A Stochastic Alternating Balance k-Means Algorithm for Fair Clustering

S. Liu
Dept. of Industrial and Systems Engineering
Lehigh University
Bethlehem, PA 18015
sul217@lehigh.edu

L. N. Vicente
Dept. of Industrial and Systems Engineering
Lehigh University
Bethlehem, PA 18015
lnv@lehigh.edu

Abstract

In the application of data clustering to human-centric decision-making systems, such as loan applications and advertisement recommendations, the clustering outcome might discriminate against people across different demographic groups, leading to unfairness. A natural conflict occurs between the cost of clustering (in terms of distance to cluster centers) and the balance representation of all demographic groups across the clusters, leading to a bi-objective optimization problem that is nonconvex and nonsmooth. To determine the complete trade-off between these two competing goals, we design a novel stochastic alternating balance fair k-means (SAfairKM) algorithm, which consists of alternating classical mini-batch k-means updates and group swap updates. The number of k-means updates and the number of swap updates essentially parameterize the weight put on optimizing each objective function. Our numerical experiments show that the proposed SAfairKM algorithm is robust and computationally efficient in constructing well-spread and high-quality Pareto fronts both on synthetic and real datasets. Moreover, we propose a novel companion algorithm, the stochastic alternating bi-objective gradient descent (SA2GD) algorithm, which can handle a smooth version of the considered bi-objective fair k-means problem, more amenable for analysis. A sublinear convergence rate of $\mathcal{O}(1/T)$ is established under strong convexity for the determination of a stationary point of a weighted sum of the two functions parameterized by the number of steps or updates on each function.

1 Introduction

Clustering is a fundamental task in data mining and unsupervised machine learning with the goal of partitioning data points into clusters, in such a way that data points in one cluster are very similar and data points in different clusters are quite distinct [16]. It has become a core technique in a huge amount of application fields such as feature engineering, information retrieval, image segmentation, targeted marketing, recommendation systems, and urban planning. Data clustering problems take on many different forms, including partitioning clustering like k-means and k-median, hierarchical clustering, spectral clustering, among many others [7, 16]. Given the increasing impact of automated decision-making systems in our society, there is a growing concern about algorithmic unfairness, which in the case of clustering may result in discrimination against minority groups. For instance, female may receive proportionally fewer job recommendations with high salary [13] due to their under-representation in the cluster of high salary recommendations. Such demographic features like gender and race are called *sensitive* or *protected* features, which we wish to be fair with respect to.

Related work An extensive literature work studying algorithmic fairness has been focused on developing universal fairness definitions and designing fair algorithms for supervised machine

Preprint. Under review.

learning problems. Among the broadest representative fairness notions proposed for classification and regression tasks are *disparate impact* [5] (also called *demographic parity* [10]), *equalized odds* [19], and individual fairness [15], based on which the fairness notions in clustering were proposed accordingly. There are a number of classes of fairness definitions proposed and investigated for the clustering task [1, 11, 12, 18, 23, 27]. The most widely used fairness notion is called *balance*. It was proposed by [12], and it has been extended in several subsequent works [6, 20, 30]. As a counterpart of the disparate impact concept in fair supervised machine learning, balance essentially aims at ensuring that the representation of protected groups in each cluster preserves the global proportion of each protected group.

Depending on the stage of clustering in which the fairness requirements are imposed, the prior works on fair clustering are categorized into three families, namely pre-processing, in-processing, and post-processing. A large body of the literature work [4, 12, 20, 30] falls into the pre-processing category. The whole dataset is first decomposed into small subsets named *fairlets*, where the desired balance can be guaranteed. Any resulting solution from classical clustering algorithms using the set of fairlets will then be fair. Chierichetti et al. [12] focused on the case of two demographic groups and formulated explicit combinatorial problems (such as perfect matching and minimum cost flow problems) to decompose the dataset into minimal fair sets defining the fairlets. Their theoretical analysis gave strong guarantees on the quality of the fair clustering solutions for k-center and k-median problems. Following that line of work, Backurs et al. [4] embedded the whole dataset into a hierarchical structure tree and improved the time complexity of the fairlet decomposition step from quadratic to nearly linear time (in the dataset size). Schmidt et al. [30] introduced the notion of fair *coresets* and proposed an efficient streaming fair clustering algorithm for k-means. They introduced a near-linear time algorithm to construct coresets that helps reduce the input data size and hence speeds up any fair clustering algorithm. Huang et al. [20] further boosted the efficiency of coresets construction and made a generalization to multiple non-disjoint demographic groups for both k-means and k-median.

On the contrary, post-processing clustering methods [3, 6, 22, 29] modify the resulting clusters from classical clustering algorithms to improve fairness. For example, Bera et al. [6] proposed a fair re-assignment problem as a linear relaxation of an integer programming model given the clustering results from any vanilla k-means, k-median, or k-center algorithms. They showed how to derive a $(\rho+2)$-approximation fair clustering algorithm from any ρ-approximation vanilla clustering algorithm within a theoretical bound of fairness constraints violation. Moreover, their framework works for datasets with multiple and potentially overlapping demographic groups. Lastly, in-processing methods incorporate the fairness constraints into the clustering process [2, 24, 32]. Our approach falls into this category and allows for the determination of the trade-offs between clustering costs and fairness. To our knowledge, the only such other in-processing approach is the one of Ziko et al. [32], where the clustering balance is approximately measured by the KL-divergence and imposed as a penalty term in the fair clustering objective function. The penalty coefficient is then used to control the trade-offs between clustering cost/fairness.

Our contribution The partitioning clustering model, also referred to as the center-based clustering model, consists of selecting a certain number K of centers and assigning data points to their closest centers. In this paper, we will focus on the well-known k-means model, and we will introduce a novel fair clustering algorithm using the balance measure. The main challenge of the fair clustering task comes from the violation of the assignment routine, which then indicates that a data point is no longer necessarily assigned to its closest cluster. The higher the balance level one wants to achieve, the more clustering cost is added to the final clustering. Hence, there exists a natural conflict between the fairness level, when measured in terms of balance, and the classical k-means clustering objective.

We explicitly formulate the trade-offs between the k-means clustering cost and the fairness as a bi-objective optimization problem, where both objectives are written as nonconvex and nonsmooth functions of binary assignment variables defining point assignments in the clustering model (see (2) further below). Our goal is to construct an informative approximation of the Pareto front for the proposed bi-objective fair k-means clustering problem, without exploring exhaustively the binary nature of the assignment variables. The most widely used method in solving general bi-objective optimization problems is the so-called weighted-sum method [17]. There, one considers a set of single objective problems, formed by convex linear combinations of the two functions, and (a portion of) the Pareto front might be approximated by solving the corresponding weighted-sum problems.

However, this methodology has no rigorous guarantees due to the nonconvexity of both objective functions. Also, the non-smoothness of the fairness objective poses an additional difficulty to the weighted-sum method, as one function will be smooth and the other one no. Moreover, even ignoring the nonconvexity and non-smoothness issues, the two objectives, namely the clustering cost and the clustering balance, can have significantly different magnitudes. One can hardly preselect a good set of weights corresponding to decision-makers' preferences to capture a well-spread Pareto front.

Therefore, we were motivated to design a novel stochastic alternating balance fair k-means (SAfairKM) algorithm, inspired from the classical mini-batch k-means algorithm, which essentially consists of alternatively taking pure mini-batch k-means updates and swap-based balance improvement updates. In fact, the number of k-means updates (denoted by n_a) and the number of swap updates (denoted by n_b) play a role similar to the weights in the weighted-sum method, parameterizing the efforts of optimizing each objective. In the pure mini-batch k-means updates, we focus on minimizing the clustering cost. A mini-batch of points is randomly drawn and assigned to their closest clusters, after which the set of centers are updated using mini-batch stochastic gradient descent. In the swap-based balance improvement steps, we aim at increasing the overall clustering balance. For this purpose, we propose a simple swap routine that is guaranteed to increase the overall clustering balance by swapping data points between the minimum balance cluster and a target well-balanced cluster. Similarly to the k-means updates, the set of centers are updated using the batch of data points selected to swap. While the k-means updates reproduce the stochastic gradient descent directions for the clustering cost function, the swap updates can be seen as taking steps along some increasing directions for the clustering balance objective (not necessarily the best ascent direction).

To provide a better understanding of the proposed algorithm, we develop a novel convergence theory for a companion algorithm named the stochastic alternating bi-objective gradient descent (SA2GD) algorithm. Such an algorithm is designed to handle a smooth version of the type of bi-objective optimization problems we considered in the fair clustering context. Besides, one can apply it to any smooth bi-objective optimization problems when knowing exactly how to optimize each single objective separately. It is shown that the SA2GD algorithm has a sublinear convergence rate of $\mathcal{O}(1/T)$ when determining a stationary point of some weighted-sum function of the two objectives, under strong convexity and classical assumptions of stochastic gradients. The derived convergence theory gives us insight into the numerical behavior of the SAfairKM algorithm, in particular in what regards the determination of a Pareto front by parameterizing the optimization effort put into optimizing each function at every iteration of the algorithm.

