Robust Fair Clustering with Group Membership Uncertainty Sets

Sharmila Duppala
University of Maryland

Juan Luque
University of Maryland

John P. Dickerson
Arthur AI

Seyed A. Esmaeili
University of Chicago

Abstract

We study the canonical fair clustering problem where each cluster is constrained to have close to population-level representation of each group. Despite significant attention, the salient issue of having incomplete knowledge about the group membership of each point has been superficially addressed. In this paper, we consider a setting where the assigned group memberships are noisy. We introduce a simple noise model that requires a small number of parameters to be given by the decision maker. We then present an algorithm for fair clustering with provable *robustness* guarantees. Our framework enables the decision maker to trade off between the robustness and the clustering quality. Unlike previous work, our algorithms are backed by worst-case theoretical guarantees. Finally, we empirically verify the performance of our algorithm on real world datasets and show its superior performance over existing baselines.

1 INTRODUCTION

Machine learning and algorithmic-based decision-making systems have seen a remarkable proliferation in the last few decades. These systems are used in financial crime detection (Nicholls et al., 2021; Kumar et al., 2022), loan approval (Sheikh et al., 2020; Arun et al., 2016), automated hiring systems (Mahmoud et al., 2019; Van den Broek et al., 2021), and recidivism prediction (Travaini et al., 2022; Ghasemi et al., 2021). The clear effect of these applications on the welfare of individuals and groups coupled with recorded instances of algorithmic bias and harm (Danks and London, 2017; Panch et al., 2019) has made fairness—in its many forms under different interpretations, with its many definitions—a prominent consideration in algorithm design.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

Thus, it is unsurprising that fair *unsupervised learning* has received great interest in the AI/ML, statistics, operations research, and optimization communities—including *fair clustering*. Clustering is a central problem in AI/ML and operations research and arguably the most fundamental problem in unsupervised learning writ large. The literature in fair clustering has produced a significant number of publications spanning a wide range of fairness notions (see, e.g., Awasthi et al., 2022, for an overview). However, the most prominent of the fairness notions that were introduced is the *group fairness* notion due to Chierichetti et al. (2017), Bercea et al. (2018), and Bera et al. (2019). Since our paper is concerned with this notion in particular, for ease of exposition we will simply refer to it as fair clustering. In fair clustering, each point belongs to a demographic group and therefore each demographic group has some percentage representation in the entire dataset.[1] Like in agnostic (ordinary or "unfair") clustering the dataset is partitioned into a collection of clusters. However, unlike agnostic clustering each cluster must have a proportional representation of each group that is close to the representation in the entire dataset. For example, if the dataset consists of groups A and B at 30% and 70% representation, respectively. Then each cluster in the fair clustering should have a $(30 \pm \epsilon)\%$ and $(70 \pm \epsilon)\%$ representation of groups A and B, respectively.

One can see a significant advantage behind group fairness. Each cluster has close to dataset-level representation[2] of each group, so any outcome associated with any cluster will affect all groups proportionally, satisfying the disparate impact doctrine (Feldman et al., 2015), as discussed by Chierichetti et al. (2017). Despite the attractive properties of group fairness, it requires complete knowledge of each point's membership in a group. In

[1] As a simple example, the demographic groups could be based on income. Therefore, each point would belong to an income bracket and each income bracket would have some percentage representation (e.g., 20% belong to group "$\leq$ USD\$30k," 10% belong to group "$\geq$ USD\$250k," and so on) of the dataset.

[2] And, ideally, *population-level* representation—yet this may not hold in common machine learning applications, where proportionally sampling an underlying population to form a training dataset requires deep nuance. For a discussion of biased sampling and its implications in fair machine learning, we direct the reader to Barocas et al. (2023, Chapters 4 & 6).

practice—say, in an advertising setting where membership is estimated via a machine learning model, or in a lending scenario where membership may be illegal to estimate at train time—knowledge of group membership may range from noisy, to adversarially corrupted, to completely unknown. This salient problem has received significant attention in fair *classification* (see, e.g., Awasthi et al., 2020, 2021; Wang et al., 2020; Hashimoto et al., 2018; Kallus et al., 2022; Lamy et al., 2019). However, this important consideration has not received significant attention in fair *clustering* with the exception of the theoretical work of Esmaeili et al. (2020) that introduced uncertain group membership and the empirical work of Chhabra et al. (2023), who provide a data-driven approach to achieve robustness against adversarial perturbations on fair clustering systems.

Our paper addresses the practical modeling shortcomings of both Esmaeili et al. (2020) and Chhabra et al. (2023) and gives new theoretical worst-case guarantees. In short, the model of Esmaeili et al. (2020) makes the strong assumption of having probabilistic information about the group membership of each point in the dataset and the weak guarantee of having proportional representation of each group in every cluster but only in expectation. Further, Chhabra et al. (2023) looks into *black-box adversarial perturbations* on fair clustering. However, in their model it is assumed that only a fixed subset of points in the dataset will have their memberships perturbed and it is not clarified how this fixed subset is exactly decided. In Section 3.2 and Appendix D we give a more detailed comparison to these prior works and demonstrate their weaknesses.

Outline and Contributions: In Section 2, we briefly go over some prior work in fair clustering and other works in fair classification with emphasis on papers that tackle the incomplete/noisy group membership case. Then in Section 3, we formally describe the basic clustering setting and introduce our notation then we give an overview of the prior noise models of (Esmaeili et al., 2020) and (Chhabra et al., 2023). In Section 4, we present our noise model. Our model requires a small number of parameters as input instead of full probabilistic information for each point. In fact, as a special case it can be given only one parameter that represents the bound on the maximum number of incorrectly assigned group memberships in the dataset. Based on the framework of robust optimization, we then define the *robust fair clustering* problem for the k-center objective. In Section 5, we present our theoretically grounded algorithm to solve the robust fair k-center problem. Our algorithms require making careful observations about the structure of a robust fair solution and represents a novel addition to the existing fair clustering algorithms. Finally, in Section 6 we validate the performance of our algorithm on real world datasets and show that it has superior performance in comparison to the existing methods.

2 ADDITIONAL RELATED WORK

We will focus on the fairness notion most relevant to us in fair clustering, specifically where the solution is constrained to have proportional group representation in each cluster (e.g., Chierichetti et al., 2017; Bercea et al., 2018; Bera et al., 2019; Dickerson et al., 2023; Wang et al., 2023; Zeng et al., 2023, and others). Under the assumption that group memberships are perfectly known, this notion is well-investigated. For example, Backurs et al. (2019) gives faster scalable algorithms for this problem to handle large datasets. Bera et al. (2019) have considered a variant of this problem when each point is allowed to belong to more than one group simultaneously. Further, variants of this notion in non-centroid based clustering have also been considered. Ahmadian et al. (2020) address the same fairness notion in correlation clustering whereas Kleindessner et al. (2019) address it in spectral clustering, and Knittel et al. (2023b,a) address it in hierarchical clustering.

The problem of incomplete and imperfect knowledge of group memberships has received significant attention in fair classification. Awasthi et al. (2020) study the effects on the equalized odds notion of Hardt et al. (2016) when the group memberships are perturbed. Awasthi et al. (2021) study the effects of using a classifier to predict the group membership of a point on the bias of downstream ML tasks. Kallus et al. (2022) study a similar problem but focusing mainly on assessing the disparate impact in various applications when the group memberships are not unavailable and have to be predicted instead. Robust optimization methods were used to obtain fair classifiers under the setting of noisy group memberships by Wang et al. (e.g., 2020) and unavailable group memberships by Hashimoto et al. (e.g., 2018). While our problem falls under the robust optimization framework, our techniques are very different.

3 PRELIMINARIES AND PREVIOUS NOISE MODELS

In this section we go through preliminary background, notation, and previously introduced noise models in clustering.

3.1 Preliminaries

Let $\mathcal{P}$ be a set of n points in a metric space with distance function $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_{\geq 0}$. In k-center clustering, the goal is to select a set of centers S from $\mathcal{P}$ of at most k points and an assignment $\phi : \mathcal{P} \to S$ minimizing the clustering cost which is $\text{cost}(S, \phi) := \max_{j \in \mathcal{P}} d(j, \phi(j))$, i.e., the cost is the maximum distance between a point and its assigned

center. Since the clustering cost is the maximum distance between a point and its center, we also refer to that cost as the clustering radius or just radius. Clearly, in the ordinary k-center problem, ϕ will assign each point $j \in \mathcal{P}$ to its closest center in S, i.e., $\phi(j) = \arg\min_{i \in S} d(j, i)$. On the other hand, finding ϕ is non-trivial in more general k-center variants when constraints are imposed, as ϕ may assign points to centers that are further away to satisfy the imposed constraint.

We index the set of ℓ many demographic groups that exist in the dataset by $\mathcal{H} = \{1, 2, \ldots, \ell\}$. Following the fair clustering literature we associate a specific color with each group (Chierichetti et al., 2017; Bercea et al., 2018; Bera et al., 2019). Therefore, we use the words group and color interchangeably. Let $\mathcal{P}_h$ denote the subset of points in $\mathcal{P}$ that are assigned color h. Each point $j \in \mathcal{P}$ belongs to exactly one color from the set of colors $\mathcal{H}$. We can equivalently describe this assignment using the function $\chi : \mathcal{P} \to \mathcal{H}$, such that for any $j \in \mathcal{P}_h$, the color assignment for j would be $\chi(j) = h$. We denote the total number of points of color h by $n_h = |\mathcal{P}_h|$, it follows that $\sum_{h \in \mathcal{H}} n_h = n$. Further, given a solution (S, ϕ), for each $i \in S$, C_i denotes the set of points assigned to center i (i.e., cluster i) and $C_{i,h}$ denotes the subset of points in that cluster belonging to group h. The FAIR-k-CENTER problem (e.g., Chierichetti et al., 2017; Bercea et al., 2018; Bera et al., 2019; Esmaeili et al., 2020) adds the following fairness constraint to the k-center objective, formally the optimization problem is:

$$\min_{S:|S| \leq k, \phi} \max_{j \in \mathcal{P}} d(j, \phi(j)) \tag{1a}$$

$$\forall i \in S, \forall h \in \mathcal{H} : l_h \leq \frac{|C_{i,h}|}{|C_i|} \leq u_h \tag{1b}$$

where l_h and u_h are proportion bounds that satisfy $0 < l_h \leq r_h \leq u_h < 1$ with r_h being the ratio (proportion) of group h in the entire set of points, i.e., $r_h := \frac{n_h}{n}$. Therefore, an instance of FAIR-k-CENTER is parametrized by the tuple $(\mathcal{P}, \chi, k, \mathcal{H}, \vec{l}, \vec{u})$.

3.2 Previous Noise Models in Fair Clustering

In this section we give more details about (Esmaeili et al., 2020) and (Chhabra et al., 2023), the two prior works which have considered robustness in fair clustering. (Esmaeili et al., 2020) introduced a probabilistic noise model where each point $j \in \mathcal{P}$ has a probability $p_{j,h} \in [0, 1]$ of belonging to group h, with $\sum_{h \in \mathcal{H}} p_{j,h} = 1$. While their algorithms satisfy proportional fairness constraints in expectation[3], the worst-case realization can significantly violate these constraints as noted earlier. In fact, in Appendix D.1 we show an example where a clustering of the given points satisfies fairness in expectation, but violates it completely in realization. This highlights a core deficiency in this model.

Chhabra et al. (2023) introduced an adversarial model where the adversary has access to a subset of points whose group memberships can be modified. However, they do not specify how this subset is selected. In their experiments, they independently sample points with equal probability and add them to this subset. We can construct instances where with high probability certain point combinations are never sampled in the subset, thereby heavily restricting the model's capability. We give a concrete discussion of this in Appendix D.2. Moreover, their algorithm does not have theoretical guarantees.

4 OUR NOISE MODEL AND PROBLEM STATEMENT

In our model we assume that there exists a number of points whose group memberships (colors) have been incorrectly assigned to other groups. The two main considerations in our model are that: (1) in general these incorrect assignments exhibit a heterogeneity across the groups and (2) that the incorrect assignments can be arbitrarily allocated across the dataset.

The first consideration is based on the fact that in many settings there exist group memberships that are more desirable than others and therefore individuals may misreport their group memberships as other more favorable groups (Krumpal, 2013). Further, the mechanism through which the group memberships were assigned may exhibit higher error rates for particular groups. For example, the method used to elicit group memberships may fail with higher rates on some particular groups. Therefore, noise exhibits heterogeneity across the groups and an effective noise model should capture that.

The second consideration is based on the fact that incorrect group assignments could arise from a set of possibilities and therefore unlike Chhabra et al. (2023) we should not assume knowledge of these particular noisy points. The (noise) perturbations in the group assignments could have resulted from a process similar to iid noise as done in Mehrotra and Vishnoi (2022) and Mehrotra and Celis (2021). At another extreme, the group memberships could have been assigned using a machine learning classifier which predicts the group memberships, in that case if the classifier's errors are localized to a specific region in the feature space[4] then clearly the group membership perturbations do not act similar to random noise. Further, note that both scenarios are empirically well-motivated and could possibly occur in the same dataset simultaneously.

[3] Since each point has some probability of belonging to each specific group, one can calculate the expected number of points belonging to a specific group in a clustering by simply adding the points' probabilities in that cluster.

