Max-Min Diversification and Monotone Nonnegative Submodular Functions under Fairness Constraints

Zafeiria Moumoulidou
SID: 31799075

1 Introduction and problem definition

Result diversification is an important concept frequently used in the context of search engines, information retrieval and summarization among others. As a motivating example, consider a query over the *#megexit* hashtag. [1] A *qualitative* result set would be a collection of documents that are *relevant* to the query and *diverse* in context. In this work, we capture the diversity requirement using the Max-Min diversification model, which is also known as the *p-dispersion* or *remote-edge* problem [11, 12]. In the Max-Min diversification problem, we aim to select a set of k elements whose minimum pairwise distance in some metric space is maximized. Namely, in the *#megexit* example we want to maximize the *dissimilarity* in context of any two documents. Moreover, in the literature a common way to model *relevance*, or some *utility* function properly defined for an application of interest, is via submodular functions [2, 4, 10]. In this work, we focus on monotone nonnegative submodular functions which *reward* a set of elements according to their *relevance* or *utility*; namely an *irrelevant* set is not penalized but instead its value is equal to zero. Nonetheless, *diversity* in context alone is not always sufficient to produce *good* result sets. An orthogonal requirement is for the document collection to cover the available sources (news channels) in order to guarantee an objective representation of the information. In order to address this additional requirement, we introduce the notion of *fairness* constraints defined over a categorical attribute with m different values and we request that the final result set contains k_i representatives from the i-th group. Formally, the problem definition is as follows:

Problem Definition: We assume a dataset $\mathcal{X}$ of size n partitioned into m disjoint groups; $\mathcal{X} = \bigcup_{i=1}^{m} \mathcal{X}_i$, a metric distance function $d(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_0^+$ along with a monotone nonnegative submodular function $f(\cdot) : 2^{\mathcal{X}} \to \mathbb{R}_0^+$ and a set of *fairness* constraints $\langle k_1, k_2, \cdots, k_m \rangle$ with $k_i \leq |\mathcal{X}_i|, \forall\, i \in [m]$. Then the problem we are studying is as follows:

$$\begin{aligned} \underset{\mathcal{S} \subseteq \mathcal{X}}{\text{maximize}} \quad & f(\mathcal{S}) + \lambda\, div(\mathcal{S}) \\ \text{subject to} \quad & |\mathcal{S}| = k,\ |\mathcal{S} \cap \mathcal{X}_i| = k_i,\ \forall\, i \in [m] \end{aligned} \tag{1}$$

where $\lambda \in \mathbb{R}^+$ is a trade-off parameter between the two objectives, $k = \sum_{i=1}^{m} k_i$ and $div(\mathcal{S})$ is equal to:

$$\min_{s_i, s_j \in \mathcal{S}, s_i \neq s_j} d(s_i, s_j)$$

The Max-Min diversification problem was shown to be NP-hard using a reduction from the Max-Clique problem [11]. This fact immediately implies that problem 1 is also NP-hard so we can only hope for approximate solutions.

2 Background and related work

Fairness constraints as a partition matroid [2] ***:*** The key observation that guides the algorithmic framework we propose for problem 1 is the fact that the *fairness* constraints can be modelled by a partition matroid. Below we briefly provide the definition of a partition matroid:

Definition 1. (Partition Matroid) *A matroid $\mathcal{M} = (\mathcal{X}, \mathcal{I})$, where $\mathcal{X}$ is the ground set and $\mathcal{I}$ is the collection of the independent sets, is a partition matroid if $\mathcal{X}$ can be decomposed into m disjoint sets $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_m$ and $\mathcal{I}$ is defined as*

$$\mathcal{I} = \{\mathcal{S} \subseteq \mathcal{X} : |\mathcal{S} \cap \mathcal{X}_i| \leq k_i\ \forall\, i \in [m]\}$$

Definition 2. *A maximal independent set in $\mathcal{I}$ (also called a basis for a matroid) is a set for which there is no element outside of the set that can be added so that the set still remains independent.*

[1] #megexit is the hashtag referring to the announcement that the Dutch and Dutchess of Sussex made regarding their intention to step back as senior members of the royal family and become financially independent.