We have evaluated the performance of the proposed SAfairKM algorithm using both synthetic datasets and real datasets. To endow SAfairKM with the capability of constructing a Pareto front in a single run, we use a list of nondominated points updated at every iteration. The list is randomly generated at the beginning of the process. At every iteration, and for every point in the current list, we apply SAfairKM for all considered pairs of (n_a, n_b). For each pair (n_a, n_b), one does n_a k-means updates and n_b swap updates. At the end of each iteration, we remove from the list all dominated points (those for each there exists another one with higher clustering cost and lower clustering balance). Such a simple mechanism is also beneficial for excluding bad local optima, considering that the two objectives are nonconvex. We will present the full trade-offs between the two conflicting objectives for four synthetic datasets and two real datasets. A numerical comparison with the fair k-means algorithm proposed in [32] further confirms the robustness and efficiency of the proposed algorithm in constructing informative and high-quality trade-offs.

2 The mini-batch k-means algorithm

In the classical k-means problem, one aims to choose K centers (representatives) and to assign a set of points to their closest centers. The k-means objective is the sum of the minimum (squared Euclidean) distance of all points to their corresponding centers. Given a set of N points $P = \{x_p\}_{p=1}^N$, where x_p is the non-sensitive feature vector, the goal of clustering is to assign N points to K clusters identified by K centroids $C = [c_1, \ldots, c_K]^\top$. Let $[K]$ denote the set of positive integers up to K. The k-means clustering problem is formulated as the minimization of a nonsmooth function of the set of centroids:

$$\min \; f_1^{KM}(C) \;=\; \frac{1}{2}\sum_{p=1}^{N} \min_{k\in[K]} \|x_p - c_k\|^2. \tag{1}$$

Since each data point is assigned to the closest cluster, the K cluster centroids are implicitly dependent on the point assignments. Let $s_{p,k} \in \{0,1\}$ be an assignment variable who takes the value 1 if point x_p is assigned to cluster k, and 0 otherwise. For simplicity, we denote s_k, $k \in [K]$, as an N-dimensional assignment vector for cluster k, and $s_p, p = 1, \ldots, N$, as a K-dimensional assignment vector for point x_p. Let $X \in \mathbb{R}^{N\times d}$ be the data matrix stacking N data points of dimension d and $e_N \in \mathbb{R}^N$ be an all-ones vector. Then one can compute each centroid using $c_k = X^\top s_k / e_N^\top s_k$.

In practice, Lloyd's heuristic algorithm [26], also known as the standard batch k-means algorithm, is the simplest and most popular k-means clustering algorithm, and converges to a local minimum but without worst-case guarantees [21, 31]. The main idea of Lloyd's heuristic is to keep updating the K cluster centroids and assigning the full batch of points to their closest centroids.

In the standard batch k-means algorithm, one can compute the full gradient of the objective function (1) with respect to k-th center by $\nabla_{c_k} f_1^{KM}(C) = \sum_{x_p \in \mathcal{C}_k} (c_k - x_p)$, where $\mathcal{C}_k, k = [K]$, is the set of points assigned to cluster k. Whenever there exists a tie, namely a point that has the same distance to more than one cluster, one can randomly assign the point to any of such clusters. A full batch gradient descent algorithm would iteratively update the centroids by $c_k^{t+1} - c_k^t = \alpha_k^t \sum_{x_p \in \mathcal{C}_k} (x_p - c_k), \forall k \in [K]$, where $\alpha_k^t > 0$ is the step size. Let N_k^t be the number of points in cluster k at iteration t. It is known that the full batch k-means algorithm with $\alpha_k^t = 1/N_k^t$ converges to a local minimum as fast as Newton's method, with a superlinear rate [8].

The standard batch k-means algorithm is proved to be slow for large datasets. Bottou and Bengio [8] proposed an online stochastic gradient descent (SGD) variant that takes a gradient descent step using one sample at a time. Given a new data point x_p to be assigned, a stochastic gradient descent step would look like $c_k^{t+1} = c_k^t + \alpha_k^t (x_p - c_k^t)$ if x_p is assigned to cluster k. While the SGD variant is computationally cheap for large datasets, it finds solutions of lower quality than the batch algorithm due to the stochasticity. The mini-batch version of the k-means algorithm uses a mini-batch sampling to lower stochastic noise and, in the meanwhile, speed up the convergence. The detailed mini-batch k-means is given in Algorithm 1 of Appendix A.

3 A new stochastic alternating balance fair k-means method

3.1 The bi-objective balance k-means formulation

Balance [12] is the most widely used fairness measure in the literature of fair clustering. Consider J disjoint demographic groups. Let $[J]$ denote the set of positive integers up to J. Let V_j represent the set of points in demographic group $j \in [J]$. Then, $v_{p,j}$ takes the value 1 if point $x_p \in V_j$. We denote v_j as an N-dimensional indicator vector for the demographic group $j \in [J]$. The balance of cluster k is formally defined as $b_k = \min_{j\neq j'} v_j^\top s_k / v_{j'}^\top s_k \leq 1, \forall k \in [K]$, which calculates the minimum ratio among different pairs of protected groups. The overall clustering balance is the minimum balance over all clusters, i.e., $b = \min_{k=1}^K b_k$. The higher the overall balance, the fairer the clustering.

By the definition of cluster balance given above, the balance function can be easily computed only using the assignment variables. The k-means objective (1) can be rewritten as a function of the assignment variables as well. Hence, one can directly formulate the inherent trade-off between clustering cost and balance as a bi-objective optimization problem, i.e.,

$$\min \quad (f_1(S), -f_2(S)) \quad \text{s.t.} \quad s_p \in \Delta_K, \forall p = 1, \ldots, N, \quad S \in \{0,1\}^{N\times K}, \tag{2}$$

where

$$f_1(S) \;=\; \frac{1}{N} \sum_{k=1}^{K} \sum_{p=1}^{N} s_{p,k} \|x_p - c_k\|^2, \quad \text{with } c_k \;=\; \frac{X^\top s_k}{e_N^\top s_k} = \frac{\sum_{p=1}^N x_p s_{p,k}}{\sum_{p=1}^N s_{p,k}},$$

$$f_2(S) \;=\; \min_{k\in[K]} \min_{\substack{j\neq j' \\ j,j'\in[J]}} \frac{v_j^\top s_k}{v_{j'}^\top s_k},$$

and $\Delta_K \subset \mathbb{R}^K$ denotes a simplex set. The two constraints in (2) ensure that one point can only be assigned to one cluster. Note that both objectives are nonconvex functions of the binary assignment variables.

3.2 The stochastic alternating balance fair k-means method

We propose a novel stochastic alternating balance fair k-means clustering algorithm to compute a nondominated solution on the Pareto front. We will use a simple but effective alternating update mechanism, which consists of improving *either* the clustering objective *or* the overall balance, by iteratively updating cluster centers and assignment variables. Specifically, every iteration of the proposed algorithm contains two sets of updates, namely pure k-means updates and pure swap-based balance improvement steps. The pure k-means updates were introduced in Section 2, and will consist of taking a certain number of stochastic k-means steps. In the balance improvement steps, a certain batch of points is selected and swapped between the minimum balanced cluster and a target well-balanced cluster.

Balance improvement steps At the current iteration, let $\mathcal{C}_l$ be the cluster with the minimum balance. Then $\mathcal{C}_l$ is the bottleneck cluster that defines the overall clustering balance. Without loss of generality, we assume that $b_l = |\mathcal{C}_l \cap V_1|/|\mathcal{C}_l \cap V_2|$, which then implies that the pair of demographic groups (V_1, V_2) forms a key to improve the balance of cluster $\mathcal{C}_l$, as well as the overall clustering balance. In terms of the assignment variables, we have

$$b \;=\; b_l \;=\; \frac{v_1^\top s_l}{v_2^\top s_l} \;=\; \frac{\sum_{p=1}^N v_{p,1} s_{p,l}}{\sum_{p=1}^N v_{p,2} s_{p,l}}. \tag{3}$$

One way to determine a target well-balanced cluster $\mathcal{C}_h$ is to select it as the one with the maximum ratio between V_1 and V_2, i.e.,

$$h \;\in\; \operatorname*{argmax}_{k\in[K]} \; \left\{ v_1^\top s_k / v_2^\top s_k, v_2^\top s_k / v_1^\top s_k \right\}. \tag{4}$$

Another way to determine such a target cluster is to select a cluster $\mathcal{C}_h$ that is closest to $\mathcal{C}_l$, i.e.,

$$h \;\in\; \operatorname*{argmin}_{k\in[K], k\neq l} \; \|c_k - c_l\|. \tag{5}$$

We call the target cluster computed by (4) a *global* target and the one selected by (5) a *local* target. Using a global target cluster makes the swap updates more efficient and stable in the sense that the target cluster is only changed when the minimum balanced cluster changes. Instead, swapping according to the local target leads to less increase in clustering costs.

To improve the overall balance, one swaps a point in cluster $\mathcal{C}_l$ belonging to V_2 with a point in cluster $\mathcal{C}_h$ belonging to V_1. Each of these swap updates will guarantee an increase in the overall balance. The detailed stochastic alternating balance fair k-means clustering algorithm is given in Algorithm 2. At each iteration, we alternate between taking k-means updates using a drawn batch of points (denote the batch size by n_a) and "swap" updates using another drawn batch of points (denote the batch size by n_b). The generation of the two batches is independent. The choice of n_a and n_b influences the nondominated point obtained at the end, in terms of the weight put into each objective.

Instead of randomly selecting points to swap in line 11 of Algorithm 2, in our experiments we have used a more accurate swap strategy by increasing the batch size. Basically, we randomly sample a batch of points from $\mathcal{C}_l \cap V_2$ (resp. $\mathcal{C}_h \cap V_1$) and select x_p (resp. x'_p) as the one closest to $\mathcal{C}_h$ (resp. $\mathcal{C}_l$). The batch size could be increased as the algorithm proceeds. Our numerical experiments show that the combination of local target clusters and the increasingly accurate swap strategy result in better numerical performance.

