the surrogate. Recall that $\mathrm{CR}(\mathbf{A})$ has two discrete components: the worst-case subgroup membership s (relaxed by $\mathbf{v} \in \mathbb{S}^{M_+}$) and the hard cluster assignment $\mathbf{A}$ (relaxed by replacing $\mathbb{I}(A_i = k)$ with A_{ik}). Depending on which component is relaxed, we have four variants:

$$\mathrm{CR}(\mathbf{A}) := \max_{\boldsymbol{s}\in\mathcal{S}_+} \max_{\beta\in\mathcal{B}_\infty} \tfrac{1}{2n} \textstyle\sum_{i=1}^n \left(\sum_{k=1}^K \beta_k \mathbb{I}(A_i = k)\right) (c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}}),$$

$$\mathrm{CR}_{\mathbf{v}}(\mathbf{A}) := \max_{\mathbf{v}\in\mathbb{S}^{M_+}} \max_{\beta\in\mathcal{B}_\infty} \tfrac{1}{2n} \textstyle\sum_{i=1}^n \left(\sum_{k=1}^K \beta_k \mathbb{I}(A_i = k)\right) \mathbf{v}^\top (c_i - \bar{c}),$$

$$\mathrm{CR}_{\mathbf{A}}(\mathbf{A}) := \max_{\boldsymbol{s}\in\mathcal{S}_+} \max_{\beta\in\mathcal{B}_\infty} \tfrac{1}{2n} \textstyle\sum_{i=1}^n \left(\sum_{k=1}^K \beta_k A_{ik}\right) (c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}}),$$

$$\overline{\mathrm{CR}}(\mathbf{A}) := \max_{\mathbf{v}\in\mathbb{S}^{M_+}} \max_{\beta\in\mathcal{B}_\infty} \tfrac{1}{2n} \textstyle\sum_{i=1}^n \left(\sum_{k=1}^K \beta_k A_{ik}\right) \mathbf{v}^\top (c_i - \bar{c}).$$

CR is the original formulation (both hard), $\mathrm{CR}_{\mathbf{v}}$ relaxes only the worst-subgroup selection using $\mathbf{v}$, $\mathrm{CR}_{\mathbf{A}}$ relaxes only $\mathbf{A}$, and $\overline{\mathrm{CR}}$ relaxes both. For the two variants with relaxed worst-subgroup selection, $\mathrm{CR}_{\mathbf{v}}$ and $\overline{\mathrm{CR}}$, we implement the adversarial term using the equivalent matrix variable $U \in \mathcal{U} := \mathrm{conv}\{\mathbf{v}\beta^\top : \mathbf{v} \in \mathbb{S}^{M_+},\ \beta \in \mathcal{B}_\infty\}$, as in Eq. (10). The same naming applies to the general $\mathcal{W}$ by replacing the max over s with a max over $m \in \mathcal{M}_+$ and using the corresponding c_{im} and $\bar{c}_m$. $\overline{\mathrm{CR}}$ is the final surrogate used in COVA-FC; the other three serve as ablations below.

$\mathrm{CR}_{\mathbf{v}}$ vs. $\mathrm{CR}_{\mathbf{A}}$. We compare $\mathrm{CR}_{\mathbf{v}}$ (hard $\mathbf{A}$) with $\mathrm{CR}_{\mathbf{A}}$ (soft $\mathbf{A}$) on the four benchmark datasets. Table 10 shows that $\mathrm{CR}_{\mathbf{A}}$ produces fair clusters, whereas $\mathrm{CR}_{\mathbf{v}}$ often fails to reduce the fairness gap due to the non-smooth indicator $\mathbb{I}(A_i = k)$.

Table 10: Comparison between $\mathrm{CR}_{\mathbf{v}}$ and $\mathrm{CR}_{\mathbf{A}}$ in terms of Cost and SP.