[4] This could be the case, if the training dataset happens to be particularly scare in that region.

Therefore, an effective noise model should not assume knowledge of the spatial noise distribution and allow noise to be arbitrarily allocated across the dataset.

Now, we delve into the formal description of the model. For a given color $h \in \mathcal{H}$ we associate two parameters m_h^+ and m_h^-, the first is the maximum number of points that were mistakenly assigned group memberships other than h and the second is the maximum number of points that were mistakenly assigned to group h.

Further, for a given set of value m_h^+ and m_h^-, by definition the consistency of the values requires that the following inequalities should be satisfied:

$$\forall h \in \mathcal{H} : m_h^+ \leq \sum_{g \in \mathcal{H}, g \neq h} m_g^- \tag{2}$$

$$\forall h \in \mathcal{H} : m_h^- \leq \sum_{g \in \mathcal{H}, g \neq h} m_g^+ \tag{3}$$

In words, the first inequality (2) simply states that no color should "gain" more points than the total number of points "lost" by the other groups. Similarly, the second inequality (3) states that if a color loses some number of points than the rest of the colors must gain at least the same amount in total. Note that since no color can lose more points than it has, an additional set of inequalities is also implied, namely $\forall h \in \mathcal{H} : m_h^- \leq n_h$. We did not list it as we assume that any given set values of m_h^- always satisfy it.

The values of m_h^+ and m_h^- lead to new possible group membership assignments (colorings) $\hat{\chi}$ other than the original coloring χ. Following the language of robust optimization (Ben-Tal et al., 2009), the set of all possible colorings $\hat{\chi} : \mathcal{P} \to \mathcal{H}$ that result from a given set of values $\{m_h^+, m_h^-\}_{h \in \mathcal{H}}$ is referred to as the *uncertainty set* $\mathcal{U}$. We use $\{\mathcal{P}_h\}_{h \in \mathcal{H}}$ to denote the color partition of $\mathcal{P}$ where $j \in \mathcal{P}_h$ has color h and we define $\chi^{-1}(h) = \mathcal{P}_h$. We can analogously define the partition $\{\hat{\mathcal{P}}_h\}_{h \in \mathcal{H}}$ for the assignment $\hat{\chi}$ where $\hat{\chi}^{-1}(h) = \hat{\mathcal{P}}_h$. More formally, given a valid set of noise parameters $\{m_h^+, m_h^-\}_{h \in \mathcal{H}}$ satisfying inequalities (2) and (3) the uncertainty set $\mathcal{U}$ is

$$\mathcal{U} = \{\hat{\chi} \mid \hat{\chi} \text{ satisfies (5a) – (5b)}\} \tag{4}$$

$$\forall h \in \mathcal{H} : |\mathcal{P}_h \setminus \hat{\mathcal{P}}_h| \leq m_h^- \tag{5a}$$

$$\forall h \in \mathcal{H} : |\hat{\mathcal{P}}_h \setminus \mathcal{P}_h| \leq m_h^+ \tag{5b}$$

Proposition 4.1. *For any instance with noise parameters $\{m_h^+, m_h^-\}_{h \in \mathcal{H}}$, the group uncertainty set $\mathcal{U}$ is defined as the set of all group assignments $\hat{\chi} : \mathcal{P} \to \mathcal{H}$ that satisfy the constraints in* (5a) *and* (5b).

Proof. Let $\hat{\chi}$ be a feasible assignment in the uncertainty set and let $\hat{\mathcal{P}}_h = \hat{\chi}^{-1}(h)$ be the collection of points assigned color h by $\hat{\chi}$. For any color h, $\mathcal{P}_h \setminus \hat{\mathcal{P}}_h$ denotes the set of points that are assigned color h by χ and a different color $g \neq h$ by $\hat{\chi}$. Clearly, by definition of m_h^- the first constraint should hold, i.e.

$$|\mathcal{P}_h \setminus \hat{\mathcal{P}}_h| \leq m_h^- \tag{6}$$

Furthermore, $\hat{\mathcal{P}}_h \setminus \mathcal{P}_h$ denotes the set of points that are assigned color h by $\hat{\chi}$ but were assigned a different color $g \neq h$ by χ. By definition of m_h^+ the second constraint holds, i.e.

$$|\hat{\mathcal{P}}_h \setminus \mathcal{P}_h| \leq m_h^- \tag{7}$$

□

The above simply states that any $\hat{\chi}$ coloring (element) in the uncertainty set should result in an assignment where (i) the number of points that are assigned a color h by χ but actually belong to a color $g \neq h$ should not exceed m_h^- and (ii) the number of points that are assigned a color $g \neq h$ by χ but actually have color h is no more than m_h^+.

For the case of two colors (denote them by red and blue), we naturally have $m_{\text{red}}^+ = m_{\text{blue}}^-$ and $m_{\text{red}}^- = m_{\text{blue}}^+$. To see that, note that using inequalities (2) and (3) it follows that $m_{\text{red}}^+ \leq m_{\text{blue}}^- \leq m_{\text{red}}^+$ and therefore $m_{\text{red}}^+ = m_{\text{blue}}^-$. Similarly, one can show that $m_{\text{red}}^- = m_{\text{blue}}^+$. The new implied equalities are natural as they simply state that what one color loses is gained by the other and vice versa.

To give a sense of the resulting group assignments implied by a given set of values m_h^+ and m_h^- consider the two color toy example shown in Figure 1. In this example, we have $m_{\text{red}}^- = 2$ whereas $m_{\text{blue}}^- = 1$. Further, from the previous discussion since we have two colors then we immediately have $m_{\text{red}}^+ = 1$ and $m_{\text{blue}}^+ = 2$. The figure shows the many possible colorings that can result from the given noise values, note that even for this simple example there are a total of 12 possibilities which is larger than the number of given points 4. A robust fair clustering has to achieve fairness over all possible colorings (all colorings in the uncertainty set).

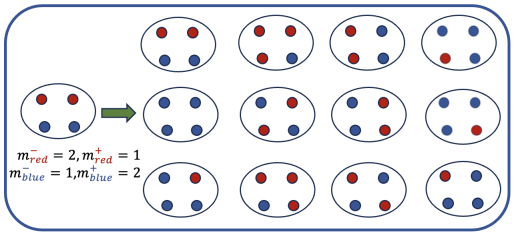

Figure 1: All colorings in the uncertainty set of a toy example with 4 points and $m_{\text{red}}^- = 2, m_{\text{blue}}^- = 1$ are shown.

We further note that a simple possible assignment of the color parameters would set them all to the same value, i.e., $\forall h \in \mathcal{H} : m_h^+ = m_h^- = m$. This essentially states that any

color can increase or decrease by m points. Figure 2 shows the same previous example where all noise parameters have been set to 2. Clearly, the number of possible colorings (size of the uncertainty set) has increased from 12 to 16.

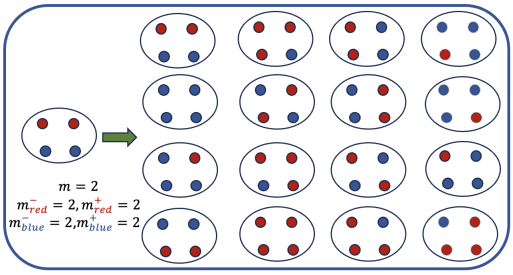

Figure 2: All colorings in the uncertainty set for the same toy example, now with $m = 2$. Note the new color assignments in the bottom row where the two (originally) blue points become red.

We are now ready to state our problem. Given an instance of *fair clustering* $(\mathcal{P}, \chi, k, \mathcal{H}, \vec{l}, \vec{u})$ along with noise parameters $\{m_h^+, m_h^+\}_{h\in\mathcal{H}}$. The objective of the RobustFair-k-Center problem is to find a clustering that minimizes the k-center objective while ensuring that the fairness constraints are satisfied for every possible coloring in the uncertainty set $\mathcal{U}$. Formally, the optimization problem of RobustFair-k-Center is

$$\min_{S:|S|\leq k,\phi} \max_{j\in\mathcal{P}} d(j,\phi(j)) \tag{8a}$$

$$\forall \hat{\chi}\in\mathcal{U}, \forall i\in S, h\in\mathcal{H}: l_h \leq \frac{|C_{i,h}(\hat{\chi})|}{|C_i|} \leq u_h \tag{8b}$$

where $C_{i,h}(\hat{\chi}) := \hat{\chi}^{-1}(h)\cap C_i$ denotes the subset of points in C_i which have been assigned to group h by a coloring $\hat{\chi}\in\mathcal{U}$. We also define a λ-violating solution as one where the fairness constraints of (8b) are violated by at most λ, formally a λ-violating solution satisfies

$$\forall \hat{\chi}\in\mathcal{U}, \forall i\in S, h\in\mathcal{H}: l_h - \lambda \leq \frac{|C_{i,h}(\hat{\chi})|}{|C_i|} \leq u_h + \lambda \tag{9}$$

Clearly, smaller λ implies a smaller violation of the fairness constraints and any value of $\lambda \geq 1$ is vacuous.

Finally, we note that our model requires the decision maker to specify a total of at most 2ℓ many parameters unlike Esmaeili et al. (2020) which needs a total of at least $(\ell-1)\cdot n$ many parameters (necessarily growing with the size of the dataset). The values in our model can be simply set using prior knowledge and statistics. Furthermore, if the decision maker does not posses fine-grained knowledge about the noise parameters m_h^+, m_h^- for each group then deciding one value m and setting $m_h^+ = m_h^- = m$ to upper bound the total change in any group would be sufficient. While this would increase the size of the uncertainty set, a clustering that is robust fair under an uncertainty set remains robust fair under a more restricted one. More formally, given a problem instance and two uncertainty sets $\mathcal{U}'$ and $\mathcal{U}$, if $\mathcal{U}'\subset\mathcal{U}$ then it follows immediately that a λ-violating solution under $\mathcal{U}$ is also a λ-violating solution under $\mathcal{U}'$. However, we note that while one may always expand the uncertainty set to ensure fairness for higher noise values it would come at an expense. Specifically, a robust solution would have to ensure fairness over a larger uncertainty set and that would in general lead the optimization objective (which is the clustering cost/quality) to be degraded.

5 ALGORITHM AND THEORETICAL ANALYSIS

We start this section by making a collection of mathematical observations that are essential for our algorithm. First, note that the size of the uncertainty set $|\mathcal{U}|$ can grow exponentially in the size of the dataset n. For example, for the two color case with $m_h^+ = m_h^- = m$ for both colors, the size of $\mathcal{U}$ is $\sum_{0\leq i,j\leq m}\binom{n_1}{i}\binom{n_2}{j} \geq \left(\frac{n}{2m}\right)^m$ where n_1 and n_2 are the number of points for the first and second color. For a reasonable choice of $m = \alpha n$ where α is a fractional constant in $(0,1)$, it is straight forward to see that the size of $|\mathcal{U}| = \Omega(c^n)$ where c is a constant strictly greater than 1. This implies that we can have an exponential set of constraints in (8b). Our first critical observation is that we can replace this exponential set by an equivalent polynomially sized set of constraints.

Lemma 5.1. *For any instance of* RobustFair-k-Center *the constraints* (8b) *are equivalent to* (10a) *and* (10b).

$$\forall i\in S, h\in\mathcal{H}: \frac{|C_{i,h}(\chi)| + m_h^+}{|C_i|} \leq u_h \tag{10a}$$

$$\forall i\in S, h\in\mathcal{H}: \frac{|C_{i,h}(\chi)| - m_h^-}{|C_i|} \geq l_h \tag{10b}$$

Proof. To show that the constraints (8b) are equivalent to (10a) and (10b), we have to show that for any feasible clustering $\{C_1, C_2, \ldots, C_{k'}\}$ with $k' \leq k$ satisfying (8b) must also satisfy (10a) and (10b) and vice versa.