One could have converted the bi-objective optimization problem (2) into a weighted-sum function using the weights associated with the decision-maker's preference. However, optimizing such a weighted-sum function hardly reflects the desired trade-off due to significantly different magnitudes of the two objectives. Moreover, the existing k-means algorithm frameworks, including Lloyd's heuristic algorithm, are not capable of directly handling the weighted-sum objective function. In our proposed SAfairKM algorithm, the pair (n_a, n_b) plays a role similar to the weights in the weighted-sum method. Later, we will show in Section 5 that a companion algorithm for solving a smooth version of the bi-objective optimization problem generates a sequence converging to a stationary point of the weighted-sum function composed by the weights defined by (n_a, n_b).

Algorithm 2 Stochastic alternating balance fair k-means clustering (SAfairKM) algorithm

Input: The set of points P, an integer K, and parameters n_a, n_b.
Output: The set of clustering labels $\Delta = \{\delta_1, \ldots, \delta_N\}$ for all points, where $\delta_p \in [K]$.
Randomly generate an initial label $\{\delta_1, \ldots, \delta_N\}$, and compute k-means centers $\{c_1, \ldots, c_K\}$ and balances $\{b_1, \ldots, b_K\}$ for all clusters.
for $t = 1, 2, \ldots$ **do**
  Randomly sample a batch of n_a points $B_t \subseteq P$.
  for $x_p \in B_t$ **do**
    Identify its closest center index i_p. Update clustering label $\delta_p = i_p$.
    Update counter $N_{\delta_p} = N_{\delta_p} + 1$ and center $c_{\delta_p} = c_{\delta_p} + \frac{1}{N_{\delta_p}}(x_p - c_{\delta_p})$.
  for $r = 1, 2, \ldots, n_b$ **do**
    Identify $\mathcal{C}_l$, $\mathcal{C}_h$, and the pair of demographic groups (V_1, V_2) according to (3) and (5).
    Randomly select points $x_p \in \mathcal{C}_l \cap V_2$ and $x_{p'} \in \mathcal{C}_h \cap V_1$.
    Swap points: set $\delta_p = h$ and $\delta_{p'} = l$.
    Update centers $c_l = c_l + \frac{1}{N_l}(x_{p'} - c_l)$ and $c_h = c_h + \frac{1}{N_h}(x_p - c_h)$.
    Update balance for clusters $\mathcal{C}_l$ and $\mathcal{C}_h$.

4 Numerical experiments

4.1 Pareto front SAfairKM algorithm

In our implementation[1], to obtain a well-spread Pareto front, we frame the SAfairKM algorithm into a Pareto front version using a list updating mechanism. See Algorithm 3 in Appendix B for a detailed description. In the initialization phase, we specify a sequence of pairs of the number of k-means updates and swap updates $\mathcal{W} = \{(n_a, n_b) : n_a + n_b = n_{\text{total}}, n_a, n_b \in \mathbb{N}_0\}$, and we generate a list of random initial clustering labels $\mathcal{L}_0$. Then we run Algorithm 2 for a certain number of iterations ($q = 1$ in our experiments) parallelly for each label in the current list $\mathcal{L}_t$, resulting in a new list of clustering labels $\mathcal{L}_{t+1}$. At the end of each iteration, the list is cleaned up by removing all the dominated points from $\mathcal{L}_{t+1}$. Using this algorithm, the list of nondominated points is refined towards the true real Pareto front. The process can be terminated when either the number of nondominated points is greater than a certain budget (1500 in our experiments) or when the total number of iterations exceeds a certain limit (depending on the size of the dataset).

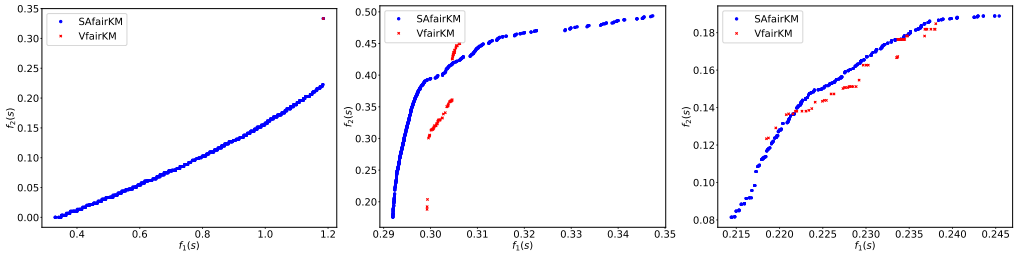
To the best of our knowledge, the only approach in the literature providing a mechanism of controlling trade-offs between the two conflicting objectives was suggested by [32] and briefly described in Appendix C. Their approach (here called VfairKM) consists of solving (7) for different penalty coefficients μ, resulting in a set of solutions from which we then remove dominated solutions to obtain an approximated Pareto front. To ensure a fair comparison, we select a set of penalty coefficients evenly from 0 to an upper bound $\mu_{\max}$, which is determined by pre-experiments such that the corresponding fairness error is less than 0.01 or no longer possibly decreased when further increasing its value. In some cases, we found that VfairKM is not able to produce a fairer clustering outcome when the penalty coefficient is greater than $\mu_{\max}$ due to numerical instability.

4.2 Numerical results

Trade-offs for synthetic datasets We randomly generated four synthetic datasets from Gaussian distributions, and their demographic compositions are given in Figure 2 of Appendix D.1. Each synthetic dataset has 400 data points in the $\mathbb{R}^2$ space and two demographic groups ($J = 2$) marked by black/circle and purple/triangle.

Using the list update mechanism (described by Algorithm 3 in Appendix B), we are able to obtain a well-spread Pareto front with comparable quality for each of the synthetic datasets. Recall that we are minimizing the clustering cost and maximizing the clustering balance. The closer the Pareto front is

[1] Our implementation code is available at `https://github.com/sul217/SAfairKM`. All the experiments were conducted on a MacBook Pro Intel Core i5 processor.



(a) Syn_unequal_ds2 ($K = 2$). (b) Adult ($K = 10$). (c) Bank ($K = 10$).

Figure 1: Pareto fronts: SAfairKM: 400 iterations for Syn_unequal_ds2, 2500 iterations for Adult, and 8000 iterations for Bank, 30 starting labels, and 4 pairs of (n_a, n_b); VfairKM: $\mu_{\max} = 0$ for Syn_unequal_ds2, $\mu_{\max} = 3260$ for Adult, and $\mu_{\max} = 2440$ for Bank.

to the upper left corner, the higher its quality. In particular, Figure 1 (a) gives the approximated Pareto front for the Syn_unequal_ds2 dataset with $K = 2$, which confirms the natural conflict between the clustering cost and the clustering balance. One can see that the VfairKM algorithm is not able to output any trade-off information as it always finds the fairest solution regardless of the value of μ. Due to the special composition of this dataset, the Pareto front generated by SAfairKM is disconnected (the point at the right upper corner is both VfairKM and SAfairKM). More results are given in Appendix D.1.

Trade-offs for real datasets Two real datasets *Adult* [25] and *Bank* [28] are taken from the UCI machine learning repository [14]. The *Adult* dataset contains $32,561$ samples. Each instance is characterized by 12 nonsensitive features (including age, education, hours-per-week, capital-gain, and capital-loss, etc.). For the clustering purpose, only five numerical features among the 12 features are kept. The demographic proportion of the *Adult* dataset is $[0.67, 0.33]$ in terms of gender ($J = 2$), which corresponds to a dataset balance of 0.49. The *Bank* dataset contains $41,108$ data samples. Six nonsensitive numerical features (age, duration, number of contacts performed, consumer price index, number of employees, and daily indicator) are selected for the clustering task. Its demographic composition in terms of marital status ($J = 3$) is $[0.11, 0.28, 0.61]$, and hence the best clustering balance one can achieve is 0.185.

For the purpose of a faster comparison, we randomly select a subsample of size 5000 from the original datasets and set the number of clusters to $K = 10$. The resulting Pareto fronts from the two algorithms are given in Figure 1 (b)-(c). For both datasets, SAfairKM is able to produce more spread-out Pareto fronts which capture a larger range of balance, and hence provide more complete trade-offs between the two conflicting goals. In terms of Pareto front quality (meaning dominance of one over the other), SAfairKM also performs better than VfairKM. In fact, we can see from Figure 1 that the Pareto fronts generated by SAfairKM dominate most of the solutions given by VfairKM. The Pareto fronts corresponding to $K = 5$ are also given in Figure 8 of Appendix D.2. SAfairKM results in a Pareto front of higher spread and slightly lower quality than VfairKM for the Adult dataset, while the Pareto front output from SAfairKM has better spread and higher quality for the Bank dataset.

Performance in terms of spread and quality of Pareto fronts SAfairKM is able to generate more spread-out and higher-quality Pareto fronts regardless of the data distribution (see the trade-off results for the four synthetic datasets). The robustness partially comes from the list update mechanism which establishes a connection among parallel runs starting from different initial points and pairs (n_a, n_b), and thus helps escape from bad local optima.

Performance in terms of computational time Since the two algorithms (SAfairKM and VfairKM) generally produce Pareto fronts of different cardinalities, we evaluate their computational efforts by the average CPU time spent per computed nondominated solution (see Table 1). Our algorithm was shown to be clearly more computationally efficient than VfairKM.