Dataset	Variant	Cost	SP
`Adult`	$\mathrm{CR}_{\mathbf{v}}$	1.70	0.1046
	$\mathrm{CR}_{\mathbf{A}}$	2.52	**0.0002**
`Dutch`	$\mathrm{CR}_{\mathbf{v}}$	2.85	0.0888
	$\mathrm{CR}_{\mathbf{A}}$	3.94	**0.0001**
`Civilcomments`	$\mathrm{CR}_{\mathbf{v}}$	624.99	0.0711
	$\mathrm{CR}_{\mathbf{A}}$	657.84	**0.0012**
`Bank`	$\mathrm{CR}_{\mathbf{v}}$	1.52	0.0877
	$\mathrm{CR}_{\mathbf{A}}$	1.83	**0.0001**

$\mathrm{CR}_{\mathbf{A}}$ vs. $\overline{\mathrm{CR}}$. We compare $\mathrm{CR}_{\mathbf{A}}$, which explicitly selects the worst subgroup, with $\overline{\mathrm{CR}}$, which relaxes the worst-subgroup selection and is optimized using the equivalent U-based matrix formulation in Eq. (10), avoiding explicit subgroup-wise maximization during optimization. Table 11 shows that $\overline{\mathrm{CR}}$ matches the cost and SP of $\mathrm{CR}_{\mathbf{A}}$ at substantially lower runtime, with larger gains when $|\mathcal{W}|$ is larger (`Adult`: 16 subgroups, `Civilcomments`: 24).

Table 11: Comparison between $\mathrm{CR}_{\mathbf{A}}$ and $\overline{\mathrm{CR}}$ in terms of Cost, SP, and Runtime.

Dataset (# Subgroups)	Variant	Cost	SP	Runtime (sec)
`Adult` (16)	$\mathrm{CR}_{\mathbf{A}}$	2.53	0.0002	201.66
	$\overline{\mathrm{CR}}$	2.52	0.0002	**132.80** ($\times$1.52 faster)
`Dutch` (4)	$\mathrm{CR}_{\mathbf{A}}$	3.94	0.0001	34.60
	$\overline{\mathrm{CR}}$	3.94	0.0001	**32.00** ($\times$1.08 faster)
`Civilcomments` (24)	$\mathrm{CR}_{\mathbf{A}}$	657.84	0.0012	224.80
	$\overline{\mathrm{CR}}$	657.84	0.0012	**89.40** ($\times$2.51 faster)
`Bank` (3)	$\mathrm{CR}_{\mathbf{A}}$	1.83	0.0001	42.12
	$\overline{\mathrm{CR}}$	1.83	0.0001	**39.43** ($\times$1.07 faster)

$\mathrm{CR}_{\mathbf{v}}$ vs. $\overline{\mathrm{CR}}$. We further isolate the role of relaxing $\mathbf{A}$ by comparing $\overline{\mathrm{CR}}$ with $\mathrm{CR}_{\mathbf{v}}$, which uses the relaxed worst-subgroup formulation but keeps the hard indicator on $\mathbf{A}$. With (μ, U) fixed, $\mathrm{CR}_{\mathbf{v}}$ admits the closed-form update $A_i \leftarrow \arg\min_{k\in[K]} \left(\|x_i - \mu_k\|^2 + \frac{\lambda}{2}\left[(c_i - \bar{c})^\top U\right]_k\right)$, $i \in [n]$.

Figure 9 shows that $\overline{\mathrm{CR}}$ yields a stable decrease and clear convergence, whereas $\mathrm{CR}_{\mathbf{v}}$ exhibits oscillations and fails to converge, because hard assignment updates cause abrupt changes in the objective and can also lead to unstable updates of U.

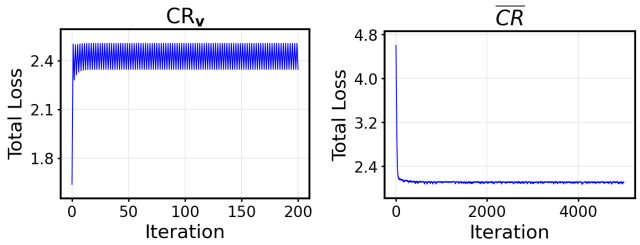


Figure 9: Plots of the total loss (clustering cost $+\lambda$ fairness penalty) across epochs on `Adult`, using $\mathrm{CR}_{\mathbf{v}}$ (left) and $\overline{\mathrm{CR}}$ (right) as the fairness penalty.