[2] A matroid $\mathcal{M}$ is a combinatorial object defined as a pair of $(\mathcal{E}, \mathcal{I})$ where $\mathcal{E}$ is a ground set and the set of independent sets $\mathcal{I}$ is composed of any subset $X \subseteq \mathcal{E}$ that satisfies the *hereditary* and *exchange* property.

For the partition matroid, a maximal independent set (or a basis) is an independent set that satisfies all the cardinality constraints with equality, namely has a size equal to $\sum_{i=1}^{m} k_i = k$. Therefore, the optimization problem can now be rephrased as maximizing the objective function over all the bases of a partition matroid $\mathcal{M}$ defined on the dataset $\mathcal{X}$. Formally, we have:

$$\underset{\mathcal{S} \in \mathcal{I} : |\mathcal{S}| = k}{\text{maximize}} \quad f(S) + \lambda \; div(\mathcal{S})$$

Diversity Maximization: The Max-Min diversification problem was first studied by Ravi et al. [11] and Tamir [12] in the context of facility location. They independently designed GMM algorithm 2 which was shown to be a $\frac{1}{2}-$approximation algorithm that at each step greedily picks to add in the set the point that is the farthest from all the other elements selected so far. Dasgupta et al. [4] study the problem of the Max-Min diversification problem, without the *fairness* constraints, in conjuction with a monotone nonnegative submodular function in the context of summarization. The authors show that a simple two-round greedy algorithm that separately optimizes each part of the objective and then selects the set for which the value of the objective function is maximized, gives a $\frac{1}{4}-$ approximation algorithm. To the best of our knowledge, the Max-Min diversification problem has not been studied under *fairness* constraints. In this work, we combine the Max-Min objective with a monotone, nonnegative submodular function and investigate whether a similar approach to the one designed by Dasgupta et al. [4] holds for the *fairness* constrained setting. In the diversification literature, the Max-Sum diversification variant where the objective to be maximized is the average pairwise distance in a set was studied under matroid constraints by Abassi et al. [1]. In this setting, the retrieved set is a basis of the matroid retrieved using a local search algorithm with a $\frac{1}{2}-$ approximation guarantee. Borodin et al. [2] formulate a bi-criteria optimization problem under matroid constraints defined as the sum of the Max-sum objective and a a monotone, nonnegative submodular function and show that the local search approach preserves the $\frac{1}{2}-$ approximation factor.

Submodular Maximization: Submodular maximization under matroid constraints was first studied in [8]. Fisher et al. [8] showed that a simple greedy heuristic that at each step adds the point with the maximum marginal gain while maintaining a solution that is an independent set provides a $\frac{1}{2}-$ approximation guarantee in the presence of a matroid. Calinesu et al. [3] and Filmus et al. [7] design algorithms that further improve the result to an $(1 - \frac{1}{e})$-approximation factor, which is shown to be the best result we can obtain [6]. For a detailed overview of the submodular maximization literature we refer the reader to [9].

3 Approach

The *fair* Max-Min diversification problem follows the definition of problem 1 when $f(S) = 0$ and $\lambda = 1$. As part of my research project, we designed the FAIR-FLOW algorithm which we are going to use as a black box for the algorithmic framework we propose. Due to space constraints, we only give a high level description of the algorithm. The main idea is that the optimal Max-Min diversity score is an edge in the dataset $\mathcal{X}$ so we can try to *guess* which edge out of the $\binom{n}{2} \simeq n^2$ edges could be the optimal one. We reduce the number of guesses to k^2m^2 by selecting a diverse set of k elements in each partition separately using GMM 2. Then we perform a binary search over the possible guesses and for each guess γ, we run the FAIR-FLOW algorithm. If the procedure fails to find a set of k elements that are at least $d_2 = \frac{\gamma}{3m-1}$ distance apart, we move towards to a guess with a smaller value and if we succeed to a guess with a larger value. We return the set with the best diversity score. The FAIR-FLOW algorithm reduces the problem of selecting k elements that are at least $d_2 = \frac{\gamma}{3m-1}$ apart to a max-flow problem.