First we show the forward direction, i.e., if $\{C_1, C_2, \ldots, C_{k'}\}$ satisfies (8b) then it also satisfies (10a) and (10b). Since the fairness constraints in (8b) must hold for all points in the uncertainty set $\mathcal{U}$, they are also valid for the worst-case $\hat{\chi}$ in $\mathcal{U}$ i.e.,

$$\forall \hat{\chi}\in\mathcal{U}, \quad \frac{|C_{i,h}(\hat{\chi})|}{|C_i|} \leq u_h \implies \max_{\hat{\chi}\in\mathcal{U}} \frac{|C_{i,h}(\hat{\chi})|}{|C_i|} \leq u_h.$$

However, in each cluster C_i, we know that the maximum number of points from color h that are incorrectly labeled as h but actually belong to a different group g given by $\min\{m_h^+, \sum_{g\in\mathcal{H},g\neq h}|C_{i,g}(\hat{\chi})|\}$. Therefore, for each $1 \leq$

$i \leq k'$ and $h \in \mathcal{H}$, we have

$$\max_{\hat{\chi} \in \mathcal{U}} \frac{|C_{i,h}(\hat{\chi})|}{|C_i|} = \frac{|C_{i,h}(\hat{\chi})| + \min\{m_h^+, \sum_{\substack{g \in \mathcal{H} \\ g \neq h}} |C_{i,g}(\hat{\chi})|\}}{|C_i|} = \frac{|C_{i,h}(\hat{\chi})| + m_h^+}{|C_i|} \leq u_h.$$

The last equality is from the fact in Observation B.1 that $\sum_{g \in \mathcal{H}, g \neq h} |C_{i,g}(\hat{\chi})| > m_h^+$

Next we prove that if a clustering satisfies constraints in (8b) then it also satisfies (10b). Since the clustering is feasible we know that,

$$\forall \hat{\chi} \in \mathcal{U}, \quad \frac{|C_{i,h}(\hat{\chi})|}{|C_i|} \geq l_h \implies \min_{\hat{\chi} \in \mathcal{U}} \frac{|C_{i,h}(\hat{\chi})|}{|C_i|} \geq l_h.$$

We know that in each C_i, the maximum number of points mistakenly assigned a color h but actually belonging to group $g \neq h$ is $\min\{m_h^-, |C_{i,h}(\chi)|\}$. However from the first part of Observation B.1 it is clear that, for any cluster C_i, we have $|C_{i,h}(\chi)| > m_h^-$ which implies that $\min\{m_h^-, |C_{i,h}(\chi)|\} = m_h^-$. Therefore, we have

$$\min_{\hat{\chi} \in \mathcal{U}} \frac{|C_{i,h}(\hat{\chi})|}{|C_i|} = \frac{|C_{i,h}(\chi)| - \min\{m_h^-, |C_{i,h}(\chi)|\}}{|C_i|} = \frac{|C_{i,h}(\chi)| - m_h^-}{|C_i|} \geq l_h.$$

This concludes the forward direction of our proof. Next, we show that if (10b) and (10a) holds, then Equation (8b) also holds. We know that for any $1 \leq i \leq k'$, C_i satisfies the upper bound constraints in (10a). This implies that any feasible assignment $\hat{\chi}$ contains at most $|C_{i,h}(\chi)| + m_h^+$ points of color h in C_i. From (5b) in Proposition 4.1 we know that, for any assignment in the uncertainty set at most m_h^+ additional points are assigned to color h. Therefore, we have

$$\forall h \in \mathcal{H} : \frac{|C_{i,h}(\chi)| + m_h^+}{|C_i|} \leq u_h$$
$$\implies \forall \hat{\chi} \in \mathcal{U} : \frac{|C_{i,h}(\hat{\chi})|}{|C_i|} \leq \frac{|C_{i,h}(\chi)| + m_h^+}{|C_i|} \leq u_h.$$

For the case of lower bound constraints we know that C_i satisfies the constraints in (10b). Therefore, we can say that any feasible assignment cannot have less than $|C_{i,h}(\chi)| - m_h^-$ points of color h in C_i. Therefore, for any $1 \leq i \leq k'$,

$$\forall h \in \mathcal{H} : \frac{|C_{i,h}(\chi)| - m_h^-}{|C_i|} \geq l_h$$
$$\implies \forall \hat{\chi} \in \mathcal{U}, \quad \frac{|C_{i,h}(\hat{\chi})|}{|C_i|} \geq \frac{|C_{i,h}(\chi)| - m_h^-}{|C_i|} \geq l_h$$

as desired. Therefore, the constraints in (8b) are equivalent to (10a) and (10b). □

We call the constraints (10a) and (10b) the *robust fairness* constraints. To see what the lemma means, consider a color h and its upper proportion bound u_h. The lemma essentially states that instead of ensuring that the solution satisfies $\frac{|C_{i,h}(\hat{\chi})|}{|C_i|} \leq u_h$ for all colorings $\hat{\chi} \in \mathcal{U}$ as done in (8b), we may instead take $C_{i,h}(\chi)$ (which uses the given coloring χ) and add the highest (worst-case) increase in the number of points that can be gained by color h which is m_h^+ and satisfy a single constraint of $\frac{|C_{i,h}(\chi)| + m_h^+}{|C_i|} \leq u_h$ in (10a) instead. A similar statement can be made about the lower bound l_h in (10b).

As a result of Lemma 5.1, it follows that a λ-violating solution of constraint (9) is only required to satisfy the following reduced constraints.

$$\forall i \in S, h \in \mathcal{H} : \frac{|C_{i,h}(\chi)| + m_h^+}{|C_i|} \leq u_h + \lambda \tag{11a}$$

$$\forall i \in S, h \in \mathcal{H} : \frac{|C_{i,h}(\chi)| - m_h^-}{|C_i|} \geq l_h - \lambda \tag{11b}$$

Another critical observation is that upper and lower proportion bounds that would have a non-empty set of feasible solutions in ordinary fair clustering might lead to an infeasible ROBUSTFAIR-k-CENTER instance unless the bounds are relaxed by a sufficient margin. To see that, consider an instance of ROBUSTFAIR-k-CENTER with two red and two blue points and noise parameters $m_{\text{red}}^+ = m_{\text{blue}}^+ = m = 1$. If we set the proportion bounds to $l_{\text{red}} = l_{\text{blue}} = 1/2$ and $u_{\text{red}} = u_{\text{blue}} = 1/2$, then clearly we would have an ordinary (non-robust) fair clustering solution. However, one can see through Lemma 5.1 that no feasible robust fair solution exists. Now, if we relax the bounds to $l_{\text{red}} = l_{\text{blue}} = 1/4$ and $u_{\text{red}} = u_{\text{blue}} = 3/4$, then a single cluster containing all four points becomes a feasible solution. More formally, the proportions bounds have to be relaxed exactly as shown in the following observation:

Observation 5.1. *For any instance of the* ROBUSTFAIR-k-CENTER *problem, a feasible solution exists if and only if for every group $h \in \mathcal{H}$, $u_h \geq \frac{n_h + m_h^+}{n}$ and $l_h \leq \frac{n_h - m_h^-}{n}$.*

5.1 Our Algorithm: ROBUSTALG

To solve our problem we employ a two-stage approach where the initial stage selects the centers and the second stage assigns the points to the centers. While various prior papers in fair clustering Bera et al. (2019); Bercea et al. (2018); Esmaeili et al. (2020, 2021) use a similar two-stage approach, the centers used in the first stage are selected by any vanilla (ordinary) k-center algorithm. In our case, it is actually critical that the centers are selected carefully by our algorithm (subroutine) GETCENTERS. In fact, in Section 5.2 we show how using another algorithm would break a critical step in our proof.

Since we are dealing with a k-center objective, the optimal radius (the maximum distance from any point to

its assigned center) for RobustFair-k-Center belongs to a finite set of possible values, i.e., the set of $\binom{n}{2}$ distance values. Therefore, our subroutine GetCenters (Algorithm 1) receives as input a "guessed" radius value R along with the entire set of points $\mathcal{P}$. GetCenters outputs a set of centers S. The procedure begins with all the points being unmarked, then in each iteration we add an arbitrary unmarked point j to S, and mark the all the points at a distance of at most $2R$ from j (this includes point j as well). We repeat this step until all the points are marked.

Algorithm 1 GetCenters

Input: Set of points $\mathcal{P}$, and a radius R
Output: Cluster centers S
$S \leftarrow \emptyset$
while $\mathcal{P} \neq \emptyset$ **do**
  Pick an arbitrary $j \in \mathcal{P}$ and $S \leftarrow S \cup \{j\}$
  $\mathcal{P} \leftarrow \mathcal{P} \setminus \text{Ball}(j, 2R)$
end while
return S

We will show that the centers returned by GetCenters when run at a sufficiently large value of R satisfy good properties that enable us to post-process them to obtain a *robust fair* clustering. Before that, it is important to introduce the feasibility linear program (LP) which takes a radius value R and a collection of centers S and assigns the points in $\mathcal{P}$ to centers in S. Since in a given fixed instance the set of centers S and radius value R can vary as inputs, we call it LP(S, R), its full details are shown below.

$$\text{LP}(S, R):$$

$$\forall j \in \mathcal{P} : \sum_{i \in S} x_{i,j} = 1 \tag{12}$$

$$\forall i \in S, h \in \mathcal{H} : \sum_{j \in \mathcal{P}^h} x_{i,j} + m_h^+ \leq u_h \sum_{j \in \mathcal{P}} x_{i,j} \tag{13}$$

$$\forall i \in S, h \in \mathcal{H} : \sum_{j \in \mathcal{P}^h} x_{i,j} - m_h^- \geq l_h \sum_{j \in \mathcal{P}} x_{i,j} \tag{14}$$

$$\forall i \in S, j \in \mathcal{P} : \text{ if } d(i,j) > 3R\text{: } x_{i,j} = 0, \text{else:} x_{i,j} \geq 0 \tag{15}$$

For each point $j \in \mathcal{P}$ and center $i \in S$, LP(S, R) has a decision variable $x_{i,j} \in [0, 1]$ which denotes the fractional assignment of point j to center i in S. LP(S, R) is more easily interpreted by considering the integral values of $x_{i,j} \in \{0, 1\}$ instead of $[0, 1]$. Therefore, constraint (12) simply states that each point should be assigned to exactly one center. Further, it follows that $\sum_{j \in \mathcal{P}^h} x_{i,j} = |C_{i,h}(\chi)|$ and $\sum_{j \in \mathcal{P}} x_{i,j} = |C_i|$, hence constraint (13) is simply imposing constraint (10a) of Lemma 5.1 to ensure that the upper proportion bounds are not violated. A similar reasoning follows for constraint (14). The last constraint (15) simply forbids assigning points j to centers i that are at a distance greater than $3R$, this is done by setting the assignment variables $x_{i,j} = 0$ if the distance $d(i, j) > 3R$ and otherwise allowing it to be in $[0, 1]$. Note that the LP(S, R) receives R as an input parameter but uses $3R$ in constraint (15).

LP(S, R) is elaborate and in fact it is not difficult to see that it might not be feasible for an arbitrary set of centers S and an arbitrary radius value R. Interestingly, we show that if GetCenters is run at a value of $R \geq R^*$ where R^* is the optimal radius (clustering cost) value then the set of centers $\hat{S}$ returned by GetCenters satisfies this LP at radius R, i.e., LP$(\hat{S}, R)$ is *feasible*. This is is shown in the following lemma. In fact, the lemma additionally shows that the number of centers in $\hat{S}$ is at most k, i.e., guaranteeing that we would not have more than k centers. The main idea in the proof is to show that an optimal robust fair solution (S^*, ϕ^*) of cost R^* can instead use the centers $\hat{S}$ at the expense of degrading the clustering cost to $3R^*$. This is done by moving clusters in (S^*, ϕ^*) to carefully chosen centers in $\hat{S}$. Note that the proof is non-constructive as it assumes knowledge of the optimal solution.

Lemma 5.2. *Let the optimal clustering cost (radius) be R^*, then if we set $R \geq R^*$ then* GetCenters *(Algorithm 1) returns a set $\hat{S}$ such that (1) $|\hat{S}| \leq k$ and (2) LP$(\hat{S}, R)$ is feasible.*

The above suggests that we may run GetCenters at different radius values R, by Lemma 5.2 once $R \geq R^*$ the returned centers $\hat{S}$ would have at most k centers and since LP$(\hat{S}, R)$ would be feasible by running it we would obtain a feasible assignment. In fact, our algorithm RobustAlg (Algorithm 2) does that and uses binary search over the set of pairwise distances to find the smallest value of R where the conditions of Lemma 5.2 are satisfied. The issue is that the feasible solution that we would obtain x^{LP} can be fractional, i.e., $\text{x}^{\text{LP}}_{i,j} \in [0, 1]$ and not necessarily $\in \{0, 1\}$. Therefore, we would have to round these fractional values into valid integral ones $\text{x}^{\text{Integ}}_{i,j} \in \{0, 1\}$. The following lemma shows that using the MaxFlow rounding scheme [5] Bercea et al. (2018); Dickerson et al. (2023) (see Appendix C for more details) we can obtain an integral assignment at no increase to the clustering cost and only for a slight change in the cluster sizes as shown in Lemma 5.3.

Lemma 5.3. *Let x^{Integ} be the integral assignment that results from running* MaxFlow *rounding over a fractional assignment x^{LP}, then (1) if $\text{x}^{LP}_{i,j} = 0$ then $\text{x}^{Integ}_{i,j} = 0$ and (2) for any center $i \in \hat{S}$ and group $h \in \mathcal{H}$,*

$$\left\lfloor |C_i^{LP}| \right\rfloor \leq |C_i^{Integ}| \leq \left\lceil |C_i^{LP}| \right\rceil,$$
$$\left\lfloor |C_{i,h}^{LP}| \right\rfloor \leq |C_{i,h}^{Integ}| \leq \left\lceil |C_{i,h}^{LP}| \right\rceil$$

where $|C_i^{LP}| = \sum_{j \in \mathcal{P}} \text{x}^{LP}_{i,j}$, $|C_i^{Integ}| = \sum_{j \in \mathcal{P}} \text{x}^{Integ}_{i,j}$, $|C_{i,h}^{LP}| = \sum_{j \in \mathcal{P}_h} \text{x}^{LP}_{i,j}$, and $|C_{i,h}^{Integ}| = \sum_{j \in \mathcal{P}_h} \text{x}^{Integ}_{i,j}$

[5] In short, in MaxFlow rounding we solve a network flow instance corresponding to a given clustering instance and fractional LP assignment x^{LP}. Then an integral flow is found and used to construct the rounded integral assignment x^{Integ}.

