Fair Diversification

Ashley Gao

December 4, 2023

Contents

1 Summary

I simply extend the result regarding part of Moumoulidou et al. [2020] from the metric space to an extended metric space, which is Fair Max-Min Diversification with $m = 2$ (i.e. Only two groups in the domain). Analysis of the approximation ratio and time complexity of the problem in extended metric space when applying Fair Swap Algorithm in Moumoulidou et al. [2020]. The results are: the time complexity remains $O(kn)$ and the approximation ratio increases to $\frac{c^2}{4}$, where c is constant regarding the property of the extended metric.
I will give the necessary definitions in Section 2 and give the formal proof in Section 3.

2 Definitions

Instead of the regular Fair Max-Min Diversification stated in Addanki et al. [2022] with metric space, I would give the reader a slightly different version of the definition, which adopts the extended metric space as the problem domain.

2.1 Fair Max-Min Diversification (Extended Metric)

Let $(\mathcal{X}, d)$ be an **extended** metric space where $\mathcal{X} = \bigcup_{i=1}^{m} \mathcal{X}_i$ is a universe of n elements positioned into m **non-overlapping** groups and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_0^+$ is a metric distance function.
We define the divergence of a set S: $div(S) = min_{u,v \in S, u \neq v} d(u, v)$.
Then we want to find a set S which satisfy the objective function $max_{S \subseteq \mathcal{X}} div(S)$

2.2 Extended Metric Space

For $(\mathcal{X}, d)$, it is an extended metric space with a factor $c \in (0, 2)$ (The selection of this range is proved and justified in Section 4) if it satisfies the following four properties:

1. (*identity*) $d(x, x) = 0, \ \forall x \in \mathcal{X}$
2. (*symmetry*) $d(x, y) = d(y, x), \ \forall x, y \in \mathcal{X}$
3. (*positive*) $d(x, y) \geq 0, \ \forall x, y \in \mathcal{X}$
4. (*triangle*) $c \cdot d(x, z) \leq d(x, y) + d(y, z), \ \forall x, y, z, \in \mathcal{X}$, where $c \in (0, 2)$

3 Proof for $m = 2$ settings for extended metric space

This proof is inspired by Moumoulidou et al. [2020].

Fair Swap Algorithm are attached in Figure 2. This algorithm is given in Moumoulidou et al. [2020], which is using the GMM algorithm (which is attached in Figure 1) that is proposed in Tamir [1991] as a building block.

Define the problem domain: I would like to show the proof of the approximation ratio and time complexity for the problem **Fair Max-Min Diversification** with the given scenario:

Algorithm 1 GMM Algorithm

Input: $\mathcal{U}$: Universe of available elements
$k \in \mathbb{Z}^+$
I: An initial set of elements

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

procedure GMM($\mathcal{U}, I, k$)
  $\mathcal{S} \leftarrow \emptyset$.
  if $I = \emptyset$ **then**
    $S \leftarrow$ a randomly chosen point in $\mathcal{U}$
  while $|\mathcal{S}| < k$ **do**
    $x \leftarrow \underset{u \in \mathcal{U}}{\text{argmax}} \min_{s \in \mathcal{S} \cup I} d(u, s)$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{x\}$
return $\mathcal{S}$

Figure 1: GMM Algorithm

Algorithm 2 FAIR-SWAP: Fair Diversification for $m = 2$

Input: $\mathcal{U}_1, \mathcal{U}_2$: Set of points of color 1 and 2
$k_1, k_2 \in \mathbb{Z}^+$

Output: k_i points in $\mathcal{U}_i$ for $i \in \{1, 2\}$

procedure FAIR-SWAP
▷Color-Blind Phase:
  $\mathcal{S} \leftarrow \text{GMM}(\mathcal{U}, \emptyset, k)$
  $\mathcal{S}_i = \mathcal{S} \cap \mathcal{U}_i$ for $i \in \{1, 2\}$
▷Balancing Phase:
  Set $U = \text{argmin}_i(k_i - |\mathcal{S}_i|)$ ▷Under-satisfied set
  $O = 3 - U$ ▷Over-satisfied set
  Compute:

$$E \leftarrow \text{GMM}(\mathcal{U}_U, \mathcal{S}_U, k_U - |\mathcal{S}_U|)$$
$$R \leftarrow \{\text{argmin}_{x \in \mathcal{S}_O} d(x, e) : e \in E\}$$

return $(\mathcal{S}_U \cup E) \cup (\mathcal{S}_O \setminus R)$

Figure 2: Fair Swap

- Two-group ($m = 2$)
- Non-overlapping
- $(\mathcal{X}, d)$ is the extended metric space with constant $c \in \mathbb{R}$
- Applying Fair Swap Algorithm (Algorithm 2)

Running Time Analysis. Since it is running the same algorithm as the non-extended version that has been proposed by Moumoulidou et al. [2020], the running time would follow the previous result, which is $O(kn)$

Approximation-Factor Analysis. Let's define the input and output.

