The Two-Stage Approach for Solving Fair Clustering Problems
How to solve fair clustering problems?

We are looking for algorithms with *theoretical guarantees*:

1-Clustering Objective:

2-The Fairness Constraint:
How to solve fair clustering problems?

- We are looking for algorithms with **theoretical guarantees**:

1- **Clustering Objective**:

\[ D = \min_{S,\varphi} \sum_{j \in C} d^2(j, \varphi(j)) \rightarrow \tilde{D} \leq \alpha D \quad (\alpha > 1, \text{recall NP-hardness}) \]

2- **The Fairness Constraint**:

\[
\begin{align*}
    l_{\text{blue}} |C_i| &\leq |C_i^{\text{blue}}| \leq u_{\text{blue}} |C_i| \\
    l_{\text{red}} |C_i| &\leq |C_i^{\text{red}}| \leq u_{\text{red}} |C_i|
\end{align*}
\]

\[
\begin{align*}
    (l_{\text{blue}} |C_i|) - \Delta &\leq |C_i^{\text{blue}}| \leq (u_{\text{blue}} |C_i|) + \Delta \\
    (l_{\text{red}} |C_i|) - \Delta &\leq |C_i^{\text{red}}| \leq (u_{\text{red}} |C_i|) + \Delta
\end{align*}
\]

- relax by \( \Delta > 0 \)
How to solve fair clustering problems?

- We are looking for algorithms with **theoretical guarantees** over:

  1- **Clustering Objective:** \( D = \min_{S,\varphi} \sum_{j \in C} d^2(j, \varphi(j)) \rightarrow \hat{D} \leq aD \) (\( a > 1 \), recall NP-hardness)

  2- **Fairness Constraint:** \( l_{\text{blue}} |C_i| \leq |C_{i \text{blue}}| \leq u_{\text{blue}} |C_i| \rightarrow (l_{\text{blue}} |C_i|) - \Delta \leq |C_{i \text{blue}}| \leq (u_{\text{blue}} |C_i|) + \Delta \)

  \( l_{\text{red}} |C_i| \leq |C_{i \text{red}}| \leq u_{\text{red}} |C_i| \rightarrow (l_{\text{red}} |C_i|) - \Delta \leq |C_{i \text{red}}| \leq (u_{\text{red}} |C_i|) + \Delta \)

- There is **NOT** a single approach to solve all fair variants.

  Unsurprising: Fair Clustering \( \subset \) Constrained Clustering,

  No generic approach to solve Constrained Clustering for different constraints.

- Even the same problem maybe solved using different algorithms, e.g. Algorithm \( \mathcal{A}_1 \) has higher clustering quality than \( \mathcal{A}_2 \), but \( \mathcal{A}_2 \) has faster run time.

- For the k-(center,median, means): A simple approach with many applications → The two-stage approach.
Two-Stage Approach

- **Step 1 (Open Centers):** Use a fairness-agnostic clustering algorithm → this gives a collection of centers $S$.

- **Step 2 (Post-processing):** process the clustering to satisfy the fairness constraint at a bounded increase to the clustering cost (often that means carefully routing the points to the centers mostly using LP methods).
Two-Stage Approach: Group Fairness Example

⚠️ Recall Group (demographic) Fairness

\[
\min \sum_{i=1}^{k} \sum_{j \in C_i} d(j, \mu_i)
\]

Agnostic Clustering

\[
\min \sum_{i=1}^{k} \sum_{j \in C_i} d(j, \mu_i) \quad \text{s.t.} \quad l_{\text{blue}} |C_i| \leq |C_i^{\text{blue}}| \leq u_{\text{blue}} |C_i|
\]

\[
l_{\text{red}} |C_i| \leq |C_i^{\text{red}}| \leq u_{\text{red}} |C_i|
\]

Group Fair Clustering
Two-Stage Approach: Group Fairness Example

Given Instance:
Two-Stage Approach: Group Fairness Example

Given Instance:

Step 1

$c_1$

$c_2$
Two-Stage Approach: Group Fairness Example

- Centers are now **open!**
- How to assign points to centers??
Two-Stage Approach: Group Fairness Example

- Centers are now open!
- How to assign points to centers??
  Cost minimizing assignment is unfair (clusters don’t mix colors)
Two-Stage Approach: Group Fairness Example

- How to assign points to centers??