Table 1: Average CPU times per nondominated solution.

Dataset	SAfairKM	VfairKM	Dataset	SAfairKM	VfairKM
Syn_equal_ds1	0.80	1.06	Adult ($K = 10$)	18.52	40.43
Syn_unequal_ds1	0.81	0.98	Bank ($K = 10$)	59.12	76.29
Syn_equal_ds2	0.70	1.29	Adult ($K = 5$)	14.88	15.08
Syn_unequal_ds2	0.80	0.10	Bank ($K = 5$)	11.97	50.31

5 Alternating gradient descent for bi-objective optimization

A continuous relaxation of the bi-objective optimization problem (2), setting $S \in [0,1]^{N\times K}$, is still a challenging problem from the viewpoint of designing rigorous solution algorithms. One difficulty in optimizing a relaxation of the bi-objective optimization problem (2) comes from the non-smoothness of the balance objective. In this section, we consider a continuous and smooth version of the bi-objective optimization problem (2) given in the general form

$$\min\ (f^a(x), f^b(x)), \quad x \in \mathbb{R}^n, \tag{6}$$

and develop a companion algorithm of SAfairKM based on alternating gradient descent for each objective function. Notice that if we smooth out the min operators in the balance objective, a continuous relaxation of (2) falls into the smooth formulation (6).

The general idea of the alternating gradient descent method for the bi-objective optimization problem (6) consists of iteratively taking n_a steps of gradient descent for the first objective and then n_b steps of gradient descent for the second objective. For simplicity, we denote such $n_a + n_b$ steps as one single iteration of the algorithm. The stochastic alternating bi-objective gradient descent (SA2GD) algorithm is formally described in Algorithm 4. At every iteration t, starting at $x_t = y^a_{0,t}$, SA2GD computes two sequences of intermediate iterates, $\{y^a_{r,t}\}_{r=1}^{n_a}$ and $\{y^b_{r,t}\}_{r=1}^{n_b}$.

Furthermore, we assume that both objectives involve randomness in its parameters, in which case we expect that the true gradient $\nabla f^i(x)$ is not available or too expensive to compute. Instead, one can generate stochastic gradients as unbiased estimates of the true gradients, denoted by $g^i(x,\xi), i \in \{a,b\}$, where ξ is some random variable.

Our main result stated below (see proof in Appendix E) shows that SA2GD drives the expected optimality gap of the weighted-sum function $S(\cdot, \lambda(n_a,n_b))$ to zero at a sublinear rate of $1/T$, where $\lambda(n_a,n_b) = n_a/(n_a+n_b)$, when using a decaying step size sequence. By varying n_a and n_b in $\{1,\ldots,n_{\text{total}}\}$, such that $n_{\text{total}} = n_a + n_b$, one can capture the entire trade-off between f^a and f^b.

Theorem 5.1 ***(sublinear convergence rate of the SA2GD algorithm)*** *Let Assumptions E.1-E.4 hold and x_* be the unique minimizer of the weighted function $S(\cdot,\lambda_*)$ where $\lambda_* = \lambda(n_a,n_b) = n_a/(n_a+n_b)$. Choosing a diminishing step size sequence $\alpha_t = \frac{2}{c(t+1)(n_a+n_b)}$, the sequence of iterates generated by the Algorithm 4 satisfies*

$$\min_{t=1,\ldots,T} \mathbb{E}[S(x_t,\lambda_*)] - S(x_*,\lambda_*) \ \le\ \frac{4}{c(T+1)}\left(\frac{n_a^2+n_b^2}{n_a+n_b}\hat{G}^2 + (n_a+n_b)L\Theta\hat{G}\right).$$

where $\hat{G} = \sqrt{G + \bar{G}L^2\Theta^2}$, $L = \max(L^a, L^b)$, $G = \max(G^a, G^b)$, and $\bar{G} = \max(\bar{G}^a, \bar{G}^b)$.

The intuition of the proof of Theorem 5.1 is described as follows. At each iteration t, the algorithm generates two sequences $\{y^a_{r,t}\}_{r=1}^{n_a}$ and $\{y^b_{r,t}\}_{r=1}^{n_b}$ related to the alternated optimization of the two objectives. One can first write the explicit form of the new iterate x_{t+1} using the sequence of stochastic gradients and the step size α_t. Assuming that the sequence converges to some Pareto stationary point x_*, the optimality gap can be measured by the expected iterate error $\mathbb{E}[\|x_{t+1} - x_*\|^2]$, and this error can be bounded by the optimality gap at the current iteration, i.e., $\mathbb{E}[\|x_t - x_*\|^2]$, and two extra terms.

The first term involves α_t^2 and the square norms of the stochastic gradients, and can be bounded above using a combination of Assumptions E.1, E.3 (b), and E.4, as in classical stochastic gradient descent. The second term, after taking expectation over random variables, becomes $-\alpha_t(x_t -$

Algorithm 4 Stochastic alternating bi-objective gradient descent (SA2GD) algorithm

Input: Initial point $x_0 = y^a_{0,0}$ and a step size sequence $\{\alpha_t\}$.
Output: A likely nondominated or Pareto stationary point $x_T = y^a_{0,T}$.
for $t = 0, 1, \ldots, T-1$ **do**
  for $r = 0, \ldots, n_a - 1$ **do**
    Generate a stochastic gradient $g^a(y^a_{r,t}, \xi^r_t)$.
    Update $y^a_{r+1,t} = y^a_{r,t} - \alpha_t g^a(y^a_{r,t}, \xi^r_t)$.
  Set $y^b_{0,t} = y^a_{n_a,t}$.
  for $r = 0, \ldots, n_b - 1$ **do**
    Generate a stochastic gradient $g^b(y^b_{r,t}, \xi^{n_a+r}_t)$.
    Update $y^b_{r+1,t} = y^b_{r,t} - \alpha_t g^b(y^b_{r,t}, \xi^{n_a+r}_t)$.
  Set $x_{t+1} = y^a_{0,t+1} = y^b_{n_b,t}$.

$x_*)^\top(\sum_{r=0}^{n_a-1} \nabla f^a(y^a_{r,t}) + \sum_{r=0}^{n_b-1} \nabla f^b(y^b_{r,t}))$. By applying a version of the Intermediate Value Theorem (given in Appendix F), there exists a point w^i_t in the convex hull of $\{y^i_{r,t}\}_{r=0}^{n_i-1}$, for both $i \in \{a, b\}$, such that taking n_i steps along the sequence of gradients is the same as taking n_i steps along the gradient at w^i_t. The second term can then be written as $-\alpha_t(x_t - x_*)^\top(n_a \nabla f^a(w^a_t) + n_b \nabla f^b(w^b_t))$.

Next, we aim at establishing a connection with a weighted-sum function $n_a f^a + n_b f^b$ by seeing how this key second term differs from $-\alpha_t(x_t - x_*)^\top(n_a \nabla f^a(x_t) + n_b \nabla f^b(x_t))$. By applying the Cauchy–Schwarz inequality and Assumption E.1, the difference is of the same order as $\mathcal{O}(\alpha_t(\|x_t - w^a_t\| + \|x_t - w^b_t\|))$. Since one can show that $\|x_t - w^i_t\| = \mathcal{O}(\alpha_t)$, the second term is finally rewritten as $-\alpha_t(x_t - x_*)^\top(n_a \nabla f^a(x_t) + n_b \nabla f^b(x_t)) + \mathcal{O}(\alpha^2_t)$. The part involving ∇f^i is further handled using the strong convexity of the weighted-sum function associated with the weight $\lambda_* = n_a/(n_a + n_b)$, where x_* is now the unique minimizer of $S(\cdot, \lambda_*)$. The part $\mathcal{O}(\alpha^2_t)$ is merged to the other alike $\mathcal{O}(\alpha^2_t)$ which appears in the first term. The rest of the proof is classical, and by plugging in the decaying step size the desired rate is established. It is in fact the use of the decaying step size that compensates for the error $\mathcal{O}(\alpha^2_t)$ generated when bundling the gradients using the IVT.

6 Concluding remarks

We have investigated the natural conflict between the k-means clustering cost and the clustering balance from the perspective of bi-objective optimization, for which we designed a novel stochastic alternating algorithm (SAfairKM). A Pareto front version of SAfairKM has efficiently computed well-spread and high-quality trade-offs, when compared to an existing approach based on a penalization of fairness. We also proposed a companion stochastic alternating bi-objective gradient descent algorithm to handle a smooth version of these fair clustering bi-objective problems. Under strong convexity, this algorithm was shown to converge at a rate of $\mathcal{O}(1/T)$ to a Pareto stationary point, which is identified by weighting differently the effort put into the optimization of each function. This result validates our numerical experiments with SAfairKM in the sense that by varying this effort in a convex combination one can aim at determining an entire Pareto front.

Note that a balance improvement routine for the SAfairKM algorithm could be derived to handle more than one demographic group. One might formulate a multi-objective problem with the clustering cost being one objective and the balance corresponding to each protected attribute (e.g., race and gender) written as separate objectives. The balance measured using each attribute can be improved via alternating swap updates with respect to each balance objective.

Limitations The computation of entire Pareto fronts poses dataset scalability issues to SAfairKM due to the increase in function value evaluations for dominance checks. Moreover, the proposed approach is limited to disjoint demographic compositions and cannot deal with multiple and overlapping demographic groups.

References

[1] M. Abbasi, A. Bhaskara, and S. Venkatasubramanian. Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 504–514, 2021.