In addition to the total loss, we also track the fairness gap $\Delta(\mathbf{A}; \mathcal{W})$ over iterations. Figure 10 presents the results across multiple values of λ : (top) $\mathrm{CR}_{\mathbf{v}}$ with hard cluster assignment as the fairness penalty. (bottom) $\overline{\mathrm{CR}}$ with soft cluster assignment as the fairness penalty. The number of iterations can differ across settings because each run is terminated once the objective in Eq. (9) satisfies the convergence criterion (see Section 5.1). We observe that $\mathrm{CR}_{\mathbf{v}}$ (top) (i) exhibits oscillations, (ii) attains substantially larger fairness gap than $\overline{\mathrm{CR}}$, and (iii) does not show a clear improvement in the gap even when using a large λ. In contrast, $\overline{\mathrm{CR}}$ (bottom) reduces the gap clearly and produces fair clusters.

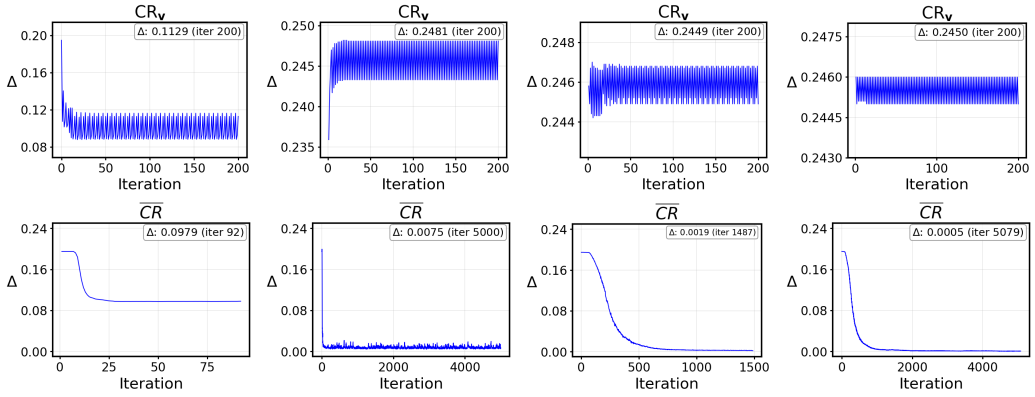


Figure 10: Plots of $\Delta(\mathbf{A}; \mathcal{W})$ across iterations on `Adult` dataset when using $\mathrm{CR}_{\mathbf{v}}$ (top) and $\overline{\mathrm{CR}}$ (bottom) as the fairness penalty. (left to right) λ ranges from 0.3 to 50.

Overall, these results indicate that using *hard cluster assignment harms optimization stability even when the worst-subgroup selection is relaxed*, and often fails to produce fair clusters in practice.

E.2 Soft-to-Hard rounding

Concentration of soft assignment. We examine how concentrated each soft cluster assignment A_i becomes during optimization. Specifically, for each $i \in [n]$, we compute the maximum probability $\max_{k\in[K]} A_{ik}$, and plot histograms of these values across all instances at 100, 300, 500, and 700 iterations in Figure 11. As optimization proceeds, the distribution of $\max_{k\in[K]} A_{ik}$ increasingly concentrates near 1, indicating that each A_i becomes sharply peaked and thus close to a hard cluster assignment. By Proposition B.1, this concentration also implies a small discrepancy between the soft surrogate and the fairness gap after argmax rounding.

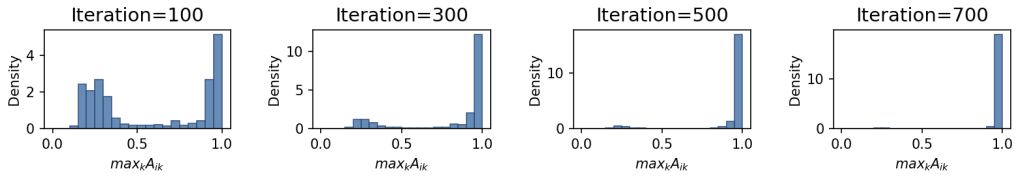


Figure 11: Histograms of the maximum soft cluster assignment, $\max_{k\in[K]} A_{ik}$, over all instances $i \in [n]$ at 100, 300, 500, and 700 iterations.

