Fair Matroid Monotone Submodular Maximization

\[
\max_{S \in \mathcal{I}, |e| \leq u_c} \sum_{c=1}^{C} (f(S) + f(V_c)) \quad \text{(FMMSM)}
\]

Set function \( f : 2^V \rightarrow \mathbb{R} \) is
- **Monotone:** for any set \( X \) and element \( e \),
  \( f(X \cup \{e\}) \geq f(X) \);
- **Submodular:** for any sets \( X \subseteq Y \) and element \( e \)
  \( f(X \cup \{e\}) - f(X) \geq f(Y \cup \{e\}) - f(Y) \);

Matroid \( \mathcal{I} \subseteq 2^V \) of rank \( k \):
- Non-empty family of sets satisfying
  - **Downward-closedness:** if \( A \subseteq B \) and \( B \in \mathcal{I} \), then \( A \in \mathcal{I} \);
  - **Augmentation:** if \( A, B \in \mathcal{I} \) with \( |A| < |B| \), then \( A + e \in \mathcal{I} \) for some \( e \in B \).

Examples: uniform matroid \( |S| \leq k \), partition matroid \( S \cap V_c \leq u_c \).

Streaming setting: Elements arrive on a stream and we have limited memory.

Fairness: Solution should be balanced with respect to some sensitive attribute.
- Each element has a color \( c \) encoding the sensitive attribute.
- \( V \) is partitioned into \( C \) disjoint color groups \( V_c \).
- We are given a lower bound \( \ell_c \) and upper bound \( u_c \) (not constants) on the number of elements we can pick from each color \( c \).

Applications: multiwinner voting, influence maximization, data summarization

Related work

- Special case of FMMSM with cardinality constraint:
  - Celis et al. [2018]: tight (1 - 1/e)-approximation in centralised setting.
  - El Halabi et al. [2020]: one-pass streaming algorithms with
    - tight 1/2-approximation with exponential in \( k \) memory
    - 1/4-approximation with \( O(k) \) memory

Monotone submodular maximization over two matroid constraints:
- Garg et al. [2021]: 1/5.828-approximation one-pass streaming algorithm with \( O(k) \) memory.

Greedy-Fair-Streaming: a one-pass heuristic algorithm based on the first pass of our two-pass algorithm

Our Results

**Theorem 1.1.** For any constant \( \eta \in (0, 1/2) \), there exists a one-pass streaming \((1/2 - \eta)\)-approximation algorithm for FMMSM that uses \( 2^{O(k^{1+1/\eta})} \cdot k \cdot \log C \) memory, where
\[
\Delta = \min_{x \in \mathcal{I}} f(\{c\}) - f(\emptyset).
\]

**What if we want to Use Less Memory?**

It is not possible to use efficient memory even if we make multiple passes.

If we violate the lower bounds we can get a high solution with quadratic memory usage in two passes over the stream.

Even with more violations, it is not possible to get efficient algorithms.

Theorem 1.2. Any \((\log C)\)-pass streaming algorithm that determines the existence of a feasible solution for FMMSM with probability at least 2/3 requires \( \Theta(k, C)^{2^{O(1)}} \) memory.

Empirical Results

**Theorem 1.3.** There exists a two-pass streaming algorithm for FMMSM that runs in polynomial time, uses \( O(k \cdot C) \) memory, and outputs a set \( S \) such that:
1. \( S \) is independent,
2. \( |S|/2 \leq |S| \leq u_c \) for any color \( c \), and
3. \( f(S) \geq \Omega(T) \).

**Theorem 1.4.** There is no one-pass semi-streaming algorithm that determines the existence of a feasible solution for FMMSM with probability at least 2/3, even if it is allowed to violate the fairness lower bounds by a factor of 2 and completely ignore the fairness upper bounds.

References