[2] S. S. Abraham and S. S. Sundaram. Fairness in clustering with multiple sensitive attributes. *arXiv preprint arXiv:1910.05113*, 2019.

[3] 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 and Data Mining*, pages 267–275, 2019.

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

[5] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, page 671, 2016.

[6] S. Bera, D. Chakrabarty, N. Flores, and M. Negahbani. Fair algorithms for clustering. In *Advances in Neural Information Processing Systems*, pages 4954–4965, 2019.

[7] P. Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

[8] L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In *Advances in neural information processing systems*, pages 585–592, 1995.

[9] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60:223–311, 2018.

[10] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

[11] X. Chen, B. Fain, L. Lyu, and K. Munagala. Proportionally fair clustering. In *International Conference on Machine Learning*, pages 1032–1041, 2019.

[12] F. Chierichetti, R. Kuma, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pages 5029–5037, 2017.

[13] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies*, 2015:92–112, 2015.

[14] D. Dua and C. Graff. UCI Machine Learning Repository, 2017.

[15] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[16] G. Gan, C. Ma, and J. Wu. *Data clustering: theory, algorithms, and applications*. SIAM, 2020.

[17] S. Gass and T. Saaty. The computational algorithm for the parametric objective function. *Nav. Res. Logist. Q.*, 2:39–45, 1955.

[18] M. Ghadiri, S. Samadi, and S. Vempala. Socially fair k-means clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 438–448, 2021.

[19] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[20] L. Huang, S. Jiang, and N. Vishnoi. Coresets for clustering with fairness constraints. In *Advances in Neural Information Processing Systems*, pages 7589–7600, 2019.

[21] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28:89–112, 2004.

[22] M. Kleindessner, P. Awasthi, and J. Morgenstern. Fair k-center clustering for data summarization. In *International Conference on Machine Learning*, pages 3448–3457. PMLR, 2019.

[23] M. Kleindessner, P. Awasthi, and J. Morgenstern. A notion of individual fairness for clustering. *arXiv preprint arXiv:2006.04960*, 2020.

[24] 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.

[25] R. Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 202–207. AAAI Press, 1996.

[26] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28:129–137, 1982.

[27] S. Mahabadi and A. Vakilian. Individual fairness for k-clustering. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6586–6596, Virtual, 13-18 Jul 2020. PMLR.

[28] S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[29] C. Rösner and M. Schmidt. Privacy preserving clustering with constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[30] M. Schmidt, C. Schwiegelshohn, and C. Sohler. Fair coresets and streaming algorithms for fair k-means. In *International Workshop on Approximation and Online Algorithms*, pages 232–251. Springer, 2019.

[31] S. Z. Selim and M. A. Ismail. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*, pages 81–87, 1984.

[32] I. M. Ziko, E. Granger, J. Yuan, and I. B. Ayed. Clustering with fairness constraints: A flexible and scalable approach. *arXiv preprint arXiv:1906.08207*, 2019.

A Mini-batch k-means algorithm

Algorithm 1 Mini-batch k-means algorithm

Input: The set of points P and an integer K.
Output: The set of centers $C = \{c_1, \ldots, c_K\}$.
Randomly select K points as initial centers.
for $t = 0, 1, 2, \ldots$ **do**
  Randomly sample a batch of points B_t.
  for $k = 1, \ldots, K$ **do**
    Identify the set of points $B_t^k \subseteq B_t$ whose closest center is c_k.
    $N_k = N_k + |B_t^k|$.
    $c_k = c_k + \frac{1}{N_k} \sum_{x_p \in B_t^k} (x_p - c_k)$.

B Pareto-Front SAfairKM Algorithm

Algorithm 3 Pareto-Front SAfairKM Algorithm

Generate a list of starting labels $\mathcal{L}_0$. Select parameter $q \in \mathbb{N}$ and a sequence of pairs $\mathcal{W} = \{(n_a, n_b) : n_a + n_b = n_{\text{total}}, n_a, n_b \in \mathbb{N}_0\}$.
for $t = 0, 1, \ldots$ **do**
  Set $\mathcal{L}_{t+1} = \mathcal{L}_t$.
  for each clustering label Δ in the list $\mathcal{L}_{k+1}$ **do**
    for $(n_a, n_b) \in \mathcal{W}$ **do**
      Apply q iterations of Algorithm 2 starting from Δ using the parameters (n_a, n_b).
      Add the final output label to the list $\mathcal{L}_{t+1}$.
  Remove all the dominated points from $\mathcal{L}_{t+1}$: **for** each label Δ in the list $\mathcal{L}_{t+1}$ **do**
  If $\exists\ \Delta' \in \mathcal{L}_{t+1}$ such that $f_1(\Delta') < f_1(\Delta)$ and $f_2(\Delta') > f_2(\Delta)$ hold, remove Δ.

C Description of an existing approach for comparison

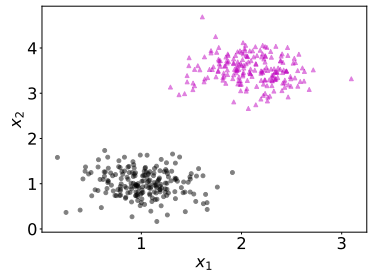
The authors in [32] considered the fairness error computed by the Kullback-Leibler (KL)-divergence, and added it as a penalized term to the classical clustering objective. When using the k-means clustering cost, the resulting problem takes the form:

$$\min\ f_1^{KM} + \mu \sum_{k=1}^{N} \mathcal{D}_{KL}(U \| \mathbb{P}_k) \quad \text{s.t. } s_p \in \Delta_K, \forall p = 1, \ldots, N, \tag{7}$$

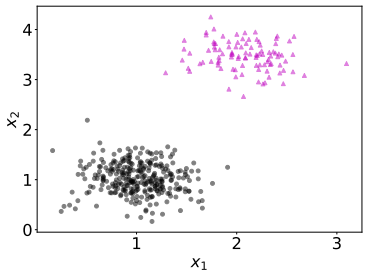
where $\mathcal{D}_{KL}$ is the KL divergence between the desired demographic proportion $U = [u_j, j = 1, \ldots, J]$ (usually specified by the demographic composition of the whole dataset) and the marginal probability $\mathbb{P}_k = [\mathbb{P}(j|k) = s_k^\top v_j / {e_N}^\top s_k, j = 1, \ldots, J]$. The penalty coefficient μ associated with the fairness error is the tool to control the trade-offs between the clustering cost and the clustering balance. To solve problem (7) for a fixed $\mu \geq 0$, the authors in [32] have developed an optimization scheme based on a concave-convex decomposition of the fairness term.

D More numerical results

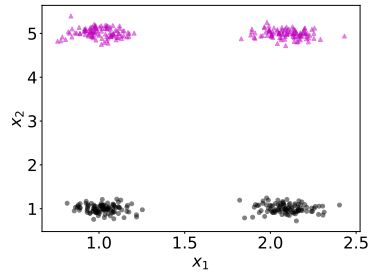
D.1 More trade-off results for synthetic datasets



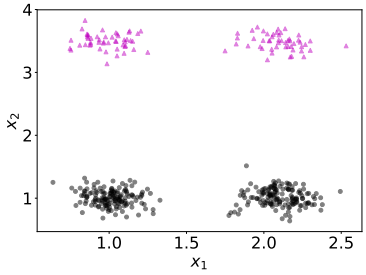
(a) Syn_equal_ds1: $|V_1| : |V_2| = 1 : 1$.



(b) Syn_unequal_ds1: $|V_1| : |V_2| = 1 : 3$.

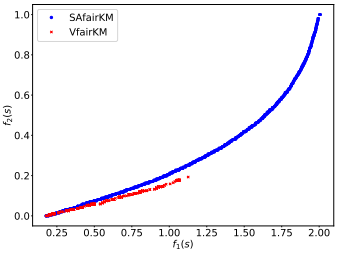


(c) Syn_equal_ds2: $|V_1| : |V_2| = 1 : 1$.

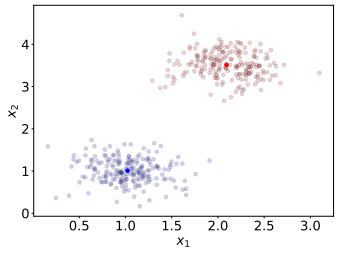


(d) Syn_unequal_ds2: $|V_1| : |V_2| = 1 : 3$.

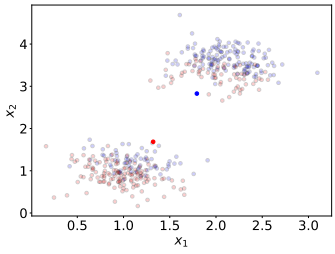
Figure 2: Demographic composition of four synthetic datasets.



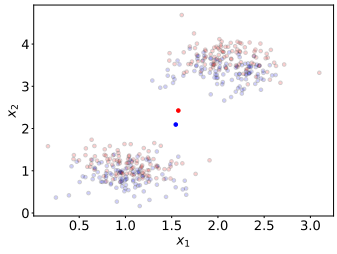
(a) Pareto front.



(b) Clust. balance $b = 0$.