Theorem 1. *The* FAIR-FLOW *algorithm is a* $3m-1$*-approximation algorithm for the* fair *Max-Min diversification problem.*

The algorithmic framework we propose for problem 1 uses the technique introduced in [3]. At first, we set $\lambda = 0$ and reduce the problem down to the problem of submodular maximization under a matroid constraint. Using algorithm 3, we retrieve the S_1 set, which is an $\frac{1}{2}-$ approximate solution [8] with respect to the nonnegative, monotone submodular function in the objective. Then, we set $f(S) = 0$ and the problem reduces down to the *fair* Max-Min diversification problem. Using the FAIR-FLOW algorithm, we retrieve the S_2 set which is a $\frac{1}{3m-1}$ approximate solution. We set the solution of problem 1 to be equal to

$$S = \text{argmax}_{S' \in \{S_1, S_2\}} \; f(S') + \lambda \; div(S')$$

Theorem 2. *The two-round algorithm is a* $\frac{1}{2(3m-1)}$*-approximation algorithm for problem 1.*

Proof. Let S_1^*, S_2^* and S^* be the optimal solution for the submodular part of the objective, for the diversity part of the objective and for problem 1 respectively. Then from the approximation guarantees of the FAIR-FLOW algorithm 1 and the greedy algorithm 3, we get that:

$$f(S_1) \geq \tfrac{1}{2} f(S_1^*)$$
$$div(S_2) \geq \tfrac{1}{3m-1} div(S_2^*)$$

Algorithm 1 FAIR-FLOW Algorithm

Input: $\mathcal{X} = \bigcup_{i=1}^{m} \mathcal{X}_i$: Universe of available elements

$k_1, \ldots, k_m \in \mathbb{Z}^+$

$\gamma \in \mathbb{R}$: A guess of the optimum fair diversity.

Output: k_i points in $\mathcal{X}_i$ for $i \in [m]$

procedure FAIR-FLOW
  for $i \in [m]$ **do**
    $Y_i \leftarrow \text{GMM}(\mathcal{X}_i, \emptyset, k)$
  $Z_i \leftarrow$ maximal prefix of Y_i such that all points in Z_i are $\geq d_1 = \frac{m\gamma}{3m-1}$ apart.
  Construct undirected graph G_Z with nodes $Z = \bigcup_i Z_i$ and edges (z_1, z_2) if $d(z_1, z_2) < d_2 = \frac{\gamma}{3m-1}$
  $C_1, C_2, \ldots C_t \leftarrow$ Connected components of G_Z
  ▷Construct flow graph
  Construct directed graph $G = (V, E)$ where

$$\begin{aligned} V &= \{a, u_1, \ldots, u_m, v_1, \ldots, v_t, b\} \\ E &= \{(a, u_i) \text{ with capacity } k_i : i \in [m]\} \\ &\quad \cup \{(v_j, b) \text{ with capacity } 1 : j \in [t]\} \\ &\quad \cup \{(u_i, v_j) \text{ with capacity } 1 : |Z_i \cap C_j| \geq 1\} \end{aligned}$$

  Compute max a-b flow.
  if flow size $< k = \sum_i k_i$ **then return** $\emptyset$ ▷Abort
  else ▷max flow is k
    $\forall (u_i, v_j)$ with flow add a node in C_j with color i to $\mathcal{S}$
  return $\mathcal{S}$

Algorithm 2 GMM Algorithm [11, 12]

Input: $\mathcal{X} = \bigcup_{i=1}^{m} \mathcal{X}_i$, $k \in \mathbb{Z}^+$

Output: $\mathcal{S} \subseteq \mathcal{X}$ of size k

procedure GMM($\mathcal{X}, I, k$)
  $s_1 \leftarrow$ a randomly chosen point in $\mathcal{X}$
  $\mathcal{S} \leftarrow s_1$
  while $|\mathcal{S}| < k$ **do**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \underset{x \in \mathcal{X} \setminus \mathcal{S}}{\text{argmax}} \min_{s \in \mathcal{S} \cup I} d(x, s)$
  return $\mathcal{S}$

Algorithm 3 Greedy Algorithm for Submodular Maximization [8, 9]