Fairness gap Δ after rounding. By Proposition B.1, $\overline{\text{CR}}$ on the soft assignment and Δ after argmax rounding are close when the learned soft assignment is close to one-hot, which we observe empirically in Figure 11. Figure 12 shows a clear positive relation: smaller values of $\overline{\text{CR}}$ are associated with smaller values of Δ, supporting $\overline{\text{CR}}$ as an effective surrogate for optimizing fair hard clustering in practice.

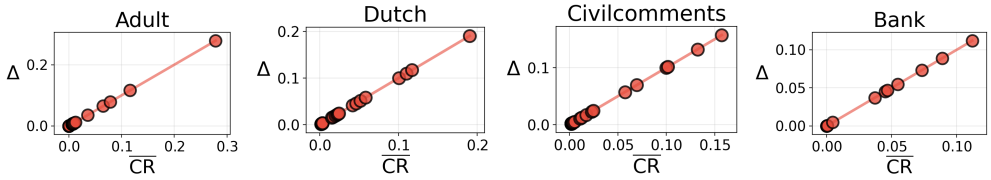


Figure 12: Plots of $\overline{\text{CR}}$ computed on the learned soft assignment versus Δ computed on the induced hard assignment across different values of λ on (left to right) `Adult`, `Dutch`, `Civilcomments`, and `Bank`.

E.3 Discussion of subgroup-size weighting and small-subgroup fairness

The weight $\pi_{\boldsymbol{s}}$ in the subgroup-fairness gap $\Delta(\mathbf{A})$ (Definition 2.1) reduces the influence of small subgroups to the objective, so substantial disparities within them may remain. We examine how weighting α affects disparities within small and large subgroups using a generalized fairness gap

For each observed subgroup $\boldsymbol{s} \in \{0,1\}^q$, let $G_{\boldsymbol{s}}(\mathbf{A}) := \sum_{k=1}^{K} |p_k(\mathbf{A}) - p_{k|\boldsymbol{s}}(\mathbf{A})|$ be a raw fairness gap. Then we consider the generalized gap

$$\Delta_\alpha(\mathbf{A}) := \max_{\boldsymbol{s}: n_{\boldsymbol{s}}>0} \pi_{\boldsymbol{s}}^{\alpha} G_{\boldsymbol{s}}(\mathbf{A}), \quad \alpha \in [0,1].$$

Setting $\alpha = 0$ corresponds to the unweighted fairness gap in Appendix A.1, whereas $\alpha = 1$ recovers the fairness gap in Definition 2.1. Thus, α controls the trade-off between reducing disparities in small subgroups and prioritizing fairness for larger subgroups.

To evaluate the effect of α, we replace each centered subgroup term $(c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}})$ in $\overline{\text{CR}}$ of Eq. (4) with $(c_{i\boldsymbol{s}} - \bar{c}_{\boldsymbol{s}})/\pi_{\boldsymbol{s}}^{1-\alpha}$. We use `Communities` with $K = 10$ because many of its full subgroups contain only a few instances. We categorize the observed subgroups by size: Small ($n_{\boldsymbol{s}} < K$), Medium ($K \le n_{\boldsymbol{s}} < 2K$), and Large ($n_{\boldsymbol{s}} \ge 2K$), and compare the distribution of $G_{\boldsymbol{s}}(\mathbf{A})$ across these groups.