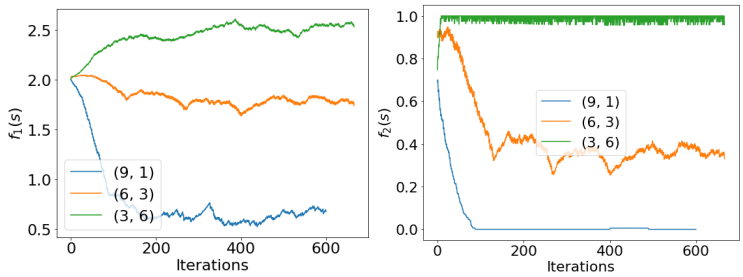


(c) Clust. balance $b = 0.5$.



(d) Clust. balance $b = 1$.

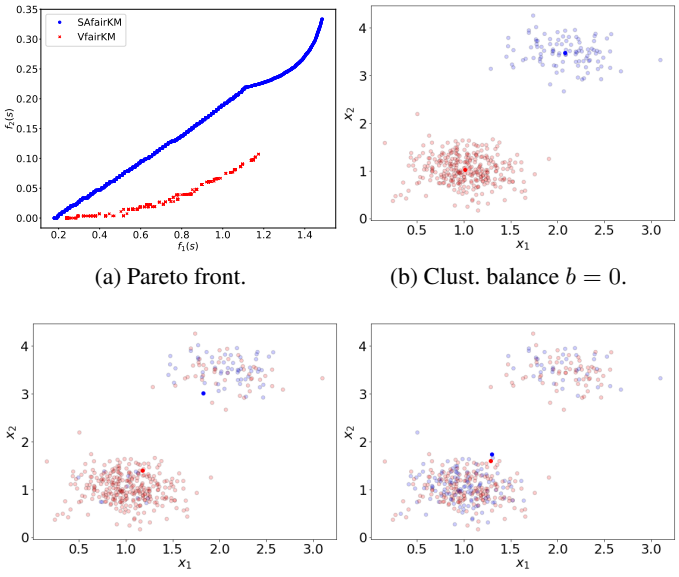
Figure 3: Syn_equal_ds1 data: SAfairKM: 400 iterations, 10 starting labels, and 3 pairs of (n_a, n_b); VfairKM: $\mu_{\max} = 202$.



(a) $f_1(s)$. (b) $f_2(s)$.

Figure 4: Syn_equal_ds1 dataset: 3 pairs of (n_a, n_b), $3,000/(n_a + n_b)$ iterations.

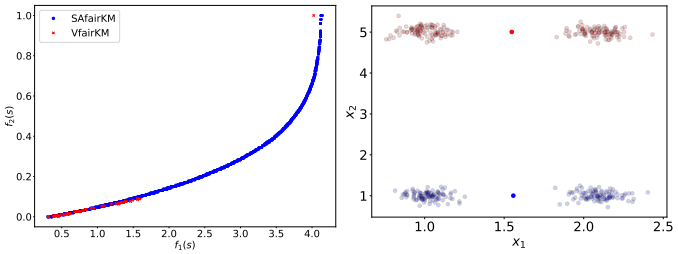
After having applied SAfairKM to the Syn_equal_ds1 dataset using three different pairs of (n_a, n_b), we show in Figure 4 how the two objective function values change along iterations. Taking far more k-means updates than swap updates, one converges to a nondominated solution of zero balance and minimum clustering cost. Using far more swap updates than k-means updates leads to a nondominated point of perfect balance and relatively higher clustering cost. In between, one can observe that after a certain number of iterations, the clustering cost and balance are driven to some intermediate range.



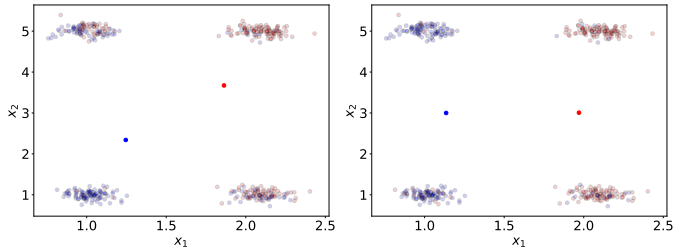
(a) Pareto front. (b) Clust. balance $b = 0$.

(c) Clust. balance $b = 0.16$. (d) Clust. balance $b = 0.33$.

Figure 5: Syn_unequal_ds1 data: SAfairKM: 400 iterations, 10 starting labels, and 3 pairs of (n_a, n_b); VfairKM: $\mu_{\max} = 223$.

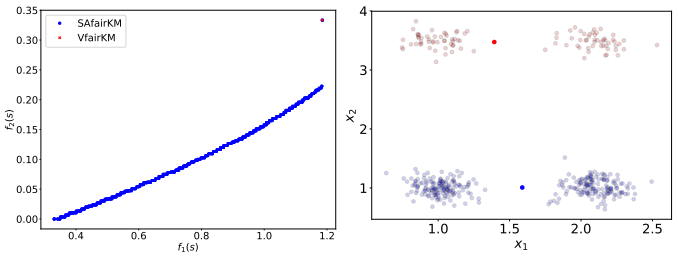


(a) Pareto front. (b) Clust. balance $b = 0$.

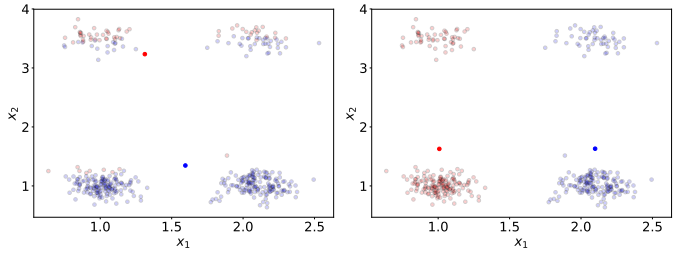


(c) Clust. balance $b = 0.5$. (d) Clust. balance $b = 1$.

Figure 6: Syn_equal_ds2 data: SAfairKM: 400 iterations, 10 starting labels, and 3 pairs of (n_a, n_b); VfairKM: $\mu_{\max} = 60$.



(a) Pareto front. (b) Clust. balance $b = 0$.

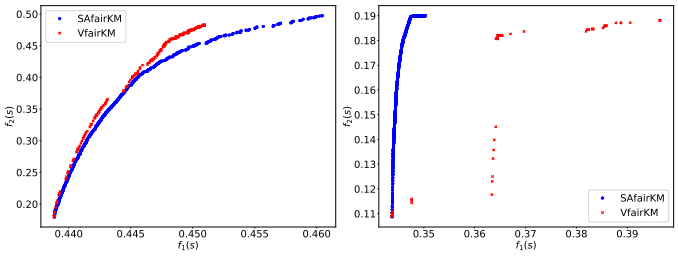


(c) Clust. balance $b = 0.17$. (d) Clust. balance $b = 0.33$.

Figure 7: Syn_unequal_ds2 data: SAfairKM: 400 iterations, 10 starting labels, and 3 pairs of (n_a, n_b); VfairKM: $\mu_{\max} = 0$.

Figure 7 (b)-(d) depicts the clustering outcomes associated with three nondominated points on the Pareto front computed by our algorithm. Specifically, Figure 7 (b) and (d) correspond to the two extreme points of zero balance and perfect balance respectively. The k-means updates tend to group the same demographic group into one cluster as points in the same demographic group are closer to each other. On the contrary, SAfairKM manages to find a clustering solution of perfect balance, where a half of each demographic group is assigned to each of the two clusters. Finally, Figure 7 (c) shows a clustering outcome of balance $b = 0.17$, where each cluster contains a mixture of points from the two demographic groups.

D.2 More trade-off results for real datasets



(a) Adult. (b) Bank.

Figure 8: Pareto fronts for $K = 5$: SAfairKM: 2500 iterations for Adult and 1500 iterations for Bank, 30 starting labels, and 4 pairs of (n_a, n_b); VfairKM: $\mu_{\max} = 6190$ for Adult and $\mu_{\max} = 4790$ for Bank.

E Rate of convergence for the stochastic alternating bi-objective gradient descent (assumptions and proof)

Let us now describe the assumptions under which SA2GD will be analyzed. We first formalize the classical smoothness assumption of Lipschitz continuity of the gradients, which is often satisfied in practice.

Assumption E.1 ***(Lipschitz continuous gradients)*** *The individual true gradients are Lipschitz continuous with Lipschitz constants $L^i > 0, i \in \{a, b\}$, i.e.,*

$$\|\nabla f^i(x) - \nabla f^i(\bar{x})\| \leq L^i \|x - \bar{x}\|, \quad \forall (x, \bar{x}) \in \mathbb{R}^n \times \mathbb{R}^n.$$

In addition to Assumption E.1, we impose strong convexity in both objective functions.

Assumption E.2 ***(Strong convexity)*** *For all $i = \{a, b\}$, there exists a scalar $c^i > 0$ such that*

$$f^i(\bar{x}_t) \geq f^i(x_t) + \nabla f^i(x_t)^\top (\bar{x}_t - x_t) + \frac{c^i}{2} \|\bar{x}_t - x_t\|^2, \quad \forall (x_t, \bar{x}_t) \in \mathbb{R}^n \times \mathbb{R}^n.$$

Note that based on the above assumption, the weighted-sum function $S(x, \lambda) = \lambda f^a(x) + (1 - \lambda) f^b(x), \lambda \in [0, 1]$ is also strongly convex with constant $c = \min(c^a, c^b)$. Given the individual stochastic gradients $g^i(x_t, \xi_t), \forall i \in \{a, b\}$, generated with random variable ξ_t, we use $\mathbb{E}_{\xi_t}[\cdot]$ to denote the conditional expectation taken with respect to ξ_t. We also impose the following two classical assumptions of stochastic gradients.