Input: $\mathcal{X} = \bigcup_{i=1}^{m} \mathcal{X}_i$, $\mathcal{M}(\mathcal{X}, \mathcal{I})$, $k \in \mathbb{Z}^+$

Output: $\mathcal{S} \subseteq \mathcal{X}$ of size k

procedure MAXMARGINALGAIN($\mathcal{X}, k$)
  $\mathcal{S} \leftarrow \emptyset$.
  while $|\mathcal{S}| < k$ **do**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \underset{x \notin \mathcal{S}: \mathcal{S} \cup \{x\} \in \mathcal{I}}{\text{argmax}} f(\mathcal{S} \cup \{x\}) - f(\mathcal{S})$
  return $\mathcal{S}$

Figure 1: Algorithmic framework for problem 1

Moreover, by the way S is selected we know that the following holds:

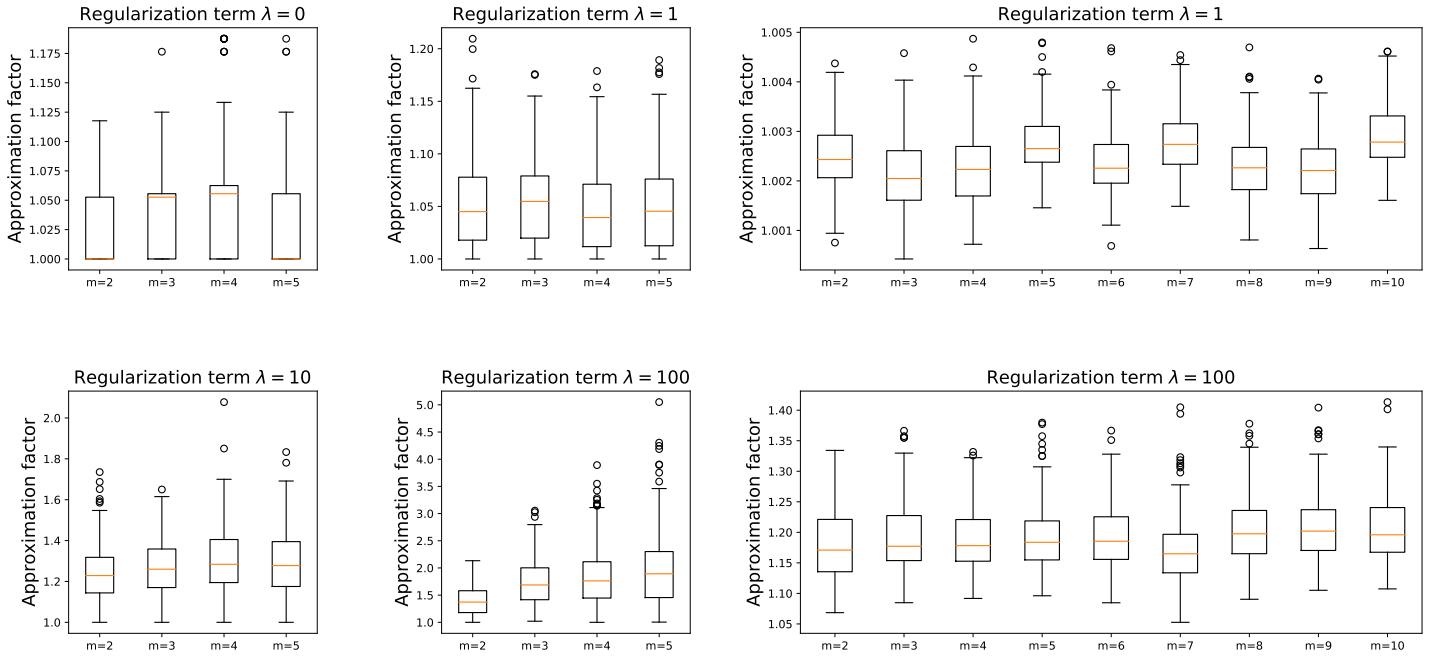
$$f(S) + \lambda\ div(S) \geq f(S_1) + \lambda\ div(S_1) \tag{2}$$
$$f(S) + \lambda\ div(S) \geq f(S_2) + \lambda\ div(S_2) \tag{3}$$