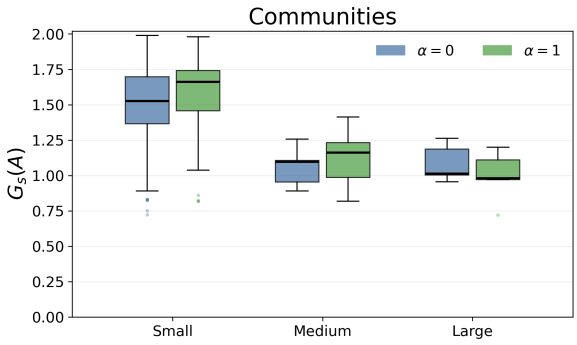


Figure 13: Distribution of the raw subgroup gap $G_{\boldsymbol{s}}(\mathbf{A})$ on `Communities` for Small, Medium, and Large subgroups, comparing $\alpha = 0$ and $\alpha = 1$.

We first compare $\alpha = 0$ and $\alpha = 1$ at comparable SP levels. As shown in Figure 13, $\alpha = 0$ yields lower median disparities for small subgroups, whereas $\alpha = 1$ yields a lower median for large subgroups. Small subgroups also exhibit a wider spread of disparities than medium and large subgroups under both settings. These results illustrate the trade-off between reducing small-subgroup disparities and prioritizing larger subgroups through α.

Table 12: Effect of the generalization parameter α on `Communities`.

α	Cost	SP	$\text{MP}^{(1)}$	$\text{MP}^{(2)}$	$n_{\boldsymbol{s}^\star_\alpha}$	$G_{\boldsymbol{s}^\star_\alpha}$
0	58.57	0.025	0.32	0.30	1	1.99
1/2	57.84	0.025	0.29	0.29	36	1.32
1	57.84	0.025	0.29	0.28	51	0.97

We further compare $\alpha \in \{0, 1/2, 1\}$, tuning λ for each α to obtain comparable SP levels. SP and MP are evaluated using the original $\pi_{\boldsymbol{s}}$-weighted metrics. For each α, choose $\boldsymbol{s}^\star_\alpha \in \arg\max_{\boldsymbol{s}: n_{\boldsymbol{s}} > 0} \pi^\alpha_{\boldsymbol{s}} G_{\boldsymbol{s}}(\mathbf{A})$, a subgroup attaining the maximum in $\Delta_\alpha(\mathbf{A})$. We report the size $n_{\boldsymbol{s}^\star_\alpha}$ and unweighted disparity $G_{\boldsymbol{s}^\star_\alpha}(\mathbf{A})$ of this subgroup in Table 12. These values show how the size and disparity of the worst-case subgroup vary with α.

As α increases from 0 to 1, the worst-case subgroup size increases from 1 to 51, while the marginal fairness gaps decrease at comparable SP levels. We use $\alpha = 1$ by default to reduce the influence of tiny subgroups on the objective, while smaller α places greater emphasis on their disparities.

E.4 SENSITIVITY AND OPTIMIZATION

We conduct the ablations on `Adult` with $K = 10$ unless varied, following the preprocessing in Section 5.1. Each comparison changes one component while keeping the remaining settings fixed. To compare complete optimization trajectories, all runs perform 10,000 updates without early stopping, using scheduled step sizes for both the assignment and centroid updates. Unless initialization itself is varied, we use k-means++ initializations with seeds $\{0, 1, 2\}$ shared across the compared settings.

Effect of λ. We first examine whether increasing the fairness penalty consistently reduces SP, using the same initializations and optimization schedule across λ values. In Figure 14, mean SP decreases monotonically from $\lambda = 1$ to 500, while cost increases. This illustrates how λ controls the cost–fairness trade-off within this range.

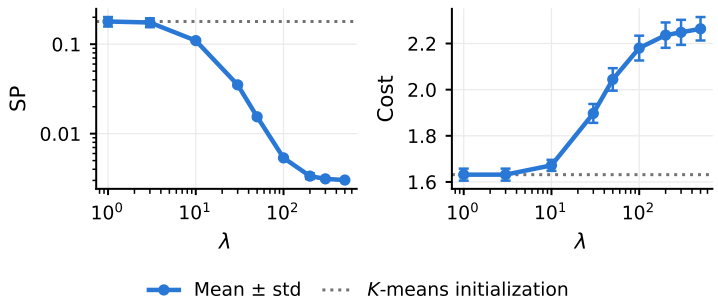


Figure 14: SP and cost versus λ on `Adult`. Error bars show mean $\pm$ standard deviation over three initializations. Dotted lines mark their initial means.