- Input: A set of points $\mathcal{U} = \mathcal{U}_1 \cup \mathcal{U}_2$ and $k_1, k_2 \in \mathbb{R}^+$, with $k_i \leq |\mathcal{U}_i|, \forall i \in \{1, 2\}$
- Output: A subset of $\mathcal{U}$ with k_i points from each $\mathcal{U}_i$ partition such that the div(S) is maximized

Let's define the terms used in the proof.

- S^*: The set of k points in $\mathcal{U}$ that maximize the diversity when there are no fairness constraints, which means that we assume all the points in the same group and don't care fairness anymore.
- l^*: The divergence of the set S^*, which is: $l^* = div(S^*)$
- $\mathcal{F}^* = \mathcal{F}_1^* \cup \mathcal{F}_2^*$: Where $\mathcal{F}^*$ is the optimal solution for the Fair Max-Min diversification, and

$$\mathcal{F}_1^* = \mathcal{F}^* \cap \mathcal{U}_1, \qquad \mathcal{F}_2^* = \mathcal{F}^* \cap \mathcal{U}_2$$

- l^*_{fair}: The divergence of the set $\mathcal{F}^*$, which is: $l^*_{\text{fair}} = div(\mathcal{F}^*)$
- $S, S_1, S_2, S_U, S_O, E, R, U, O$: Adopt the same meaning that defined in the Algorithm.
  - In detail, S is the intermediate output of the algorithm with no fairness constraints(output for the Color-Blind Phase).

- S_1 is the points that both in S and group 1, while S_2 is the points that both in S and group 2.
- S_U is one of the S_1 or S_2 which under satisfy the constraint for the group (which means that we need to add more points from S_U to satisfy the fairness constraint), while S_O is the other set which over-satisfied.
- E is the intermediate output of the GMM algorithm. This algorithm randomly and greedily chooses the points that could be added to S_U and returns the adding set. R is the set of the points that would be removed from S_O, which the points are greedily selected by traversing all the possibilities.
- U is the index number for the unsatisfied group, and O is the index for the satisfied group.

I claim that the output set $(S_U \cup E) \cup (S_O \setminus R)$ is a $\frac{c^2}{4}$-approximation solution for the Max-Min Fair Diversification.

Proof. First, note that

$$l^* \geq l^*_{\text{fair}} \tag{1}$$

Since the l^* is the div of the solution in the groupless setting, then by applying more constraints, the div of the solution $\mathcal{F}^*$ would be worse, in other words, l^*_{fair} is smaller or equal than l^*.
Then note that

$$\forall i \in S, \quad \left| \{p \in S^* \mid d(p,i) < \frac{cl^*}{2}\} \right| \leq 1$$

This means that **for each** point i in S, there's **at most one point** p in S^* that satisfies: $d(p,i) < \frac{cl^*}{2}$, in other word, there's only one point in S^*, such that distance $< \frac{cl^*}{2}$ from each point of S. We can prove this by contradiction. Assume that there exist at least two points satisfy (2), then we have $\exists i \in S$, $\exists p, q \in S^*$, s.t. $d(p,i) < \frac{cl^*}{2}$ and $d(q,i) < \frac{cl^*}{2}$.
Then from triangle inequality and the previous two inequalities, which is

$$c \cdot d(p,q) \leq d(p,i) + d(q,i)$$

$$d(p,i) < \frac{cl^*}{2}, \quad d(q,i) < \frac{cl^*}{2}$$

We have

$$c \cdot d(p,q) \leq d(p,i) + d(q,i) < \frac{cl^*}{2} + \frac{cl^*}{2} = cl^* \implies d(p,q) < l^*$$

Since $l^* = div(S^*)$, which is the Min distance between points in S^*, but now we have points p, q with $d(p,q) < l^*$, which contradicts to the Minimum distance l^*. Then we've proved that for each point in S, there's at most one point in S^* such that $d(p,i) < \frac{cl^*}{2}$.
Therefore, while GMM has picked less than k elements, in the worst case, each point in S has 1 corresponding point in S^* such that the distance between them $< \frac{cl^*}{2}$ (and different points in S may correspond to the same point in S^* as well), then we can create a bijective mapping between the points in S and the points in S^*. By the Pigeon Hole Principle, since $|S| < |S^*| = k$, there must exist a good point that can be greedily selected and added to S, with distance $\geq \frac{cl^*}{2}$ from all the points already selected. Also with the fact that the algorithm greedily picks the next point, which is the point farthest away, we guarantee that the good

point that we mentioned would be chosen if there's no other better option. Then we naturally have

$$div(S) \geq \frac{cl^*}{2} \geq \frac{cl^*_{\text{fair}}}{2} \tag{2}$$

Since S_U is a subset of S, which has less point. Then we observe that

$$div(S_U) \geq div(S) \geq \frac{cl^*_{\text{fair}}}{2}$$

Then we look at set E, which is the set that include the points that would be added to S_U later. Similar as previous reasoning, when we are running GMM with parameter $(\mathcal{U}_U, S_U, k_U - |S_U|)$, there is at most one point in $\mathcal{F}^*_U$ that is distance $< \frac{cl^*_{\text{fair}}}{2}$ from each point in $S_U \cup E$. Formally,

$$\forall i \in S_U \cup E, \quad \left| \{p \in \mathcal{F}^*_U \mid d(p,i) < \frac{cl^*}{2}\} \right| \leq 1$$