The fact that the clustering cost would not increase should be clear from guarantee (1) of the above lemma as it implies that $\mathrm{x}^{\text{Integ}}$ will only assign points j to centers i where they already had a non-zero assignment in the fractional solution, i.e., $\mathrm{x}^{\text{LP}}_{i,j} > 0$. The integral assignment $\mathrm{x}^{\text{Integ}}$ can immediately be used to construct the assignment function $\hat{\phi} : \mathcal{P} \to \hat{S}$. Therefore, our final solution is $(\hat{S}, \hat{\phi})$.

All that remains is the final guarantee on the solution $(\hat{S}, \hat{\phi})$. While it is clear that the radius is at most $3R^*$, the theorem below also shows that the violation in the fairness constraints is also bounded by a small value. Formally, we have the following theorem.

Theorem 5.1. ROBUSTALG *(Algorithm 2) is a* 3*-approximation algorithm for the* ROBUSTFAIR-k-CENTER *with fairness violation* $\lambda = \frac{2}{\sum_{h \in \mathcal{H}} m_h^-}$.

From the above theorem it is clear that for large values of $\sum_{h \in \mathcal{H}} m_h^-$ the violations would become smaller. In fact, if $\sum_{h \in \mathcal{H}} m_h^- = \omega(1)$ then $\lambda \to 0$ as $n \to \infty$.

Algorithm 2 ROBUSTALG

Input: An instance of FAIR-k-CENTER, $m_h^+, m_{h\ h \in \mathcal{H}}^-$.

Output: Clustering of points $(\hat{S}, \hat{\phi})$

1: Perform binary search to find the smallest radius R for which the set S returned by GETCENTERS$(\mathcal{P}, R)$ has at most k centers and LP(S, R) is feasible.
2: We call the set of centers returned at the smallest radius $\hat{S}$ and we call its associated fractional assignment x^{LP}.
3: Round x^{LP} to $\mathrm{x}^{\text{Integ}}$ using MAXFLOW rounding (see full details in Appendix C).
3: Construct the assignment $\hat{\phi} : \mathcal{P} \to \hat{S}$ as follows, for each $j \in \mathcal{P}$ set $\hat{\phi}(j) = i$ if $\mathrm{x}^{\text{Integ}}_{i,j} = 1$.
4: **return** $(\hat{S}, \hat{\phi})$

5.2 Failure When Using a Vanilla Clustering Algorithm

Here we show that LP(S, R) would not be feasible using centers selected by a vanilla clustering algorithm S_{vll}[6]. This is shown in the theorem below. The main idea behind this is that a vanilla clustering algorithm lacks adaptivity and therefore may select too many centers and since the constraints in LP(S, R) (specifically (13) and (14)) implicitly impose a lower bound on the cluster size there would not be enough points to assign to each center. While closing a subset of centers in S_{vll} might lead to a feasible LP(S_{vll}, R), knowing which ones to close without degrading the clustering cost is not straightforward to do. Our algorithm GETCENTERS avoids all of this and gives a simple way to find the set of centers and construct the final clustering solution.

[6]By a vanilla clustering algorithm we mean one which has some α approximation ratio for the ordinary clustering objective.

Theorem 5.2. *Given a set of centers selected by a vanilla* k*-center algorithm* S_{vll} *then for any arbitrarily large* R*, there may not exist a feasible solution to LP*(S_{vll}, R).

6 EXPERIMENTS

We conduct experiments on a commodity laptop (Ryzen 7 5800, 16GB RAM) using Python 3.6 and cplex 12.8 to solve LPs. Additional details and plots are available in Appendix E.

Datasets. We experiment on three datasets from the UCI repository (Dua and Graff, 2017): **Adult**, **Bank**, and **Census1990**. The datasets have 32k, 32k and 4.5k points with 5, 3, and 66 numerical features, respectively. Further, the colors in each dataset, i.e., sensitive attributes, are respectively a binary sex, a binary marital status, and a membership in one of three age buckets. The distances between any pair of points is set to the Euclidean distance between their normalized numerical features.

Parameters. The experiments use noise parameters $m_h^+ = m_h^- = m, \forall h \in \mathcal{H}$, where m ranges from $\frac{n}{100}$ to $\frac{n}{10}$. We set the proportions l_h and u_h to the respectively greatest and least values leading to feasible ROBUSTFAIR-k-CENTER instances, in accordance with Observation 5.1. Therefore, l_h and u_h are the same across instances in the same dataset. The number of centers is fixed at 10, i.e., $k = 10$.

We benchmark our proposed algorithm ROBUSTALG against two relevant baselines. Namely, *probabilistic* fair clustering (Esmaeili et al., 2020) and *deterministic* fair clustering (Bera et al., 2019). These three fair clustering algorithms are denoted as ROBUSTALG, PROBALG, and DETALG in the plots. Note that probabilistic fair clustering is an algorithm for two colors only so it is tested solely on **Adult** and **Bank**.

We evaluate the algorithms on their attained k-center objectives and fairness violation given by (9). For deterministic (and robust) fair clustering we obtain fairness violations by directly finding the corruption of point colors leading to the greatest fairness violation as described in Inequalities (11a) and (11b). Specifically, for each color, up to m points can essentially be added or subtracted.

In the probabilistic fair clustering work of Esmaeili et al. (2020), the noise model is different. Specifically, each point's color is corrupted with some probability. To ensure that our evaluation is fair we evaluate the probabilistic instances under their assumed noise model. Instances are set up so that each point's color is corrupted with probability m/n, leading to an *expected* m corruptions dataset-wide. In contrast, robust and deterministic fair clustering allow for up to m corruptions in *each color*. Finally, we sample 200 realizations of colors and directly report the mean fairness violations. However, probabilistic fair clustering says nothing about point correlations.

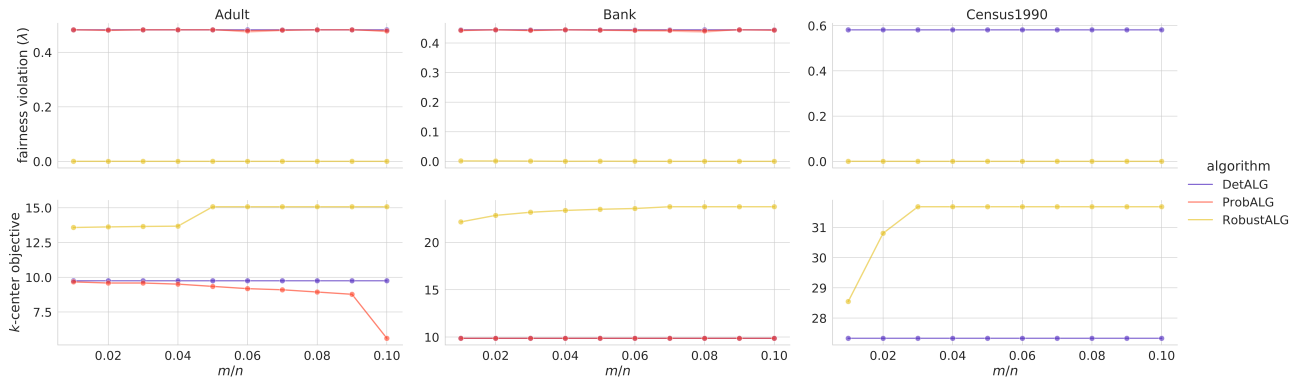

Figure 3: Plots for k-center objective and fairness violation as m/n increases. Over half of RobustAlg's pictured fairness violations are exactly zero; the rest fall between 10^{-5} and 10^{-3}. The 95% confidence interval around ProbAlg is shaded; however, it is faint because the fairness violations are very sharply concentrated around their plotted mean.

We exploit this fact as follows. First, sample $S_{i,h} \sim \text{Bernoulli}(m/n)$ for each cluster i and color h, and then, if $S_{i,h} = 1$, corrupt the colors of all points of color h assigned cluster i. Note that this is a generous evaluation, since an adversarial color assignment (as done in deterministic and robust fair clustering) can only lead to a higher violation.

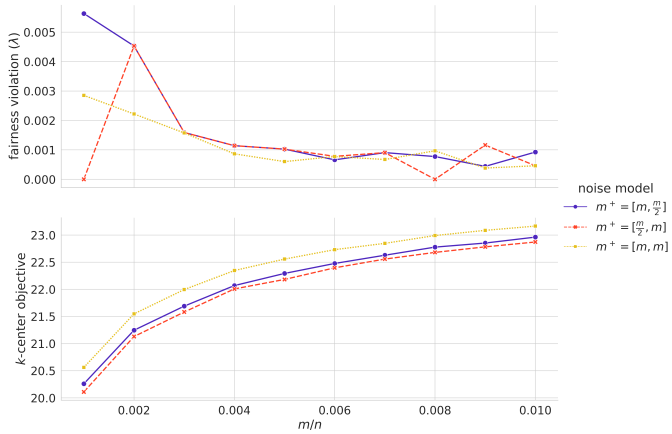

Figure 4: Comparison of the effect of increasing noise, given different noise model configurations.

Figure 3 shows the results of our experiments. RobustAlg *maintains fairness violations of zero and near-zero* across the board; unlike deterministic and probabilistic baselines which have fairness violations as large as $0.6 = 60\%$. In fact, in DetAlg the post-corruption ratio of some colors becomes 0 in **Adult** and in **Bank** (i.e., a color is completely absent from the cluster) or 1 in **Census1990** (i.e., a color is fully dominating the cluster). Note that these are the worst-possible fairness violations in all three datasets. ProbAlg nearly hits these worst-case violations as well.

The objective (clustering cost) of RobustAlg is greater and increases with m/n, as expected when targeting a more stringent notion of robustness. The break in the objective plot of Figure 3 occurs when RobustAlg opens fewer centers. As m increases, centers must have a greater number of points and thus fewer centers can receive points; otherwise applying all m corruptions to the smallest center results in (nearly) absent colors or (nearly) monochromatic centers, which produce large fairness violations.

Finally, Figure 4 concludes with experiments on **Bank** using three settings of noise parameters. We fix the l_h and u_h for all plots as described before. We denote the two colors in bank by 0 and 1. Further, we take m from $\frac{1}{1000}n$ to $\frac{1}{100}n$ but consider the (m_0^+, m_1^+) choices of $(m, m/2)$, $(m/2, m)$, and (m, m). The experiments validate our intuitions. First, (m, m) achieves the highest objective as it has the most corruptions (largest uncertainty set). Moreover, $(m/2, m)$ and $(m, m/2)$ both have the same number of total corruptions and their objectives are indeed comparable. In all cases, the fairness violations border on zero, agreeing with Theorem 5.1.

References

S. Ahmadian, A. Epasto, R. Kumar, and M. Mahdian. Clustering without over-representation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 267–275, 2019.

S. Ahmadian, A. Epasto, R. Kumar, and M. Mahdian. Fair correlation clustering. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4195–4205. PMLR, 2020.

K. Arun, G. Ishan, and K. Sanmeet. Loan approval prediction based on machine learning approach. *IOSR Journal of Computer Engineering*, 18(3):18–21, 2016.

P. Awasthi, M. Kleindessner, and J. Morgenstern. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1770–1780. PMLR, 2020.

P. Awasthi, A. Beutel, M. Kleindessner, J. Morgenstern, and X. Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 206–214, 2021.

P. Awasthi, B. Brubach, D. Chakrabarty, J. P. Dickerson, S. A. Esmaeili, M. Kleindessner, M. Knittel, J. Morgenstern, S. Samadi, A. Srinivasan, and L. Tsepenekas. Fairness in clustering. In *Conference on Artificial Intelligence (AAAI)*, 2022.

A. Backurs, P. Indyk, K. Onak, B. Schieber, A. Vakilian, and T. Wagner. Scalable fair clustering. In *International Conference on Machine Learning (ICML)*, pages 405–413. PMLR, 2019.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009.

S. Bera, D. Chakrabarty, N. Flores, and M. Negahbani. Fair algorithms for clustering. *Conference on Neural Information Processing Systems (NeurIPS)*, 32, 2019.

I. O. Bercea, M. Groß, S. Khuller, A. Kumar, C. Rösner, D. R. Schmidt, and M. Schmidt. On the cost of essentially fair clusterings. *arXiv preprint arXiv:1811.10319*, 2018.

A. Chhabra, P. Li, P. Mohapatra, and H. Liu. Robust fair clustering: A novel fairness attack and defense framework. *International Conference on Learning Representations (ICLR)*, 2023.

F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. *Conference on Neural Information Processing Systems (NeurIPS)*, 30, 2017.

D. Danks and A. J. London. Algorithmic bias in autonomous systems. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 17, pages 4691–4697, 2017.

J. Dickerson, S. A. Esmaeili, J. Morgenstern, and C. J. Zhang. Doubly constrained fair clustering. *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

D. Dua and C. Graff. Uci machine learning repository. 2017.

S. Esmaeili, B. Brubach, L. Tsepenekas, and J. Dickerson. Probabilistic fair clustering. *Conference on Neural Information Processing Systems (NeurIPS)*, 33:12743–12755, 2020.

S. Esmaeili, B. Brubach, A. Srinivasan, and J. Dickerson. Fair clustering under a bounded cost. *Advances in Neural Information Processing Systems*, 34:14345–14357, 2021.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 259–268, 2015.

M. Ghasemi, D. Anvari, M. Atapour, J. Stephen Wormith, K. C. Stockdale, and R. J. Spiteri. The application of machine learning to a general risk–need assessment instrument in the prediction of criminal recidivism. *Criminal Justice and Behavior*, 48(4):518–538, 2021.