Combining 2, 3 and using the fact that both f and the Max-Min diversity are nonnegative functions we get that:

$$\begin{aligned} f(S) + \lambda\ div(S) &\geq \frac{1}{2}\big(f(S_1) + \lambda\ div(S_1) + f(S_2) + \lambda\ div(S_2)\big), \text{ due to nonnegativity of } f \text{ and } div \\ &\geq \frac{1}{2}\big(f(S_1) + \lambda\ div(S_2)\big) \\ &\geq \frac{1}{2}\Big(\frac{1}{2}f(S_1^*) + \lambda\ \frac{1}{3m-1}div(S_2^*)\Big) \\ &\geq \frac{1}{4}f(S_1^*) + \lambda\ \frac{1}{2(3m-1)}div(S_2^*) \\ &\geq \frac{1}{4}f(S^*) + \lambda\ \frac{1}{2(3m-1)}div(S^*) \\ &\geq \frac{1}{2(3m-1)}\Big(f(S^*) + \lambda\ div(S^*)\Big) \end{aligned} \tag{4}$$

where the final steps follow from the fact that $f(S_1^*) \geq f(S^*)$ and $div(S_2^*) \geq div(S^*)$ which is implied by the optimality of S_1^* and S_2^* for the two parts of the objective along with the fact that the two quantities, $f(S^*)$ and $\lambda div(S^*)$, are both nonnegative and $\frac{1}{4} \geq \frac{1}{2(3m-1)}, \forall m \geq 2$. □

From the analysis above, we showed that we can come up with a simple solution for problem 1 that uses prior work. Moreover, it is easy to verify that the approximation factor for the Max-Min diversity is the dominant term in the analysis. Therefore, even if we were to use a more efficient algorithm of the submodular maximization literature [3, 7] with respect to the first part of the objective, the final approximation bound would not change.



(a) Data with computable optimal solution.

(b) Data with known optimal solution.

Figure 2: Approximation performance of the proposed algorithm with respect to the number of different groups m and the regularization term λ. (a) Dataset of size $N = 25$ with computable optimal solution and $k = 5$. (b) Dataset of size $N = 10,128$ with computable optimal solution and $k = 128$.

4 Experimental Evaluation

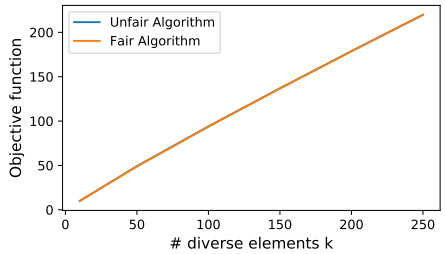
In this section, we evaluate the behavior of the proposed algorithmic framework in terms of the approximation factor using synthetic datasets with computable and known optimal solution respectively. Furthermore, we use the Adult dataset [5] to evaluate the *price of fairness*, namely how much the sacrifice in the objective function is so as to ensure fairness. Finally, we use the UTKFace dataset [13] so as to qualitatively compare the solution of the fairness oblivious algorithm with that produced by the fair algorithm.

Submodular functions: In the experiments below we use two different classes of submodular functions. In particular, for the approximation performance we use a *coverage function* which is monotone and nonnegative. As a motivating example for using a coverage function, consider a document collection $D = \bigcup_{i=1}^{m} D_i$ from m different news channels and a a set of different topics T (e.g politics, tech). Each document is associated with a subset $T' \subseteq T$ of the topics. A *fair* document collection $D' \subseteq D$ consists of $\langle k_1, k_2, \cdots, k_m \rangle$ documents that maximize the number of covered topics in T. At each step, the greedy algorithm 3 selects the document with the maximum marginal gain; namely the document which covers the maximum number of topics not seen by the previously selected documents so far.