Similarly, when GMM has picked less than $k - |S_U|$ elements, there exists at least one element that can be selected that with distance greater or equal than $\frac{cl^*_{\text{fair}}}{2}$ from the points already selected. Since the algorithm picks the next point farthest away from the points already chosen, the next point is at least $\frac{cl^*_{\text{fair}}}{2}$ from the existing points. Then we naturally have

$$div(S_U \cup E) \geq \frac{cl^*_{\text{fair}}}{2} \tag{3}$$

Our output is $(S_U \cup E) \cup (S_O \setminus R)$, from the (2) and (3), we already proved that $div(S)$ and $div(S_U \cup E)$ are greater or equal to $\frac{cl^*_{\text{fair}}}{2}$, the only thing that left for our proof is $div(E \cup S_O)$. For this case, by the algorithm's logic, we remove the closest point in S_O from E. Note that by application of the triangle inequality and the fact that $div(S_O) \geq \frac{cl^*_{\text{fair}}}{2}$, for each $x \in E$ there can be at most one point $y \in S_O$ such that $d(x,y) < \frac{c^2 l^*_{\text{fair}}}{4}$, and it is obvious that the size of E and R are the same. Hence, after the removal of the closest points the distance between all pairs is as required and we have the following Theorem 3.1. □

Theorem 3.1. *FairSwap(Algorithm2) is a $\frac{c^2}{4}$-approximation algorithm for the fair diversification problem with extended metric factor c when $m = 2$ that runs in time $O(kn)$*

4 Appendix

(Note: In the previous section, I use c as the extended metric space factor, while in this appendix, instead, α is considered as the factor for the metric space. They are the same thing, just using a different notation.)

Also, it is worth noticing that the value for **α-extended metric space**(i.e.Triangle inequality is $d(x,y) + d(y,z) \geq \alpha d(x,z)$) would satisfy $\alpha \in (0,2)$, since the problem becomes trivial for most of the settings when $\alpha = 2$, because if $\alpha = 2$, then it enforces that all non-zero distances become identical. While the maximum value for α to make sense is 2 as well.

Here is a check for the scenario that $\alpha = 2$, we have identical distance:

Proof. Assume that $\alpha = 2$, the distances between three points are a, b, c.

From this setup, we have $a+b \geq 2c,\ \ a+c \geq 2b,\ \ b+c \geq 2a$.
Then $a+b \geq 2c \geq 2(2b-a) \implies a+b \geq (4b-2a) \implies 3a \geq 3b$
Similarly, we get $3b \geq 3a$, then we have $a=b$. By symmetry, $a=b=c$. The distances in this metric are identical. □

Here is a simple proof for the statement to show why $\alpha = 2$ is the maximum value for α.

Proof. Let $\alpha \in \mathbb{R}^+$
We have

$$\begin{aligned} a+b \geq \alpha c,\ \ b+c \geq \alpha a,\ \ a+c \geq \alpha b \\ \implies a+b \geq \alpha(\alpha a - b) \\ \implies a+b \geq \alpha^2 a - \alpha b \\ \implies (1+\alpha)b \geq (\alpha^2-1)a \end{aligned}$$

We also have $(1+\alpha)a \geq (\alpha^2-1)b$
Then we have

$$\left(\frac{1+\alpha}{\alpha^2-1}a\right) \geq b$$
$$\left(\frac{1+\alpha}{\alpha^2-1}b\right) \geq a$$

We need to have $\frac{1+\alpha}{\alpha^2-1} \geq 1$ to make this possible, regardless of the value of a, b. Then

$$1+\alpha \geq \alpha^2 - 1 \implies \alpha \leq 2$$

Then we reach our conclusion, 2 is the maximum value for the α. □

References

R. Addanki, A. McGregor, A. Meliou, and Z. Moumoulidou. Improved approximation and scalability for fair max-min diversification, 2022.

Z. Moumoulidou, A. McGregor, and A. Meliou. Diverse data selection under fairness constraints, 2020.

A. Tamir. Obnoxious facility location on graphs. *SIAM Journal on Discrete Mathematics*, 4(4):550–567, 1991. doi: 10.1137/0404048. URL `https://doi.org/10.1137/0404048`.