Assumption E.3 *For both objective functions $i \in \{a, b\}$, and all iterates $t \in \mathbb{N}$, the individual stochastic gradients $g^i(x_t, \xi_t)$ satisfy the following:*

(a) ***(Unbiased gradient estimation)*** $\mathbb{E}_{\xi_t}[g^i(x_t, \xi_t)] = \nabla f^i(x_t), \ \forall i \in \{a, b\}$.

(b) ***(Bound on the second moment)*** *There exist positive scalars $G^i > 0$ and $\bar{G}^i > 0$ such that*

$$\mathbb{E}_{\xi_t}[\|g^i(x_t, \xi_t)\|^2] \leq G^i + \bar{G}^i \|\nabla f^i(x_t)\|^2, \quad \forall i \in \{a, b\}.$$

The above assumptions are commonly used ones in classical stochastic gradient methods [9], basically assuming reasonable bounds on the expectation and variance of the individual stochastic gradients. Lastly, we make a reasonable assumption that the sequence of points generated by the algorithm is bounded.

Assumption E.4 *The sequence $\{x_t\}_{t=0}^T$ generated by the Algorithm 4 is contained in a bounded set, i.e., there exists a positive constant Θ such that*

$$\max_{t,t' \in \{0,\ldots,T\}} \|x_t - x_{t'}\| \leq \Theta < \infty.$$

Notice that the above assumption together with Assumption E.1 implies that $\|\nabla f^i(x_t) - \nabla f^i(x_{t'})\| \leq L^i \Theta$ holds for any $x_t, x_{t'}$ in the sequence. Letting $x_{t'}$ be the unique minimizer of $f^i(x)$ results in an upper bound on the true gradient norm, i.e., $\|\nabla f^i(x_t)\| \leq L^i \Theta, \forall i \in \{a, b\}$.

Proof of Theorem 5.1. The proof is divided in three parts for better organization and understanding. In the first part, one obtains an upper bound on the norm of the iterates, $\mathbb{E}[\|x_{t+1} - x_*\|^2]$. Strong convexity of the weighted-sum function is applied in the second part. The third part concludes the rate using standard arguments. For simplicity, we let $z_t = y^b_{0,t} = y^a_{n_a,t}$.

Part I: Bound on the iterates error. At any iteration t, the sequence of stochastic gradients is computed from drawing the sequence of random variables $\xi_t = \{\xi_t^0, \dots, \xi_t^{n_a+n_b-1}\}$. We have

$$\begin{aligned} x_{t+1} - x_* &= z_t - x_* - \alpha_t \sum_{r=0}^{n_b-1} g^b(y^b_{r,t}, \xi_t^r), \\ &= x_t - x_* - \alpha_t \sum_{r=0}^{n_a-1} g^a(y^a_{r,t}, \xi_t^r) - \alpha_t \sum_{r=0}^{n_b-1} g^b(y^b_{r,t}, \xi_t^{n_a+r}). \end{aligned}$$

Since the sequence ξ_t is drawn independently, using Assumption E.3 (a) one has

$$\mathbb{E}_{\xi_t}[g^a(y^a_{r,t}, \xi_t^r)] = \mathbb{E}_{\xi_t^0,\dots,\xi_t^{r-1}}[\mathbb{E}_{\xi_t^r}[g^a(y^a_{r,t}, \xi_t^r)]] = \mathbb{E}_{\xi_t^0,\dots,\xi_t^{r-1}}[\nabla f^a(y^a_{r,t})] = \mathbb{E}_{\xi_t}[\nabla f^a(y^a_{r,t})],$$

where the second and last equalities hold due to the independence between $y^a_{r,t}$ and $\{\xi_t^r, \dots, \xi_t^{n_a-1}\}$. Similarly, we have $\mathbb{E}_{\xi_t}[g^b(y^b_{r,t}, \xi_t^{n_a+r})] = \mathbb{E}_{\xi_t}[\nabla f^b(y^b_{r,t})]$.

Taking square norms and expectations over the random variables ξ_t on both sides and using Assumption E.3 (a) yield

$$\begin{aligned} \mathbb{E}_{\xi_t}[\|x_{t+1} - x_*\|^2] &= \|x_t - x_*\|^2 + \alpha_t^2 \mathbb{E}_{\xi_t}[\| \sum_{r=0}^{n_a-1} g^a(y^a_{r,t}, \xi_t^r) + \sum_{r=0}^{n_b-1} g^b(y^b_{r,t}, \xi_t^{n_a+r})\|^2] \\ &\quad - \mathbb{E}_{\xi_t}[2\alpha_t (x_t - x_*)^\top \sum_{r=0}^{n_a-1} \nabla f^a(y^a_{r,t}) + \sum_{r=0}^{n_b-1} \nabla f^b(y^b_{r,t})]. \end{aligned} \tag{8}$$

We now claim, by applying a version of the Intermediate Value Theorem given in Proposition F.1, that the last term of the right-hand side in (8) can be written as $-2\alpha_t (x_t - x_*)^\top \mathbb{E}_{\xi_t}[n_a \nabla f^a(w_t^a) + n_b \nabla f^b(w_t^b)]$. In fact, we apply Proposition F.1 to the real continuous function $\phi^i(y) = 2\alpha_t (x_t - x_*)^\top \nabla f^i(y)$, from which we then know that w_t^i is a convex combination of a sequence of points $\{y^i_{r,t}\}_{r=0}^{n_i-1}$ for both $i \in \{a, b\}$.

As for the second term in the right-hand side of (8), we can use a combination of Assumptions E.1, E.3 (b), and E.4 to derive an upper bound for the second moment of the sequence of stochastic gradients $\{g^a(y^a_{r,t}, \xi_t^r)\}_{r=0}^{n_a-1}$ and $\{g^b(y^b_{r,t}, \xi_t^{n_a+r})\}_{r=0}^{n_b-1}$ at each iteration t, i.e.,

$$\begin{aligned} \mathbb{E}_{\xi_t}[\| \sum_{r=0}^{n_a-1} g^a(y^a_{r,t}, \xi_t^r) + \sum_{r=0}^{n_b-1} g^b(y^b_{r,t}, \xi_t^{n_a+r})\|^2] &\le 2\mathbb{E}_{\xi_t}[\| \sum_{r=0}^{n_a-1} g^a(y^a_{r,t}, \xi_t^r)\|^2] + 2\mathbb{E}_{\xi_t}[\| \sum_{r=0}^{n_b-1} g^b(y^b_{r,t}, \xi_t^{n_a+r})\|^2] \\ &\le 2n_a \sum_{r=0}^{n_a-1} \mathbb{E}_{\xi_t}[\|g^a(y^a_{r,t}, \xi_t^r)\|^2] \\ &\quad + 2n_b \sum_{r=0}^{n_b-1} \mathbb{E}_{\xi_t}[\|g^b(y^b_{r,t}, \xi_t^{n_a+r})\|^2] \\ &\le 2(n_a^2 + n_b^2)(G + \bar{G} L^2 \Theta^2), \end{aligned}$$

where $L = \max(L^a, L^b)$, $G = \max(G^a, G^b)$, and $\bar{G} = \max(\bar{G}^a, \bar{G}^b)$. We thus arrive at

$$\begin{aligned} \mathbb{E}_{\xi_t}[\|x_{t+1} - x_*\|^2] &\le \|x_t - x_*\|^2 + 2\alpha_t^2 (n_a^2 + n_b^2)(G + \bar{G} L^2 \Theta^2) \\ &\quad - 2\alpha_t (x_t - x_*)^\top \mathbb{E}_{\xi_t}[n_a \nabla f^a(w_t^a) + n_b \nabla f^b(w_t^b)]. \end{aligned} \tag{9}$$

By adding and subtracting $2\alpha_t (x_t - x_*)^\top (n_a \nabla f^a(x_t) + n_b \nabla f^b(x_t))$ in the right-hand side of (9), we further rewrite it as

$$\begin{aligned} \mathbb{E}_{\xi_t}[\|x_{t+1} - x_*\|^2] &\le \|x_t - x_*\|^2 + 2\alpha_t^2 (n_a^2 + n_b^2)(G + \bar{G} L^2 \Theta^2) \\ &\quad - 2\alpha_t (x_t - x_*)^\top (n_a \nabla f^a(x_t) + n_b \nabla f^b(x_t)) \\ &\quad + 2\alpha_t \|x_t - x_*\| \mathbb{E}_{\xi_t}[\|n_a \nabla f^a(w_t^a) - n_a \nabla f^a(x_t)\|] \\ &\quad + 2\alpha_t \|x_t - x_*\| \mathbb{E}_{\xi_t}[\|n_b \nabla f^b(w_t^b) - n_b \nabla f^b(x_t)\|]. \end{aligned} \tag{10}$$

Note that the last two terms are derived by the Cauchy–Schwarz and Jensen's inequalities.