A. Hagberg, D. Schult, P. Swart, D. Conway, L. Séguin-Charbonneau, C. Ellison, B. Edwards, and J. Torrents. Networkx. high productivity software for complex networks. *Webová strá nka https://networkx. lanl. gov/wiki*, 2013.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Conference on Neural Information Processing Systems (NeurIPS)*, 29, 2016.

T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, pages 1929–1938. PMLR, 2018.

N. Kallus, X. Mao, and A. Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

M. Kleindessner, S. Samadi, P. Awasthi, and J. Morgenstern. Guarantees for spectral clustering with fairness constraints. In *International Conference on Machine Learning*, pages 3458–3467. PMLR, 2019.

M. Knittel, M. Springer, J. Dickerson, and M. Hajiaghayi. Fair polylog-approximate low-cost hierarchical clustering. *Conference on Neural Information Processing Systems (NeurIPS)*, 2023a.

M. Knittel, M. Springer, J. P. Dickerson, and M. Hajiaghayi. Generalized reductions: making any hierarchical clustering fair and balanced with low cost. In *International Conference on Machine Learning (ICML)*, pages 17218–17242. PMLR, 2023b.

I. Krumpal. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4):2025–2047, 2013.

S. Kumar, R. Ahmed, S. Bharany, M. Shuaib, T. Ahmad, E. Tag Eldin, A. U. Rehman, and M. Shafiq. Exploitation of machine learning algorithms for detecting financial crimes based on customers' behavior. *Sustainability*, 14 (21):13875, 2022.

A. Lamy, Z. Zhong, A. K. Menon, and N. Verma. Noise-tolerant fair classification. *Conference on Neural Information Processing Systems (NeurIPS)*, 32, 2019.

A. A. Mahmoud, T. A. Shawabkeh, W. A. Salameh, and I. Al Amro. Performance predicting in hiring process and performance appraisals using machine learning. In *International Conference on Information and Communication Systems (ICICS)*, pages 110–115. IEEE, 2019.

A. Mehrotra and L. E. Celis. Mitigating bias in set selection with noisy protected attributes. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 237–248, 2021.

A. Mehrotra and N. K. Vishnoi. Fair ranking with noisy protected attributes. *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

J. Nicholls, A. Kuppa, and N.-A. Le-Khac. Financial cybercrime: A comprehensive survey of deep learning approaches to tackle the evolving financial crime landscape. *IEEE Access*, 9:163965–163986, 2021.

S. Nickel, C. Steinhardt, H. Schlenker, and W. Burkart. Ibm ilog cplex optimization studio—a primer. In *Decision Optimization with IBM ILOG CPLEX Optimization Studio: A Hands-On Introduction to Modeling with the Optimization Programming Language (OPL)*, pages 9–21. Springer, 2022.

T. Panch, H. Mattie, and R. Atun. Artificial intelligence and algorithmic bias: implications for health systems. *Journal of Global Health*, 9(2), 2019.

M. A. Sheikh, A. K. Goel, and T. Kumar. An approach for prediction of loan approval using machine learning algorithm. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 490–494. IEEE, 2020.

G. V. Travaini, F. Pacchioni, S. Bellumore, M. Bosia, and F. De Micco. Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction. *International Journal of Environmental Research and Public Health*, 19(17): 10594, 2022.

E. Van den Broek, A. Sergeeva, and M. Huysman. When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly*, 45(3), 2021.

J. Wang, D. Lu, I. Davidson, and Z. Bai. Scalable spectral clustering with group fairness constraints. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 6613–6629. PMLR, 2023.

S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. Jordan. Robust optimization for fairness with noisy protected groups. *Advances in neural information processing systems*, 33:5190–5203, 2020.

P. Zeng, Y. Li, P. Hu, D. Peng, J. Lv, and X. Peng. Deep fair clustering via maximizing and minimizing mutual information: Theory, algorithm and metric. In *Computer Vision and Pattern Recognition Conference (CVPR)*, pages 23986–23995, June 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [*Yes, we describe our model in Section 4 and the algorithms in Section 5.*]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [*Yes, we give the full theoretical analysis of our algorithm in Section 5.*]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [*Yes, we provide the anonymized source code along with our other supplementary material as a zip file.*]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [*Yes, we provide the theoretical analysis in Section 5. We include all our proofs in the Appendix of the supplementary material.*]
   (b) Complete proofs of all theoretical results. [*Yes, we include all our proofs in the Appendix of the supplementary material.*]
   (c) Clear explanations of any assumptions. [*Yes*]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [*Yes, we include the source code for our experiments in a .zip file which will be included in the supplementary material.*]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [*Not Applicable*]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [*Yes, we include this in our experiments section.*]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [*Yes*]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [*Yes, we include the necessary citations for our datasets.*]

(b) The license information of the assets, if applicable. [*Not Applicable*]

(c) New assets either in the supplemental material or as a URL, if applicable. [*Not Applicable*]

(d) Information about consent from data providers/curators. [*Not Applicable*]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [*Not Applicable*]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [*Not Applicable*]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [*Not Applicable*]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [*Not Applicable*]

A USEFUL FACT

Fact A.1. *For any positive real numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, the following holds*

$$\min_{i\in[n]} \frac{a_i}{b_i} \leq \frac{a_1 + a_2 + \cdots + a_n}{b_1 + b_2 + \cdots + b_n} \leq \max_{i\in[n]} \frac{a_i}{b_i} \tag{16}$$

Proof. Let $\tau_{\max} = \max_{i\in[n]} \frac{a_i}{b_i}$, therefore we have

$$\frac{a_1 + a_2 + \cdots + a_n}{b_1 + b_2 + \cdots + b_n} \leq \frac{\tau_{\max}(b_1 + b_2 + \cdots + b_n)}{b_1 + b_2 + \cdots + b_n} = \tau_{\max}$$

The lower bound can be proved similarly. □

B OMITTED PROOFS

Observation B.1. *Suppose that (S, ϕ) is robust fair solution to a given input instance and $u_h < 1$ and $l_h > 0$. Then the clustering $\{C_1, C_2, \cdots, C_{k'}\}$ of points induced by (S, ϕ) must satisfy the following,*

- *For any $i \in S$, $h \in \mathcal{H}$, $|C_{i,h}(\chi)| > m_h^-$.*
- *For any $i \in S$, $h \in \mathcal{H}$, $\sum_{g\neq h, g\in\mathcal{H}} |C_{i,g}(\chi)| > m_h^+$.*

Proof. We know that since the clustering is robust fair it must satisfy the constraints in Equation (8b) i.e.,

$$\forall i \in S, \forall h \in \mathcal{H} : \min \frac{|C_{i,h}(\hat{\chi})|}{|C_i|} \geq l_h \implies \min |C_{i,h}(\hat{\chi})| > 0$$

However, we know that there can be as much as m_h^- points that have incorrect group memberships from each color h. Therefore we have

$$\forall i \in S, \forall h \in \mathcal{H} : \min |C_{i,h}(\hat{\chi})| > 0 \implies |C_{i,h}(\chi)| - m_h^- > 0$$

Thus, we show the first part of the claim.

For each $i \in S$ and color $h \in \mathcal{H}$, using the lower bound on $|C_{i,g}(\chi)|$ obtained in the first part, we can derive a lower bound on $\sum_{g\in\mathcal{H}, g\neq h} |C_{i,g}(\chi)|$ by summing over all the groups $g \in \mathcal{H}$ and $g \neq h$ as follows,

$$\sum_{g\in\mathcal{H}, g\neq h} |C_{i,g}(\chi)| > \sum_{g\in\mathcal{H}, g\neq h} m_g^- \geq m_h^+$$

The last inequality is from the inequality (2) which says that the number of points gained by any group h is at most the total number of points lost by the remaining groups $g \neq h$ i.e., $\sum_{g\in\mathcal{H}, g\neq h} m_g^- \geq m_h^+$. Therefore, we get the desired bound. □

Observation 5.1. *For any instance of the* RobustFair-k-Center *problem, a feasible solution exists if and only if for every group $h \in \mathcal{H}$, $u_h \geq \frac{n_h + m_h^+}{n}$ and $l_h \leq \frac{n_h - m_h^-}{n}$.*

Proof. Suppose we have a clustering $\{C_1, \ldots, C_{k'}\}$) with $k' \leq k$. As before $C_{i,h}(\chi)$ denotes the set of points belonging to group $\mathcal{H}$ in the original coloring. Applying the Fact A.1, we get the following,

$$\min_{i\in[k']} \frac{|C_{i,h}(\chi)|}{|C_i|} \leq \frac{\sum_{i\in[k']} C_{i,h}(\chi)}{\sum_{i\in[k']} C_i} = \frac{|\mathcal{P}_h|}{|\mathcal{P}|} \leq \max_{i\in[k']} \frac{|C_{i,h}(\chi)|}{|C_i|} \tag{17}$$

Suppose that $u_h < \frac{n_h + m_h^+}{n}$ for some color, then by the definition of the robust optimization problem, it is possible to have $n_h + m_h^+$ many points of color h in the dataset. Accordingly, we it must be that for some color assignment $\frac{|\mathcal{P}_h|}{|\mathcal{P}|} = \frac{n_h + m_h^+}{n}$ but (17) implies that there exists some value $i \in [k']$ such that $\frac{|C_{i,h}(\chi)|}{|C_i|} \geq \frac{|\mathcal{P}_h|}{|\mathcal{P}|} = \frac{n_h + m_h^+}{n}$. Therefore, $|C_{i,h}(\chi)| > u_h |C_i|$ and therefore the solution is infeasible. The same argument can be made for the lower bound as well.

We will now show that if $\forall h \in \mathcal{H} : u_h \geq \frac{n_h+m_h^+}{n}, l_h \leq \frac{n_h-m_h^-}{n}$, then the problem must be feasible. To show feasibility we simply show one feasible solution. Specifically, a solution which is always feasible is a one cluster solution that includes all of the points, i.e. $\{C_1\} = \{\mathcal{P}\}$. Clearly, we have for any color h : $\frac{n_h-m_h^-}{n} \leq \frac{|\mathcal{P}_h|}{|\mathcal{P}|} \leq \frac{n_h+m_h^+}{n}$. Since the $u_h \geq \frac{n_h+m_h^+}{n}, l_h \leq \frac{n_h-m_h^-}{n}$, then the solution is feasible. □

Lemma 5.2. *Let the optimal clustering cost (radius) be R^*, then if we set $R \geq R^*$ then* GETCENTERS *(Algorithm 1) returns a set $\hat{S}$ such that (1) $|\hat{S}| \leq k$ and (2) LP$(\hat{S}, R)$ is feasible.*

Proof. We first show that for any $R \geq R^*$ the number of centers in $\hat{S}$ returned by Algorithm 1 is at most k. Each center $i^* \in S^*$ has at most one $i' \in \hat{S}$ from its optimal cluster. This is because any two centers in $\hat{S}$ are separated by a distance strictly greater than $2R$ and therefore no two centers in $\hat{S}$ are selected from the same optimal cluster. Therefore, we have $|\hat{S}| \leq |S^*| \leq k$.

To prove the second part of the lemma, it suffices to show that for any $R \geq R^*$, (i) there exists an assignment $\phi' : \mathcal{P} \to \hat{S}$ that assigns points in $\mathcal{P}$ to the centers in $\hat{S}$ and (ii) the cluster $\phi'^{-1}(i)$ corresponding to each $i \in \hat{S}$ is *robust fair*. In Lemma B.1, we show using a non-constructive proof that for any $R \geq R^*$ there exists a feasible solution $(\hat{S}, \phi')$ to our problem with a cost of at most $3R$ where each of the centers $i \in \hat{S}$ has a cluster that is *robust fair*.

If such a solution $(\hat{S}, \phi')$ exists then it immediately corresponds to feasible solution to the LP, this can be shown as follows: for each point $j \in \mathcal{P}$, set $x_{i,j} = 1$ if $\phi'(j) = i$ and $x_{i,j} = 0$ otherwise. Clearly, this x satisfies the constraints in (12). Moreover, each cluster $C_i = \phi'^{-1}(i)$ satisfies the constraints in (10b) and (10a) since it is robust fair, i.e.,

$$\forall i \in \hat{S}, h \in \mathcal{H} : |C_i \cap \mathcal{P}_h| - m_h^- \geq l_h |C_i|, \quad \text{and} \quad |C_i \cap \mathcal{P}_h| + m_h^+ \leq u_h |C_i|. \tag{18}$$

Furthermore, since $|C_i| = \sum_{j \in \mathcal{P}} x_{i,j}$ and $|C_i \cap \mathcal{P}_h| = \sum_{j \in \mathcal{P}_h} x_{i,j}$, the assignment x satisfies the constraints in (13) and (14) by substituting $|C_i| = \sum_{j \in \mathcal{P}} x_{i,j}$ and $|C_i \cap \mathcal{P}_h| = \sum_{j \in \mathcal{P}_h} x_{i,j}$ in the (18) we have

$$\forall i \in \hat{S}, h \in \mathcal{H} : \sum_{j \in \mathcal{P}_h} x_{i,j} - m_h^- \geq l_h \sum_{j \in \mathcal{P}} x_{i,j} \quad \text{and} \quad \sum_{j \in \mathcal{P}_h} x_{i,j} + m_h^+ \leq u_h \sum_{j \in \mathcal{P}} x_{i,j}.$$

Further note that by Lemma B.1 that each point is assigned to a center in $\hat{S}$ that is at most at a distance of $3R$ therefore (15) is satisfied as well. Therefore, we conclude that for any $R \geq R^*$, there exists a non-empty feasible solution to LP $(\hat{S}, R)$. □

Lemma B.1. *For any $R \geq R^*$ and set of centers $\hat{S}$ returned by the* GETCENTERS *subroutine, (i) there exists an assignment $\phi' : \mathcal{P} \to \hat{S}$ at a clustering cost of at most $3R$, i.e., cost$(\hat{S}, \phi') \leq 3R$ and (ii) the assignment leads each center $i \in \hat{S}$ to have a cluster that is robust fair.*

Proof. Suppose that we have an optimal solution (S^*, ϕ^*) with cost R^*. Each point j is assigned to some center $i^* \in S^*$ by the optimal assignment ϕ^*. Let C_{i^*} denote the set of points assigned to i^*, i.e., $C_{i^*} = \phi^{*-1}(i^*)$. We show the existence of an assignment $\phi' : \mathcal{P} \to \hat{S}$ from the set of points $\mathcal{P}$ to the set of centers $\hat{S}$ returned by Algorithm 1 such that (i) the assignment ϕ' has cost$(\hat{S}, \phi') \leq 3R$ and (ii) each cluster corresponding to a center $i \in \hat{S}$ is *robust fair*. Note that the proof is non-constructive since it assumes knowledge of the optimal solution.