**Step 2** Route points so as to **minimize clustering cost**
subject to satisfying **color-proportional (fairness)** → Setup an integer program

**Integer Program:**

\[
\min \sum_{i \in S} \sum_{j \in C} d(i, j) x_{ij}
\]

\(x_{ij} \in \{0, 1\}\)

0-1 decision variable

\(\sum_{i \in S} x_{ij} = x_{1j} + x_{2j} = 1\)

point must be assigned to some center

\[l_{\text{blue}} \left( \sum_{j \in C} x_{1j} \right) \leq \sum_{j \in C} p_j^{\text{blue}} x_{1j} \leq u_{\text{blue}} \left( \sum_{j \in C} x_{1j} \right)\]

\[l_{\text{red}} \left( \sum_{j \in C} x_{1j} \right) \leq \sum_{j \in C} p_j^{\text{red}} x_{1j} \leq u_{\text{red}} \left( \sum_{j \in C} x_{1j} \right)\]

\[l_{\text{blue}} \left( \sum_{j \in C} x_{2j} \right) \leq \sum_{j \in C} p_j^{\text{blue}} x_{2j} \leq u_{\text{blue}} \left( \sum_{j \in C} x_{2j} \right)\]

\[l_{\text{red}} \left( \sum_{j \in C} x_{2j} \right) \leq \sum_{j \in C} p_j^{\text{red}} x_{2j} \leq u_{\text{red}} \left( \sum_{j \in C} x_{2j} \right)\]
Two-Stage Approach: Group Fairness Example

How to assign points to centers??

(Step 2) Route points so as to minimize clustering cost subject to satisfying color-proportional (fairness) → Setup an integer program

Integer Program:

\[
\begin{align*}
\min_{x_{ij}} & \sum_{i \in S} \sum_{j \in C} d(i, j)x_{ij} \\
x_{ij} & \in \{0,1\} & \text{0-1 decision variable} \\
\sum_i x_{ij} = x_{1j} + x_{2j} & = 1 & \text{point must be assigned to some center} \\
\end{align*}
\]

\[
\begin{align*}
\ell_{\text{blue}} \left( \sum_{j \in C} x_{1j} \right) & \leq \sum_{j \in C} p_j^{\text{blue}} x_{1j} \leq u_{\text{blue}} \left( \sum_{j \in C} x_{1j} \right) \\
\ell_{\text{red}} \left( \sum_{j \in C} x_{1j} \right) & \leq \sum_{j \in C} p_j^{\text{red}} x_{1j} \leq u_{\text{red}} \left( \sum_{j \in C} x_{1j} \right) \\
\ell_{\text{blue}} \left( \sum_{j \in C} x_{2j} \right) & \leq \sum_{j \in C} p_j^{\text{blue}} x_{2j} \leq u_{\text{blue}} \left( \sum_{j \in C} x_{2j} \right) \\
\ell_{\text{red}} \left( \sum_{j \in C} x_{2j} \right) & \leq \sum_{j \in C} p_j^{\text{red}} x_{2j} \leq u_{\text{red}} \left( \sum_{j \in C} x_{2j} \right)
\end{align*}
\]

Integer Programs Generally Take Exponential Time!
Two-Stage Approach: Group Fairness Example

How to assign points to centers??

(Step 2) Route points so as to minimize clustering cost

subject to satisfying color-proportional (fairness) → Setup an integer program → Relax to LP

Linear Program:

\[
\min_{x_{ij}} \sum_{i \in S} \sum_{j \in C} d(i, j) x_{ij}
\]

\[x_{ij} \in \{0, 1\} \quad x_{ij} \in [0, 1]\]

\[\sum_{i \in S} x_{ij} = x_{1j} + x_{2j} = 1 \quad \text{point must be assigned to some center}\]

\[
l_{blue}(\sum_{j \in C} x_{1j}) \leq \sum_{j \in C} p_{j}^{blue} x_{1j} \leq u_{blue}(\sum_{j \in C} x_{1j})
\]

\[
l_{red}(\sum_{j \in C} x_{1j}) \leq \sum_{j \in C} p_{j}^{red} x_{1j} \leq u_{red}(\sum_{j \in C} x_{1j})
\]

\[
l_{blue}(\sum_{j \in C} x_{2j}) \leq \sum_{j \in C} p_{j}^{blue} x_{2j} \leq u_{blue}(\sum_{j \in C} x_{2j})
\]

\[
l_{red}(\sum_{j \in C} x_{2j}) \leq \sum_{j \in C} p_{j}^{red} x_{2j} \leq u_{red}(\sum_{j \in C} x_{2j})
\]
Two-Stage Approach: Group Fairness Example

- Resulting solution $x_{ij}$ is possibly fractional (not 0 or 1)
Two-Stage Approach: Group Fairness Example

Resulting solution $x_{ij}$ is possibly fractional (not 0 or 1)

Applying a rounding technique
Two-Stage Approach: Group Fairness Example

- Resulting solution $x_{ij}$ is possibly fractional (not 0 or 1)
  - Applying a rounding technique

- Choice of rounding technique is non-trivial and often the most difficult step.
Two-Stage Approach

- Previous was for demographic fairness [Bera et al 2019; Bercea et al 2019; Esmaeili 2020].

- Other post processing approaches:
  - Combinatorial approach [Chakrabarti et al, AISTATS 2022]
  - Randomized approach [Brubach et al, ICML 2020]
THANK YOU!