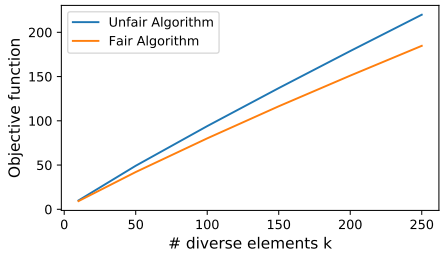
For the experiments on the *price of fairness* and the qualitative comparison, we use the function $f(S) = \sum_{s_i \in S} w(s_i)$, where $w(s_i)$ is a nonnegative score indicating the *utility* or *relevance* of an item in the result set. Notice that $f(S)$ is a special case of submodular functions for which the greedy algorithm is optimal even under fairness constraints. (e.g by sorting the elements in each group and selecting the k_i top results to the optimal solution). As a result, if were to use the modular function for the evaluation of the approximation factor the greedy algorithm would always the find the optimal solution for the submodular part of the objective in problem 1.

Real-world datasets: We use two real world datasets, the Adult dataset [5] and the UTKFace dataset [13]. The Adult dataset consists of 30,162 tuples with **no** missing attributes. For the experiments on the *price of fairness* we use only the continuous attributes (e.g age, capital-gain etc), normalize the data and select race as the sensitive attribute. Race defines $m = 5$ different groups: White (25, 933), Asian-Pac-Islander (895), Amer-Indian-Eskimo (286), Black (2,817) and Other (231). The UTKFace dataset consists of 23,708 RGB images of size 200×200 size and is frequently used for gender classification. Gender defines $m = 2$ different groups: Male (12,391), Female (11,317). We perform a dimensionality reduction using PCA to select the top $\ell = 330$ eigenfaces that accumulate 96% of the total variance of the data.

Approximation factor: We evaluate the approximation factor of the proposed approach using a synthetic dataset $X \sim \mathcal{N}(0, 1)$ with $N = 25$ items and computable optimal solution via exhaustive search. We use coverage as the submodular part of the objective and set $k = 5$. The *fairness* constraints are defined following a proportional representation rule; the sample preserves the balance ratio of the groups as observed in X. For readability, we report the inverse of the approximation factor, namely how much better the optimal solution is with respect to the solution retrieved by the algorithm. In figure 2 we give



(a) Proportional representation of the m demographic groups.



(b) Equal representation of the m demographic groups.

Figure 3: *Price of fairness* while k increases under: (a) Proportional representation of the $m = 5$ demographic groups present in the dataset. The sample of size k preserves the balance ratio observed in the original dataset. (b) Equal representation of the $m = 5$ demographic groups present in the dataset. The sample of size k contains the same number of elements per group.

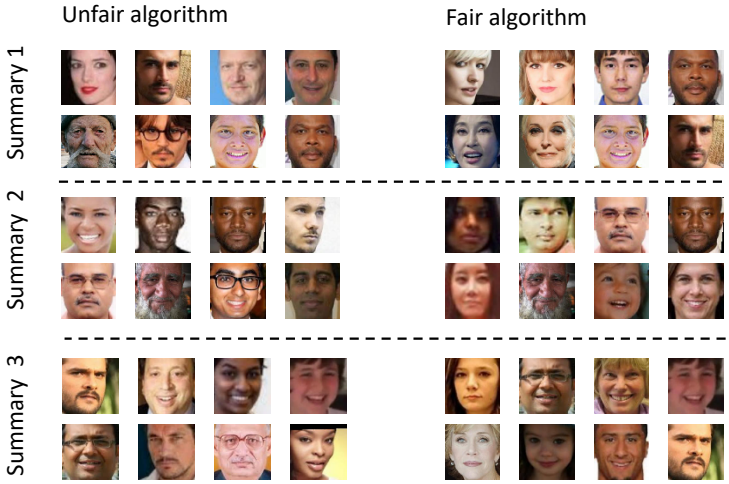


Figure 4: Qualitative comparison of three summaries produced by a fairness oblivious and a fair algorithm when the sensitive attribute is gender. Each summary is collected from a set of 16 images, with 8 male and 8 female faces, randomly selected from the UTKFace dataset [13]. We set $\langle k_1, k_2 \rangle = \langle 4, 4 \rangle$. We observe that the first set of summaries under-represent female faces while. The second set of summaries though equally represent genders while maintaining diversity.