We construct the assignment ϕ' as follows: ϕ' assigns all the points in each cluster C_{i^*} to the center $i' \in \hat{S}$ if i' belongs to the cluster C_{i^*}, i.e., if $\phi^*(i') = i^*$. It is possible that there are clusters C_{i^*} with no point $i' \in \hat{S}$. Therefore, we assign such clusters to a center i' in $\hat{S}$ where $d(i^*, i') \leq 2R$, i.e., i' is at a distance of at most $2R$ from the cluster's center. Since every point has at least one center in $\hat{S}$ at a distance of at most $2R$. This concludes the description of our assignment ϕ'.

Notice that each $i' \in \hat{S}$ gets assigned all the points associated with at least one center $i^* \in S^*$. This is because any two centers in $\hat{S}$ are separated by a distance strictly greater than $2R$ and therefore no two centers in $\hat{S}$ are selected from the same optimal cluster. Further, each center $i^* \in S^*$ gets assigned to some $i' \in \hat{S}$ at a distance of at most $2R$. We now prove that this new assignment has a cost of at most $3R$, i.e., cost$(\hat{S}, \phi') \leq 3R$. This holds because for any point $j \in \mathcal{P}$ we have:

$$\begin{aligned} d(j, \phi'(j)) &\leq d(j, \phi^*(j)) + d(\phi^*(j), \phi'(j)) && (19)\\ &\leq R^* + d(\phi^*(j), \phi'(j)) && (20)\\ &\leq R^* + 2R && (21)\\ &\leq 3R. && (22) \end{aligned}$$

Inequalities (19) follows from triangle inequality. Inequality (20) follows from the fact that ϕ^* is an optimal *robust fair* assignment. Finally, inequality (21) is from the fact that $d(\phi^*(j), \phi'(j)) \leq 2R$ since each center $i^* \in S^*$ is assigned to a center $i' \in \hat{S}$ at a distance of at most $2R$. It remains to show that for each $i' \in \hat{S}$ the corresponding cluster $\phi'^{-1}(i')$ is *robust fair*, i.e., satisfying the constraints (10b) and (10a). The following claim concludes the proof of this lemma.

Claim B.1. *The clustering induced by $(\hat{S}, \phi')$ leads each center $i' \in \hat{S}$ to be robust fair.*

Proof. Recall that C_{i^*} denotes the set of points that are assigned to center $i^* \in S^*$ in the optimal clustering (S^*, ϕ^*). Since (S^*, ϕ^*) is an optimal clustering to our problem, it must satisfy the constraints in (10a) and (10b). Therefore, we have the following:

$$\forall i^* \in S^*, \forall h \in \mathcal{H} : \frac{|C_{i^*,h}(\chi)| + m_h^+}{|C_{i^*}|} \leq u_h, \text{ and } \frac{|C_{i^*,h}(\chi)| - m_h^-}{|C_{i^*}|} \geq l_h.$$

For each $i' \in \hat{S}$, let $N(i')$ denote the set of centers $i^* \in S^*$ that are assigned to i' by ϕ'. For each $i' \in \hat{S}$ we can upper bound the proportion of any color as follows:

$$\begin{aligned}
\frac{|C_{i',h}| + m_h^+}{|C_{i'}|} &= \frac{\left(\sum_{i^* \in N(i')} |C_{i^*,h}|\right) + m_h^+}{\sum_{i^* \in N(i')} |C_{i^*}|} \\
&\leq \frac{\sum_{i^* \in N(i')} \left(|C_{i^*,h}| + m_h^+\right)}{\sum_{i^* \in N(i')} |C_{i^*}|} \\
&\leq \max_{i^* \in N(i')} \frac{|C_{i^*,h}| + m_h^+}{|C_{i^*}|} \leq u_h
\end{aligned}$$

Similarly, we can also lower bound the proportions:

$$\begin{aligned}
\frac{|C_{i',h}| - m_h^-}{|C_{i'}|} &= \frac{\left(\sum_{i^* \in N(i')} |C_{i^*,h}|\right) - m_h^-}{\sum_{i^* \in N(i')} |C_{i^*}|} \\
&\geq \frac{\sum_{i^* \in N(i')} \left(|C_{i^*,h}| - m_h^-\right)}{\sum_{i^* \in N(i')} |C_{i^*}|} \\
&\geq \min_{i^* \in N(i')} \frac{|C_{i^*,h}| - m_h^-}{|C_{i^*}|} \geq l_h
\end{aligned}$$

Note that the above inequalities follow from the Fact A.1 since for any center $i^* \in N(i')$, $\frac{|C_{i^*,h}| - m_h^-}{|C_{i^*}|} \geq l_h$ and $\frac{|C_{i^*,h}| + m_h^+}{|C_{i^*}|} \leq u_h$.

Furthermore, note that from the definition of ϕ' we know that $|N(i')| \geq 1$ for any i' in $\hat{S}$, i.e., there exists at least one cluster C_{i^*} in the optimal clustering that is assigned to i'. This concludes that each center $i' \in S$ is indeed *robust fair*, i.e., satisfies the constraints (10a) and (10b). □

□

Lemma B.2. *$(\hat{S}, \hat{\phi})$ returned by Algorithm 2 is a λ-violating solution where λ is at most $\frac{2}{\sum_{h \in \mathcal{H}} m_h^-}$.*

Proof. According to Equation (9) a λ-violating solution is only required to satisfy the following reduced constraints.

$$\forall i \in S, h \in \mathcal{H} : \frac{|C_{i,h}(\chi)| + m_h^+}{|C_i|} \leq u_h + \lambda \text{ and } \frac{|C_{i,h}(\chi)| - m_h^-}{|C_i|} \geq l_h - \lambda$$

Therefore, for any given clustering $\{C_i\}_{i \in S}$, we have the fairness violation λ as follows,

$$\lambda \leq \max_{h \in \mathcal{H}, i \in S} \left\{ \frac{l_h |C_i| - |C_{i,h}(\chi)| + m_h^-}{|C_i|}, \frac{|C_{i,h}(\chi)| + m_h^+ - u_h |C_i|}{|C_i|} \right\}$$

Before we delve into the proof we note that any LP solution x^{LP} that satisfies constraints Equation (12)-Equation (15), the following holds:

$$l_h|C_i^{\mathrm{LP}}| - |C_{i,h}^{\mathrm{LP}}| + m_h^- \leq 0 \tag{23}$$
$$|C_{i,h}^{\mathrm{LP}}| - u_h|C_i^{\mathrm{LP}}| + m_h^+ \leq 0 \tag{24}$$

Where Inequalities (23) and (24) follow from constraints (14) and (13), respecitvely, by simple algebreic manipulation.

We note further by Observation B.1 that $|C_i^{\mathrm{LP}}|$ satisfies:

$$|C_i^{\mathrm{LP}}| \geq \sum_{h\in\mathcal{H}} m_h^- \tag{25}$$

Now, let C_i^{Integ} denote the cluster corresponding to the set of points that are assigned to center $i \in S$. The fairness violation from the lower bound can be bounded as follows

$$\begin{aligned}
\frac{l_h|C_i^{\mathrm{Integ}}| - |C_{i,h}^{\mathrm{Integ}}| + m_h^-}{|C_i^{\mathrm{Integ}}|} &\leq \frac{l_h(|C_i^{\mathrm{LP}}|+1) - (|C_{i,h}^{\mathrm{LP}}|-1) + m_h^-}{\lfloor |C_i^{\mathrm{LP}}| \rfloor} \qquad \text{(by the bounds in in Lemma 5.3)} \\
&= \frac{l_h|C_i^{\mathrm{LP}}| - |C_{i,h}^{\mathrm{LP}}| + m_h^-}{\lfloor |C_i^{\mathrm{LP}}| \rfloor} + \frac{l_h+1}{\lfloor |C_i^{\mathrm{LP}}| \rfloor} \\
&\leq 0 + \frac{l_h+1}{\lfloor |C_i^{\mathrm{LP}}| \rfloor} \qquad \text{(by Inequality (23))} \\
&\leq \frac{l_h+1}{\sum_{h\in\mathcal{H}} m_h^-} \qquad \text{(by Inequality(25) since } \sum_{h\in\mathcal{H}} m_h^- \text{ is an integer)}
\end{aligned}$$

Similarly, we can bound the violation from the upper bound as follows

$$\begin{aligned}
\frac{|C_{i,h}^{\mathrm{Integ}}| + m_h^+ - u_h|C_i{}^{\mathrm{Integ}}|}{|C_i{}^{\mathrm{Integ}}|} &\leq \frac{(|C_{i,h}^{\mathrm{LP}}|+1) + m_h^+ - u_h(|C_i^{\mathrm{LP}}|-1)}{\lfloor |C_i^{\mathrm{LP}}| \rfloor} \qquad \text{(by the bounds in in Lemma 5.3)} \\
&= \frac{|C_{i,h}^{\mathrm{LP}}| - u_h|C_i^{\mathrm{LP}}| + m_h^+}{\lfloor |C_i^{\mathrm{LP}}| \rfloor} + \frac{1+u_h}{\lfloor |C_i^{\mathrm{LP}}| \rfloor} \\
&\leq 0 + \frac{u_h+1}{\lfloor |C_i^{\mathrm{LP}}| \rfloor} \qquad \text{(by Inequality (24))} \\
&\leq \frac{u_h+1}{\sum_{h\in\mathcal{H}} m_h^-} \qquad \text{(by Inequality(25) since } \sum_{h\in\mathcal{H}} m_h^- \text{ is an integer)}
\end{aligned}$$

Finally, we have the following bound on λ,

$$\lambda \leq \max_{h\in\mathcal{H}} \left\{ \frac{l_h+1}{\sum_{h\in\mathcal{H}} m_h^-}, \frac{1+u_h}{\sum_{h\in\mathcal{H}} m_h^-} \right\} < \frac{2}{\sum_{h\in\mathcal{H}} m_h^-}$$

The last inequality is due to the fact that $l_h \leq u_h < 1$. □

Theorem 5.1. ROBUSTALG *(Algorithm 2) is a* 3*-approximation algorithm for the* ROBUSTFAIR-k-CENTER *with fairness violation* $\lambda = \frac{2}{\sum_{h\in\mathcal{H}} m_h^-}$.

Proof. For any given instance of ROBUSTFAIR-k-CENTER, any non-empty solution $(\hat{S}, \hat{\phi})$ returned by ROBUSTALG (Algorithm 2) has a cost of at most $3R^*$ guaranteed by the fractional assignment returned by LP $(\hat{S}, \hat{\phi})$ as no point is fractionally assigned to any center at a distance more than $3R^*$. Further, part (1) of Lemma 5.3 guarantees that the radius would not increase, therefore the cost would still be at most $3R^*$. Moreover, from Lemma B.2 it follows that $(\hat{S}, \hat{\phi})$ is a λ-violating solution with $\lambda \leq \frac{2}{\sum_{h\in\mathcal{H}} m_h^-}$. This concludes the proof of the theorem. □

Theorem 5.2. *Given a set of centers selected by a vanilla k-center algorithm S_{vll} then for any arbitrarily large R, there may not exist a feasible solution to LP(S_{vll}, R).*

Proof. We prove this theorem using a counter example. Specifically, consider the instance in Figure 5 with $n = 4$ points belonging to red and blue groups. Let $k = 2$ and $l_{\text{red}} = l_{\text{blue}} = 1/4$ and $u_{\text{red}} = u_{\text{blue}} = 3/4$. The noise parameters are set to $m^+_{\text{red}} = m^-_{\text{red}} = 1$. The vanilla k-center algorithm selects the centers $S_{\text{vll}} = \{s_1, s_2\}$. Consider the lower bound constraints (14) in the corresponding LP with $S = S_{\text{vll}}$. This constraint requires that each center s_1 and s_2 must have a fractional assignment strictly greater than 1 from each color. This follows from the lower bound constraints in Equation (14) where each center in $s \in S_{\text{vll}}$ and $h \in \{\text{red,blue}\}$ must satisfy

$$|C^{\text{LP}}_{s,h}| \geq m^-_h + l_h|C^{\text{LP}}_s| \geq 1 + l_h|C^{\text{LP}}_s| > 1$$

where $|C^{\text{LP}}_{s,h}|$ denotes the fractional assignment of strictly greater than one point of color h to center s and $|C^{\text{LP}}_s|$ denotes the fractional assignment of strictly greater than two points to center s. However, both s_1 and s_2 cannot simultaneously be fractionally assigned strictly greater than 2 points each since there are only 4 points in total. Therefore, this shows that there exists no feasible solution to LP (12)-(15) for any arbitrarily large R. □

Figure 5: Toy example showing that for the set of centers S_{vll} from vanilla k-center algorithm, there does not exist a feasible solution to LP (12)-(15) for any arbitrarily large R.