Part II: Using strong convexity. Selecting $\lambda_* = \lambda(n_a, n_b) = n_a/(n_a + n_b)$, by the strong convexity of the weighted-sum function, one has

$$\nabla_x S(x_t, \lambda_*)^\top (x_t - x_*) \;\geq\; S(x_t, \lambda_*) - S(x_*, \lambda_*) + \frac{c}{2}\|x_t - x_*\|^2,$$

which is equivalent to

$$(x_t - x_*)^\top (n_a \nabla f^a(x_t) + n_b \nabla f^b(x_t)) \;\geq\; (n_a + n_b)(S(x_t, \lambda_*) - S(x_*, \lambda_*) + \frac{c}{2}\|x_t - x_*\|^2). \tag{11}$$

From Assumption E.1, we obtain a bound for the last two terms of (10) in the form

$$2\alpha_t \|x_t - x_*\| \mathbb{E}_{\xi_t}[\|n_i \nabla f^i(w_t^i) - n_i \nabla f^i(x_t)\|] \;\leq\; 2\alpha_t L n_i \|x_t - x_*\| \mathbb{E}_{\xi_t}[\|x_t - w_t^i\|], \forall i \in \{a, b\}. \tag{12}$$

According to Proposition F.1, w_t^i is a convex combination of a sequence of points $\{y_{r,t}^i\}_{r=0}^{n_i}$ for $i \in \{a, b\}$. One can write $w_t^i = \sum_{r=0}^{n_i - 1} \beta_r y_{r,t}^i$ with $\beta_r \geq 0$, $r = 0, \ldots, n_i - 1$, and $\sum_{r=0}^{n_i-1} \beta_r = 1$. An explicit upper bound of $\|x_t - w_t^i\|$ can then be derived using Jensen's inequality

$$\mathbb{E}_{\xi_t}[\|x_t - \sum_{r=0}^{n_i-1} \beta_r y_{r,t}^i\|] \;=\; \mathbb{E}_{\xi_t}[\|\sum_{r=0}^{n_i-1} \beta_r (x_t - y_{r,t}^i)\|] \;\leq\; \sum_{r=0}^{n_i-1} \beta_r \mathbb{E}_{\xi_t}[\|x_t - y_{r,t}^i\|]. \tag{13}$$

Using $y_{r,t}^i = y_{0,t}^i - \sum_{j=0}^{r-1} \alpha_t g^i(y_{j,t}^i, \xi_t^j)$ and applying the triangle inequality, we have

$$\mathbb{E}_{\xi_t}[\|x_t - y_{r,t}^a\|] \;\leq\; \alpha_t \sum_{j=0}^{r-1} \mathbb{E}_{\xi_t}[\|g^a(y_{j,t}^a, \xi_t^j)\|], \tag{14}$$

and

$$\begin{aligned} \mathbb{E}_{\xi_t}[\|x_t - y_{r,t}^b\|] \;&=\; \mathbb{E}_{\xi_t}[\|x_t - y_{n_a,t}^a + \sum_{j=0}^{r-1} \alpha_t g^b(y_{j,t}^b, \xi_t^{n_a+j})\|] \\ &\leq\; \alpha_t \sum_{j=0}^{n_a-1} \mathbb{E}_{\xi_t}[\|g^a(y_{j,t}^a, \xi_t^j)\|] + \alpha_t \sum_{j=0}^{r-1} \mathbb{E}_{\xi_t}[\|g^b(y_{j,t}^b, \xi_t^{n_a+r})\|]. \end{aligned} \tag{15}$$

Recall that using a combination of Assumptions E.1, E.3 (b), and E.4, the bound for the second moment of the stochastic gradients is given by $\mathbb{E}_\xi[\|g^i(y_{r,t}^i)\|^2] \leq G + \bar{G}L^2\Theta^2$, where L, G, and $\bar{G}$ are constants defined in Part I. Plugging (14) into (13) with $i = a$ results in

$$\begin{aligned} \mathbb{E}_{\xi_t}[\|x_t - w_t^a\|] \;&\leq\; \sum_{r=1}^{n_a-1} \beta_r \alpha_t \sum_{j=0}^{r-1} \mathbb{E}_{\xi_t}[\|g^a(y_{j,t}^a, \xi_t^j)\|] \\ &\leq\; \alpha_t \sqrt{G + \bar{G}L^2\Theta^2} \sum_{r=0}^{n_a-1} \beta_r\, r \;\leq\; \alpha_t (n_a - 1)\sqrt{G + \bar{G}L^2\Theta^2}. \end{aligned} \tag{16}$$

Similarly, plugging (15) into (13) with $i = b$ leads to

$$\mathbb{E}_{\xi_t}[\|x_t - w_t^b\|] \;\leq\; \alpha_t (n_a + n_b - 1)\sqrt{G + \bar{G}L^2\Theta^2}. \tag{17}$$

Finally, combining (12) and (16)-(17) yields

$$2\alpha_t \|x_t - x_*\| \sum_{i \in \{a,b\}} \mathbb{E}_{\xi_t}[\|n_i \nabla f^i(w_t^i) - n_i \nabla f^i(x_t)\|] \;\leq\; 2\alpha_t^2 L\Theta (n_a + n_b)^2 \sqrt{G + \bar{G}L^2\Theta^2}. \tag{18}$$

Part III: Bound on the optimality gap in terms of weighted-sum function. Applying inequalities (11) and (18) to (10) leads to

$$\begin{aligned} \mathbb{E}_{\xi_t}[\|x_{t+1} - x_*\|^2] \;&\leq (1 - \alpha_t(n_a + n_b)c)\|x_t - x_*\|^2 + 2\alpha_t^2((n_a^2 + n_b^2)(G + \bar{G}L^2\Theta^2) \\ &\quad + (n_a + n_b)^2 L\Theta\sqrt{G + \bar{G}L^2\Theta^2}) - 2\alpha_t(n_a + n_b)(S(x_t, \lambda_*) - S(x_*, \lambda_*)). \end{aligned}$$

To simplify, let $\bar{M} = 2((n_a^2 + n_b^2)(G + \bar{G}L^2\Theta^2) + (n_a + n_b)^2 L\Theta\sqrt{G + \bar{G}L^2\Theta^2})$. Plugging in $\alpha_t = \frac{2}{c(n_a+n_b)(t+1)}$ and rearranging the last inequality results in

$$\begin{aligned} S(x_t, \lambda_*) - S(x_*, \lambda_*) \;&\leq \frac{(1 - \alpha_t(n_a + n_b)c)\|x_t - x_*\|^2 - \mathbb{E}_{\xi_t}[\|x_{t+1} - x_*\|^2] + \alpha_t^2 \bar{M}}{2\alpha_t(n_a + n_b)} \\ &\leq \frac{c(t-1)}{4}\|x_t - x_*\|^2 - \frac{c(t+1)}{4}\mathbb{E}_{\xi_t}[\|x_{t+1} - x_*\|^2] + \frac{\tilde{M}}{c(t+1)}, \end{aligned}$$

where $\tilde{M} = \frac{\bar{M}}{n_a+n_b}$. Taking total expectation over $\{\xi_t\}$, multiplying both sides by t, and summing over $t = 1, \ldots, T$, one obtains

$$\begin{aligned}\sum_{t=1}^{T} t(\mathbb{E}[S(x_t, \lambda_*)] - S(x_*, \lambda_*)) \;&\leq\; \sum_{t=1}^{T}\left(\frac{ct(t-1)}{4}\mathbb{E}[\|x_t - x_*\|^2] - \frac{ct(t+1)}{4}\mathbb{E}[\|x_{t+1} - x_*\|^2]\right) \\ &\quad + \sum_{t=1}^{T}\frac{\tilde{M}t}{c(t+1)} \\ &\leq\; -\frac{cT(T+1)}{4}\mathbb{E}[\|x_{T+1} - x_*\|^2] + \sum_{t=1}^{T}\frac{\tilde{M}t}{c(t+1)} \;\leq\; \frac{T}{c}\tilde{M}.\end{aligned}$$

Dividing both sides of the last inequality by $\sum_{t=1}^{T} t$ yields

$$\min_{t=1,\ldots,T} \mathbb{E}[S(x_t, \lambda_*)] - S(x_*, \lambda_*) \;\leq\; \frac{2}{c(T+1)}\tilde{M},$$

which concludes the proof. □

F Proposition using Intermediate Value Theorem

Based on the Intermediate Value Theorem, we derive the following proposition for the purpose of convergence rate analysis of Algorithm 4 (SA2GD).

Proposition F.1 *Given a continuous real function $\phi(x) : \mathbb{R}^n \to \mathbb{R}$ and a set of points $\{x_j\}_{j=1}^m$, there exists $w \in \mathbb{R}^n$ such that*

$$m\phi(w) = \sum_{j=1}^{m} \phi(x_j),$$

where $w = \sum_{j=1}^m \mu_j x_j$, with $\sum_{j=1}^m \mu_j = 1, \mu_j \geq 0$, $i = 1, \ldots, m$, is a convex combination of $\{x_j\}_{j=1}^m$.

Proof. The proposition is obtained by consecutively applying the Intermediate Value Theorem to $\phi(x)$. First, for the pair of points x_1 and x_2, there exists a point $w_{12} = \mu_{12}x_1 + (1-\mu_{12})x_2, \mu_{12} \geq 0$, such that $\phi(w_{12}) = (\phi(x_1) + \phi(x_2))/2$ according to the Intermediate Value Theorem, which implies that $\sum_{j=1}^m \phi(x_j) = 2\phi(w_{12}) + \sum_{j=3}^m \phi(x_j)$. Then, there exists $w_{13} = \mu_{13}w_{12} + (1-\mu_{13})x_3, \mu_{13} \geq 0$, such that $\phi(w_{13}) = (2\phi(w_{12}) + \phi(x_3))/3$ holds given that the average function value $(2\phi(w_{12}) + \phi(x_3))/3$ lies between $\phi(w_{12})$ and $\phi(x_3)$. Notice that w_{13} can also be written as convex combination of $\{y_1, y_2, y_3\}$. The proof is concluded by continuing this process until x_m is reached. □