a boxplot showing the approximation factors observed in #runs= 200 for different λ and m values. Observe that the smaller the regularization term λ is, the more significant the submodular part of the objective is. As a result, since the approximation guarantee is stronger for the submodular part, we observe that as λ increases, the approximation factor *worsens*. Nonetheless, we observe that the approximation performance of the algorithm is better in practice that the theoretical bound we derived. We also use a synthetic dataset with known optimal solution of size $N = 10,128$ items. We construct a set of $k = 128$ optimal elements as follows: (1) We assume there are 200 different topics and randomly assign a subset of at most 20 of the available topics to each of the k elements. The the optimal value for the submodular part of the objective is equal to the number of covered topics from the k optimal elements. We associate the rest of the elements only with elements already covered by the optimal solution so as to guarantee that there does not exist a better solution. (2) For diversity part of the objective, we assume that k optimal elements are the 'k' corners of a unit hypercube in the $D = \log_2 128 = 7$ dimensional space and the rest 10^4 points are generated by randomly selecting each of their dimensions to be in the $(0, 1)$ range. (so that all the other points lie inside the hypercube)

Price of Fairness: For this set of experiments, we use the Adult dataset and assign each tuple a weight equal to the 'age' attribute. We normalize the weights so as to be in the [0, 1] range. The rest of the continuous attibutes are used so as to compute the ℓ_2-distance of the tuples present in the data. The algorithm oblivious to fairness follows the technique introduced in [4]; namely we run GMM 2 to find a solution for the diversity part of the objective and use the maximum marginal gain algorithm ignoring the constraints to find a solution for the submodular part. Then among the two solutions we select the one that maximizes the value for the combined objective. From the experimental results shown in figure 3, we observe the price of fairness is negligible when the sample of size k preserves the balance in the original dataset. However, under equal representation the price of fairness is more prominent. Notice that $\lambda = 1$ for the experiments shown above.

Qualitative Results: We set $\lambda = 1$ and for each set of experiments shown if figure 4, we randomly select a set of 16 images with 8 female and 8 male faces from the UTFace dataset. Subsequently, we randomly assign each image a score in the $[0, 1]$ range and use ℓ_2-distance as our metric. Finally, we produce a summary with 4 images per gender of the 16 sampled images

using the fair algorithm we proposed and the algorithm oblivious to fairness proposed in [4]. Observe that the first set of summaries are often unbalanced with respect to gender since contain more male faces.

References

[1] Zeinab Abbassi, Vahab S. Mirrokni, and Mayur Thakur. Diversity maximization under matroid constraints. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 32–40, New York, NY, USA, 2013. ACM.

[2] Allan Borodin, Hyun Chul Lee, and Yuli Ye. Max-sum diversification, monotone submodular functions and dynamic updates. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '12, pages 155–166, New York, NY, USA, 2012. ACM.

[3] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a submodular set function subject to a matroid constraint (extended abstract). In Matteo Fischetti and David P. Williamson, editors, *Integer Programming and Combinatorial Optimization*, pages 182–196, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[4] Anirban Dasgupta, Ravi Kumar, and Sujith Ravi. Summarization through submodularity and dispersion. In *ACL (1)*, pages 1014–1022, 2013.

[5] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[6] Uriel Feige. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, July 1998.

[7] Y. Filmus and J. Ward. A tight combinatorial algorithm for submodular maximization subject to a matroid constraint. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 659–668, 2012.

[8] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey. *An analysis of approximations for maximizing submodular set functions—II*, pages 73–87. Springer Berlin Heidelberg, Berlin, Heidelberg, 1978.

[9] Andreas Krause and Daniel Golovin. Submodular function maximization. In *Tractability*, 2014.

[10] Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, page 510–520, USA, 2011. Association for Computational Linguistics.

[11] S. S. Ravi, D. J. Rosenkrantz, and G. K. Tayi. Heuristic and special case algorithms for dispersion problems. *Oper. Res.*, 42(2):299–310, April 1994.

[12] Arie Tamir. Obnoxious facility location on graphs. *SIAM J. Discrete Math.*, 4:550–567, 11 1991.

[13] Song Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.