C MaxFlow Rounding

MaxFlow rounding is given an LP solution satisfying (12),(13),(14), and (15). The MaxFlow rounding we use is identical to the one used in Dickerson et al. (2023); Ahmadian et al. (2019); Bercea et al. (2018). Our explanation here closely follows that in Dickerson et al. (2023) with some modifications for our setting. Formally, given an LP solution $\text{x}^{\text{LP}} = \{x^{\text{LP}}_{i,j}\}_{i\in\hat{S},j\in\mathcal{P}}$, the network flow diagram is constructed as follows:

1. $V = \{s, t\} \cup \mathcal{P} \cup \{i^h | i \in \hat{S}, h \in \mathcal{H}\} \cup \{i \in \hat{S}\}$.
2. $A = A_1 \cup A_2 \cup A_3 \cup A_4$ where $A_1 = \{(s, j) | j \in \mathcal{P}\}$ with upper bound of 1. $A_2 = \{(j, i^h) | x^{\text{LP}}_{i,j} > 0\}$ with upper bound of 1. The arc set $A_3 = \{(i^h, i) | i \in \hat{S}, h \in \mathcal{H}\}$ with lower bound $\left\lfloor \sum_{j\in\mathcal{P}_h} x^{\text{LP}}_{i,j} \right\rfloor$ and upper bound of $\left\lceil \sum_{j\in\mathcal{P}_h} x^{\text{LP}}_{i,j} \right\rceil$. As for $A_4 = \{(i, t) | i \in \hat{S}\}$ the lower and upper bounds are $\left\lfloor \sum_{j\in\mathcal{P}} x^{\text{LP}}_{i,j} \right\rfloor$ and $\left\lceil \sum_{j\in\mathcal{P}} x^{\text{LP}}_{i,j} \right\rceil$.

By construction of the network flow diagram the maximum flow that can be achieved at the sink t is n (the number of points). Further, the given LP assignment $\text{x}^{\text{LP}} = \{x^{\text{LP}}_{i,j}\}_{i\in\hat{S},j\in\mathcal{P}}$ is a valid fractional flow that achieves a maximum flow of n. Since the upper and lower bound on the arcs are integral, it follows by standard result of the max flow problem that we can find an integral maximum flow $\text{x}^{\text{Integ}} = \{x^{\text{Integ}}_{i,j}\}_{i\in\hat{S},j\in\mathcal{P}}$. Moreover, from the set upper and lower bounds the following is immediate:

$$\left\lfloor \sum_{j\in\mathcal{P}} \text{x}^{\text{LP}}_{i,j} \right\rfloor \leq \sum_{j\in\mathcal{P}} \text{x}^{\text{Integ}}_{i,j} \leq \left\lceil \sum_{j\in\mathcal{P}} \text{x}^{\text{LP}}_{i,j} \right\rceil$$

$$\left\lfloor \sum_{j\in\mathcal{P}_h} \text{x}^{\text{LP}}_{i,j} \right\rfloor \leq \sum_{j\in\mathcal{P}_h} \text{x}^{\text{Integ}}_{i,j} \leq \left\lceil \sum_{j\in\mathcal{P}_h} \text{x}^{\text{LP}}_{i,j} \right\rceil$$

Further, recall from Lemma 5.3 that $|C_i^{\text{LP}}| = \sum_{j\in\mathcal{P}} \mathrm{x}_{i,j}^{\text{LP}}$, $|C_i^{\text{Integ}}| = \sum_{j\in\mathcal{P}} \mathrm{x}_{i,j}^{\text{Integ}}$, $|C_{i,h}^{\text{LP}}| = \sum_{j\in\mathcal{P}_h} \mathrm{x}_{i,j}^{\text{LP}}$, and $|C_{i,h}^{\text{Integ}}| = \sum_{j\in\mathcal{P}_h} \mathrm{x}_{i,j}^{\text{Integ}}$. Therefore, it follows that we have

$$\begin{aligned}\left\lfloor |C_i^{\text{LP}}| \right\rfloor \leq & |C_i^{\text{Integ}}| \leq \left\lceil |C_i^{\text{LP}}| \right\rceil, \\ \left\lfloor |C_{i,h}^{\text{LP}}| \right\rfloor \leq & |C_{i,h}^{\text{Integ}}| \leq \left\lceil |C_{i,h}^{\text{LP}}| \right\rceil\end{aligned}$$

This proves part (2) of Lemma 5.3. Part (1) of Lemma 5.3 simply follows from that fact that if $\mathrm{x}_{i,j}^{\text{LP}} = 0$ then there would not be an arc connecting point (vertex) j to vertex $i^h, \forall h \in \mathcal{H}$ and that immediately implies that $\mathrm{x}_{i,j}^{\text{Integ}} = 0$.

D MORE DISCUSSION ABOUT PREVIOUS NOISE MODELS AND THEIR ALGORITHMS IN FAIR CLUSTERING

D.1 Drawbacks of the Noise Model of Probabilistic Fair Clustering Esmaeili et al. (2020)

We provide an example to show the drawbacks of the probabilistic model introduced in Esmaeili et al. (2020). Their theoretical guarantees for robust fair clustering only guarantee to satisfy the fairness constraints in expectation. However, the realizations can be arbitrarily unfair. We illustrate this using an example. Consider a set of 14 points as shown in the Figure 6 where each point is assigned to either the red or blue group, each with a probability of $1/2$. Any clustering of these points is fair in expectation assuming the lower and upper bounds for both the blue and red group are close to $\frac{1}{2}$. However individual realizations can be unfair. Specifically, since the joint probability distribution is not known (in fact, it is not incorporated at all in the probabilistic model of Esmaeili et al. (2020)) the realizations could be as shown in the figure where all points in a cluster take on the same color simultaneously in a realization. This shows that the probabilistic model may return clusters that can be fair (proportional) in expectation but completely unfair (unproportional) in realization.

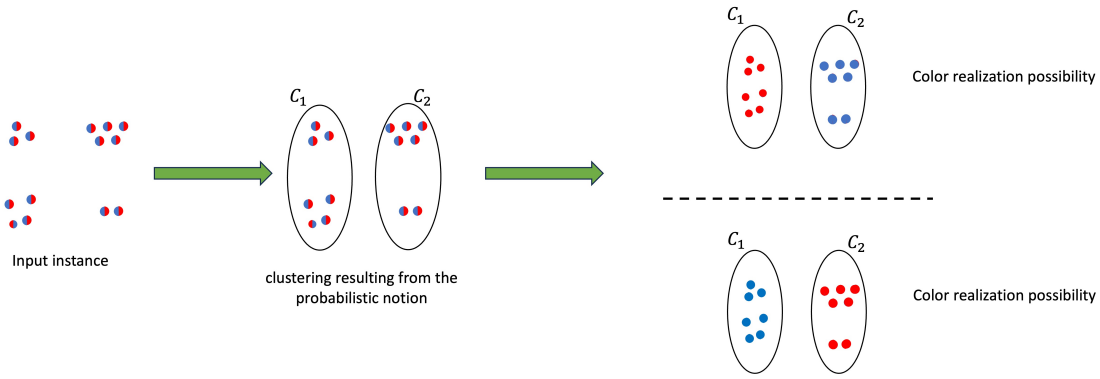

Figure 6: An instance of FAIR-k-CENTER with 14 points where each point has the probabilities $p_j^{\text{red}} = p_j^{\text{blue}} = \frac{1}{2}$. Here p_j^{red} and $p_j^{\text{blue}} \forall j \in \mathcal{P}$ denote the probability that point j belong to red and blue groups respectively. Note how the probabilistic fair clustering satisfies fairness in expectation. However, the two realizations shown demonstrate that color proportionality can be completely violated.

D.2 More Discussion About the Noise Model of Chhabra et al. (2023) and Their Algorithm

We provide an example to show the drawbacks of the noise model introduced by Chhabra et al. (2023). Their noise model assumes that only a subset of the points are affected by the adversary but they do not specify how one can access this subset. In their experiments, they generate this subset using random sampling. Specifically, they independently sample each point with probability 0.15 to obtain a subset of points $\mathcal{P}' \subseteq \mathcal{P}$. However, we can easily construct examples where their random sampling method with probability ≈ 0.99 can never return some subsets. We illustrate this using an example. Consider an instance of FAIR-k-CENTER as shown in Figure 7 where only a subset of points have incorrect memberships according to Chhabra et al. (2023). Their sampling process selects a subset (comprising 10% of the points) by randomly sampling each point independently with a probability of 0.1. As a result, out of 160 points 16 points are perturbed in expectation. However, this does not capture scenarios where all perturbations occur within a subset of points (as shown in the left side of Figure 7) as the probability of such an event is close to 0, precisely 0.0002. On the other hand, our model considers for all possible subsets with 16 points. As a result, we can model the scenarios where all the incorrect memberships occur in a single group or more generally all possible combinations across the two groups.

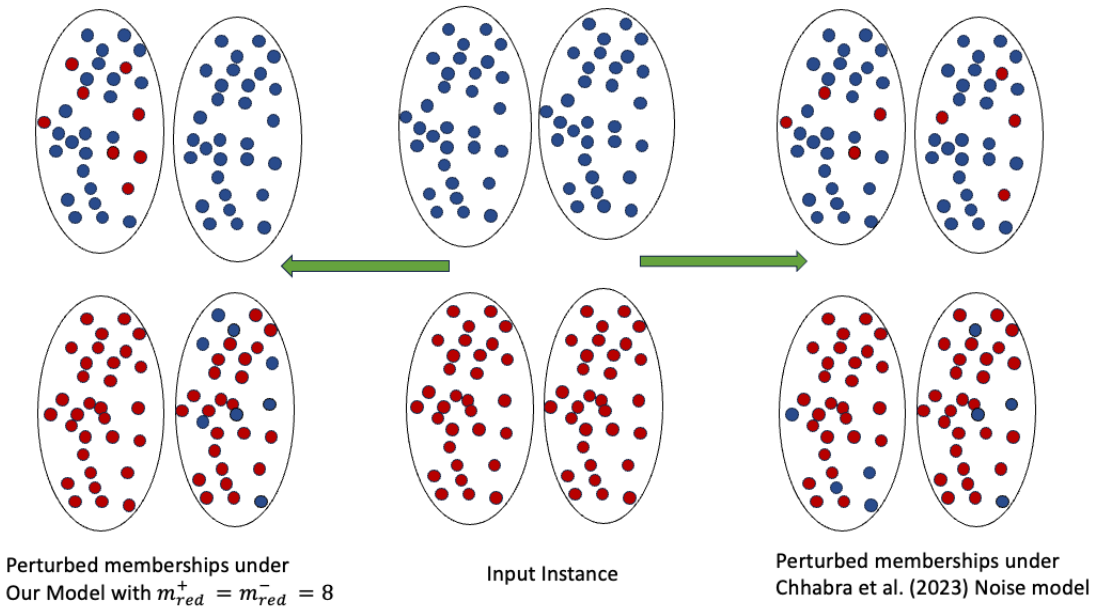

Figure 7: Consider an instance of FAIR-k-CENTER with $n = 160$ points, where 10% of the points have incorrect memberships. Chhabra et al. (2023) perturbs a subset (comprising 10% of the points) by randomly sampling each point independently with a probability of 0.1. As a result 16 points are perturbed in expectation. However, this does not capture scenarios where all perturbations occur within a subset of points as shown on the left side. This is because the probability of such an event is ≈ 0. On the other hand, our model considers for all possible subsets with 16 points, thereby covering all possible scenarios.

Furthermore, in terms of algorithms (Chhabra et al., 2023) also provide a defense algorithm by a Consensus Clustering method via k-means (Lloyd's algorithm) combined with fair constraints to achieve robustness against their proposed attack. Their algorithm trains a neural network using a loss function based on pair-wise similarity information obtained by running Consensus k-means clustering, which is different from our k-center objective. Moreover, their *fair clustering loss* fails to include the fractional proportional bounds present in our fair clustering instances. Therefore, their algorithm is inapplicable to our problem. Further, their algorithm provide no theoretical or empirical guarantees on the distance-based clustering cost. Finally, their fairness loss does not model the fairness constraints based on proportional lower and upper bounds or any other parameters provided by the stakeholder. More importantly, their robust algorithms have no theoretical guarantees on ex-post fairness violations, whereas ours do.