Sensitivity to initialization. To test whether these results depend on a favorable starting point, we repeat each setting in $\lambda \in \{10, 50, 100, 500\}$ with ten k-means++ initializations. As shown in Figure 15, different initializations yield similar final clustering costs and SP values at each λ, indicating that COVA-FC is robust to initialization across the tested penalty levels.

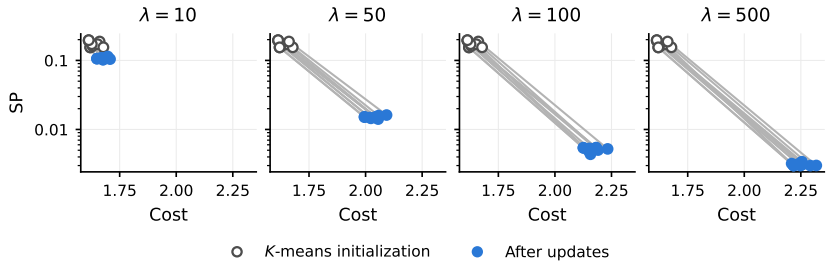


Figure 15: Initialization sensitivity for $\lambda \in \{10, 50, 100, 500\}$. Connected hollow and filled markers show the initial and final cost-SP pairs.

Effect of step-size decay. At $\lambda = 100$, we compare the decaying step sizes η_t defined in Eq. (25) of Appendix B.3 with the constant step size η_1, using the same three initializations. The stationarity residual $\|\mathcal{G}_t\|$ is the norm of the projected gradient mapping in Eq. (29) for the unregularized objective. As shown in Figure 16, constant step sizes η_1 leave the objective close to its initial value, whereas decaying step sizes η_t substantially reduce both the objective and the stationarity residual.

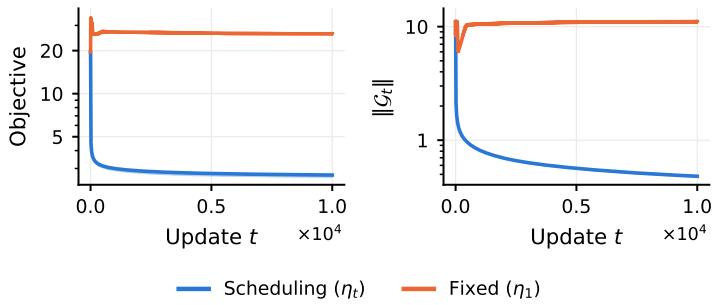


Figure 16: Step-size ablation at $\lambda = 10^2$: objective (left) and stationarity residual $\|\mathcal{G}_t\|$ from Eq. (29) (right). Lines and shaded bands show the mean and range over three matched initializations.

Sensitivity to K. To assess whether the cost–fairness trade-off observed at $K = 10$ persists for other cluster counts, we also evaluate $K = 5$ and $K = 20$, each over three K-means initializations. Across all tested values of K, increasing λ lowers SP at higher clustering cost in Table 13, indicating that this result is not specific to $K = 10$.

Table 13: Sensitivity to the number of clusters K on `Adult`. Cost and SP, mean $\pm$ standard deviation over three K-means initializations.

	$\lambda = 0$		$\lambda = 10$		$\lambda = 100$		$\lambda = 500$	
K	Cost	SP	Cost	SP	Cost	SP	Cost	SP
5	2.71±0.31	0.152±0.008	2.79±0.29	0.077±0.005	3.07±0.37	0.0026±0.0009	3.10±0.37	0.0022±0.0011
10	1.63±0.03	0.179±0.022	1.67±0.02	0.110±0.004	2.18±0.05	0.0054±0.0001	2.26±0.05	0.0030±0.0000
20	1.12±0.01	0.182±0.004	1.14±0.01	0.145±0.003	1.71±0.01	0.0111±0.0007	1.91±0.03	0.0033±0.0003