E ADDITIONAL DETAILS ON OUR EXPERIMENTAL SETUP & EXPERIMENTAL RESULTS

In Section E.1 we expand on the libraries and hardware used to complete experiments. In Section E.2 we discuss the running times of our algorithm ROBUSTALG and baselines presented in Figure 3 of Section 6. In Section E.3 we use a fourth dataset, **Diabetes**, to compare the algorithms' running times as the number of points n increases. Lastly, in Section E.4 we repeat the experiments in Figure 3 of Section 6 for a different range of m values.

E.1 Experimental Setup

The experiments are run on Python 3.6.15 on a commodity laptop with a Ryzen 7 5800U and 16GB of RAM. The linear programs (LP) in the algorithms are solved using CPLEX 12.8.0.0 (Nickel et al., 2022) and flow problems are solved using `NetworkX` 2.5.1 (Hagberg et al., 2013). In total, the experiments in Figure 3 solve 53 fair clustering instances (30 robust fair instances, 20 probabilistic fair, and 3 deterministic fair). PROBALG is not run on **Census1990** because its theoretical guarantees hold for the two-color setting only. DETALG has a single fair clustering instance per dataset because DETALG does not depend on m. Our code implementation forks the code of Dickerson et al. (2023).

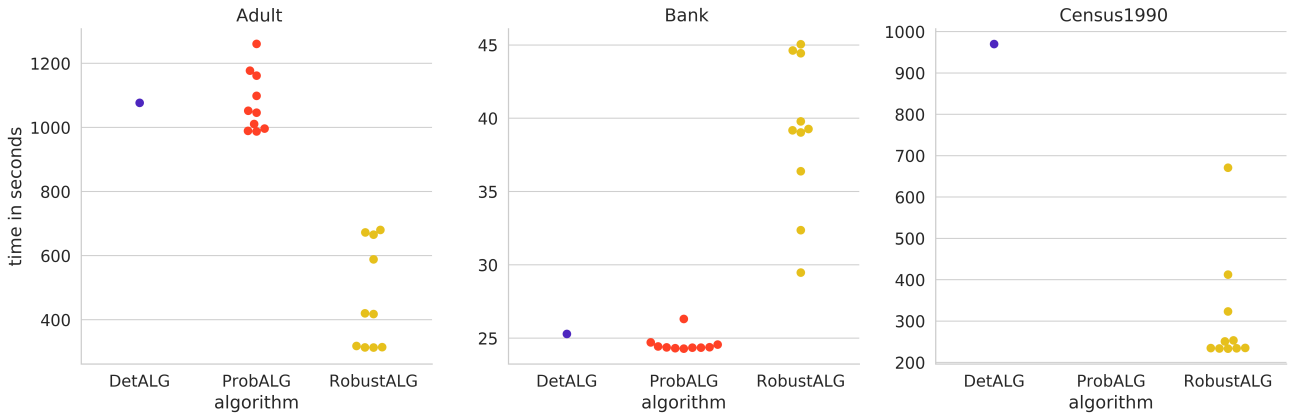

Figure 8: Wall clock time each algorithm took to solve the associated {deterministic, probabilistic, robust} fair clustering instances in Figure 3.

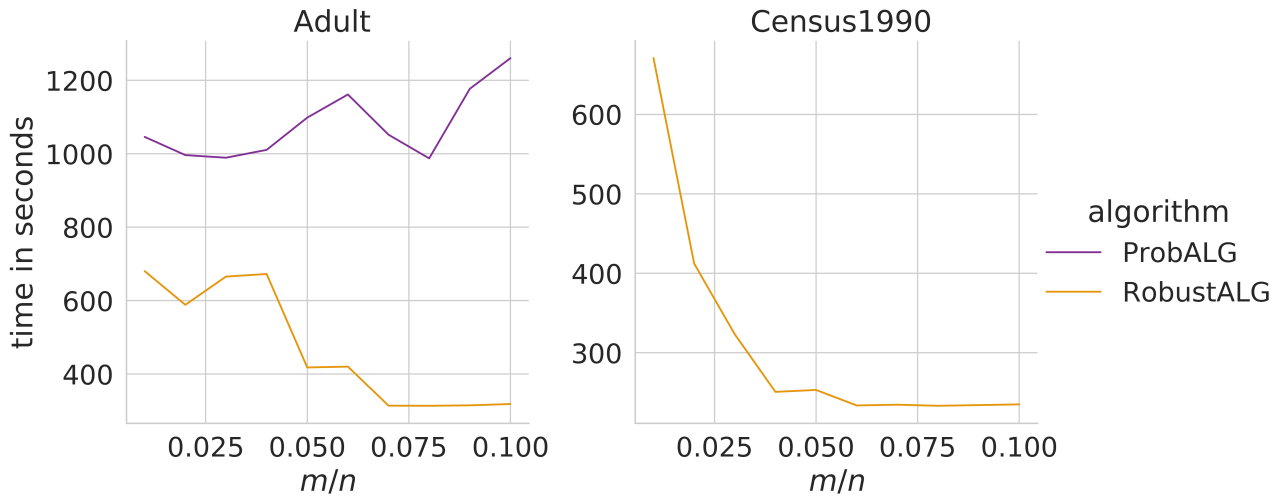

Figure 9: Running times of PROBALG and ROBUSTALG from Figure 8 but plotted against m.

E.2 Running Times of the Experiments

Figure 8 shows ROBUSTALG outperforms the baselines in both of the larger datasets (**Adult** and **Census1990**); this is not the case in **Bank**, but the difference is at most 20 seconds there. In each step of the binary search, ROBUSTALG has an additional step of GETCENTERS which contributes to the running time. However, the running time of these algorithms is largely dominated by solving the associated LPs. However, the LPs for the baselines always use k centers and thus have nk variables. However, ROBUSTALG uses the $k' \leq k$ centers selected by GETCENTERS. In practice, this can be a substantial speedup as the difference between k' and k increases and as n increases. This behavior is also observed in E.3.

Related to this phenomenon, as m grows, the binary search in ROBUSTALG discards smaller values of R because more LPs become infeasible. Therefore, the binary search moves onto greater values of R, for which GETCENTERS selects fewer centers. Indeed, in Figure 9 we can observe ROBUSTALG speeding up as m increases. On the other hand, and as expected, the running time of PROBALG does not exhibit this interaction with m.

E.3 Additional Experiments on the Diabetes dataset

We use an additional dataset **Diabetes** to supplement our running time experiments. Like our other three datasets, **Diabetes** is also from the UCI repository. It has 49 features and we set up two colors corresponding to whether the patient (i.e., point) was or was not female. Figure 10 shows the dependence of all three algorithms on k, and we can also see that the baselines are more sensitive to larger k. These plots use the first n' points of **Diabetes** for n' in $\{2000, 4000, 6000, \ldots, 23000\}$. For

each instance we take $m = 0.005n'$. The proportionality constants u_h and l_h are set to be feasible for RobustAlg exactly in the same way as in Section 6.

We also recreate the fairness violation and objective plots from Figure 3 on **Diabetes** as shown in Figure 11. For these experiments we fix $k = 5$ and $\forall h$, $l_h = 0.3$ and $u_h = 0.75$. A moment's reflection shows the worst fairness violation possible is 0.3. Indeed, DetAlg and ProbAlg, respectively, exactly reach and get very close to this violation. The objectives also only differ by at most a meager 4 units of distance.

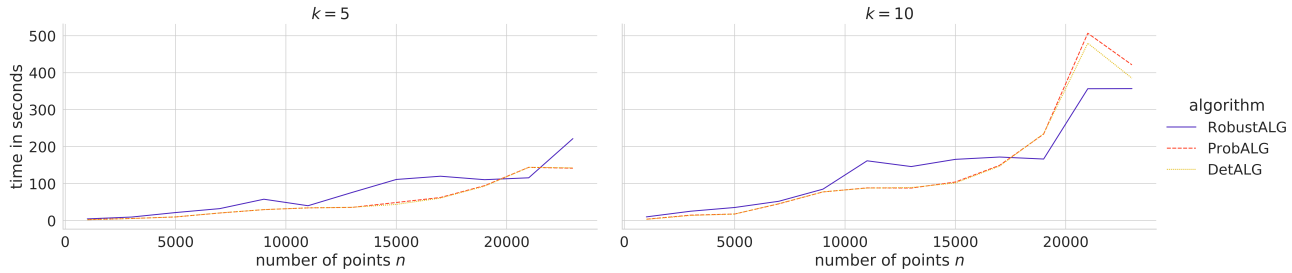

Figure 10: Running times of RobustAlg, ProbAlg, and DetAlg on dataset **Diabetes** as the number of points n increases. We run the experiments for number of clusters k of 5 and 10.

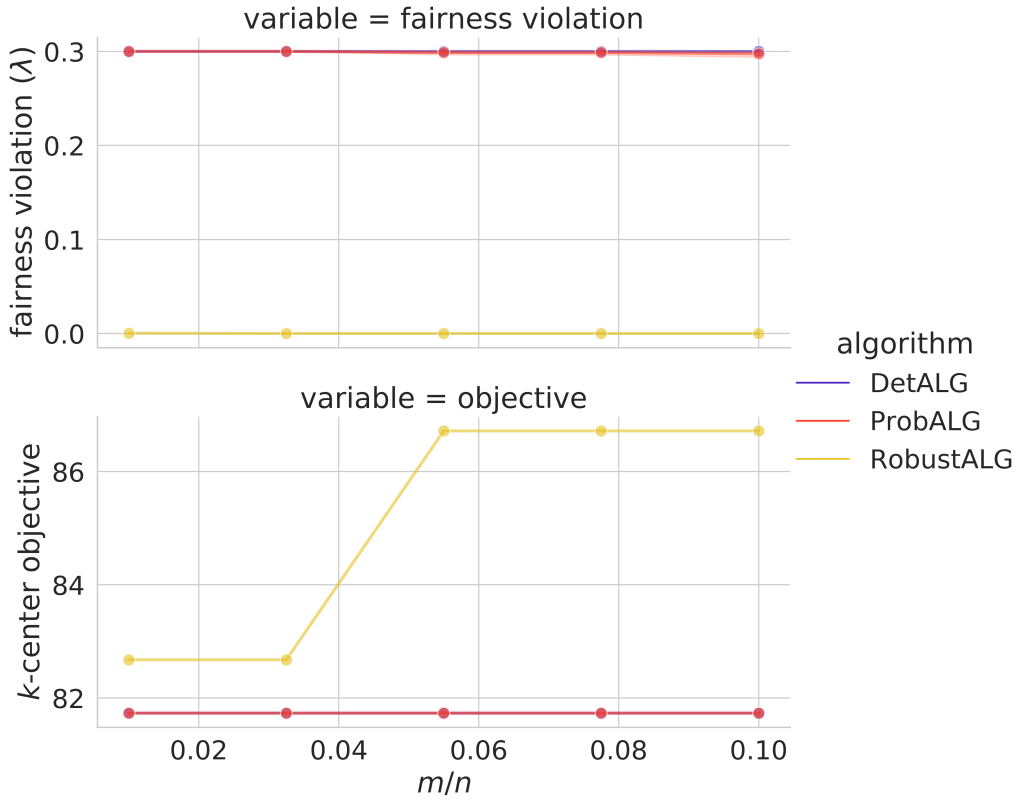

Figure 11: Plots of k-center objective and fairness violations on **Diabetes**.

E.4 Additional experiments on smaller values of m

Figure 12 recreates the experiments in Figure 3 but with values of m one order of magnitude smaller and $k = 5$. These plots show greater variance in the fairness violations of ProbAlg, which makes sense given the smaller m and that, in probabilistic fair clustering, the probability a point's label is correct is $p_{\text{acc}} = 1 - m/n$.

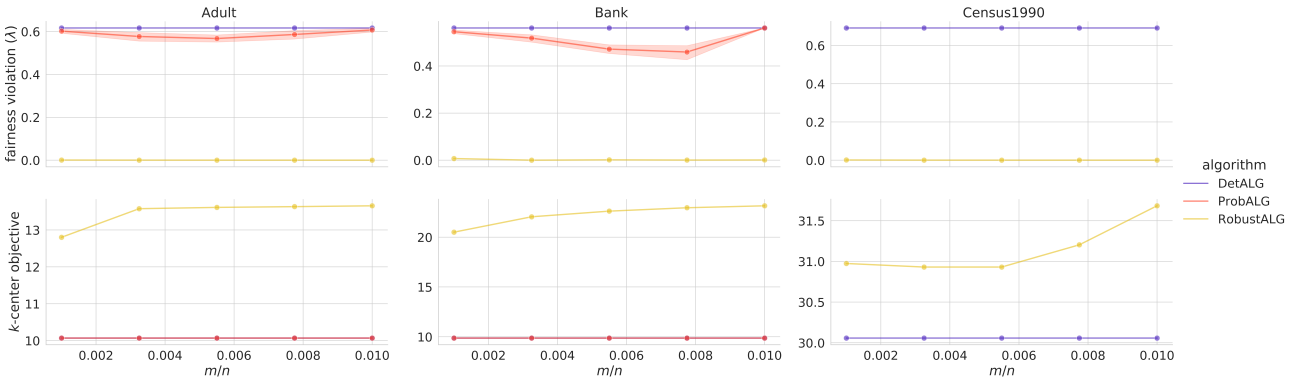

Figure 12: A repeat of Figure 3 but for smaller values of m and $k = 5$ instead of $k = 10$.