Mode Finding for SLC Distributions via Regularized Submodular Maximization

Ehsan Kazemi[1,*], Shervin Minaee[2], Moran Feldman[3] and Amin Karbasi[1]

[1]Yale University, [2]Snap Inc., and [3]University of Haifa

[*]Now at Google, Zürich, Switzerland

Regularized Submodular Maximization

- Consider the following problem:

$$S^* = \arg\max_{S \subseteq \mathcal{N}, |S| \le k} f(S), \text{ where } f(\cdot) \triangleq g(\cdot) - \ell(\cdot)$$

utility: $g(\cdot)$; Cost / Penalty: $\ell(\cdot)$

Submodularity

$$\forall A \subseteq B \subseteq \mathcal{N} \text{ and } e \in \mathcal{N} \setminus B$$

$$f(A \cup \{e\}) - f(A) \ge f(B \cup \{e\}) - f(B)$$

Monotonicity

$$\forall A \subseteq B \subseteq \mathcal{N}$$

$$f(A) \le f(B)$$

- If f is
  1. non-negative
  2. monotone
  3. submodular

  Greedy yields $(1 - e^{-1})$-approximation

- Given a general submodular function f
  Testing whether there exists S such that $f(S) > 0$ is NP-hard!

  We need further assumptions on the objective f!

- A more tractable problem:

$$S^* = \arg\max_{S \subseteq \mathcal{N}, |S| \le k} f(S) \text{ where } f(\cdot) \triangleq g(\cdot) - \ell(\cdot)$$

Non-negative Monotone: $g(\cdot)$; Non-negative Modular: $\ell(\cdot)$

- Many practical scenarios
  - The data arrives at a very fast pace
  - There is only time to read the data once
  - No random access
  - On massive data the greedy policies take a few days/weeks to complete

Is it possible to summarize a massive data set "on the fly"?

Can we parallelize the greedy approach?

Regularized Submodular Maximization

- For a positive real value r

$$h(r) = \frac{2r + 1 - \sqrt{4r^2 + 1}}{2}$$

- We define:

$$T \in \underset{S \subseteq \mathcal{N}, |S| \le k}{\arg\max} \left[(h(r) - \varepsilon) \cdot g(T) - r \cdot \ell(T) \right]$$

Algorithm 1: THRESHOLD-STREAMING

Guess a value τ such that
$k\tau \le h(r) \cdot g(T) - r \cdot \ell(T) \le (1 + \varepsilon)k\tau$.
Let $\alpha(r) \leftarrow \frac{2r+1+\sqrt{4r^2+1}}{2}$.
Let $S \leftarrow \varnothing$.
while $|S| < k$ *and there are more elements* **do**
  Let u be the next elements in the stream.
  if $g(u \mid S) - \alpha(r) \cdot \ell(\{u\}) \ge \tau$ **then**
    Add u to the set S.
return *the better solution among* S *and* $\varnothing$.

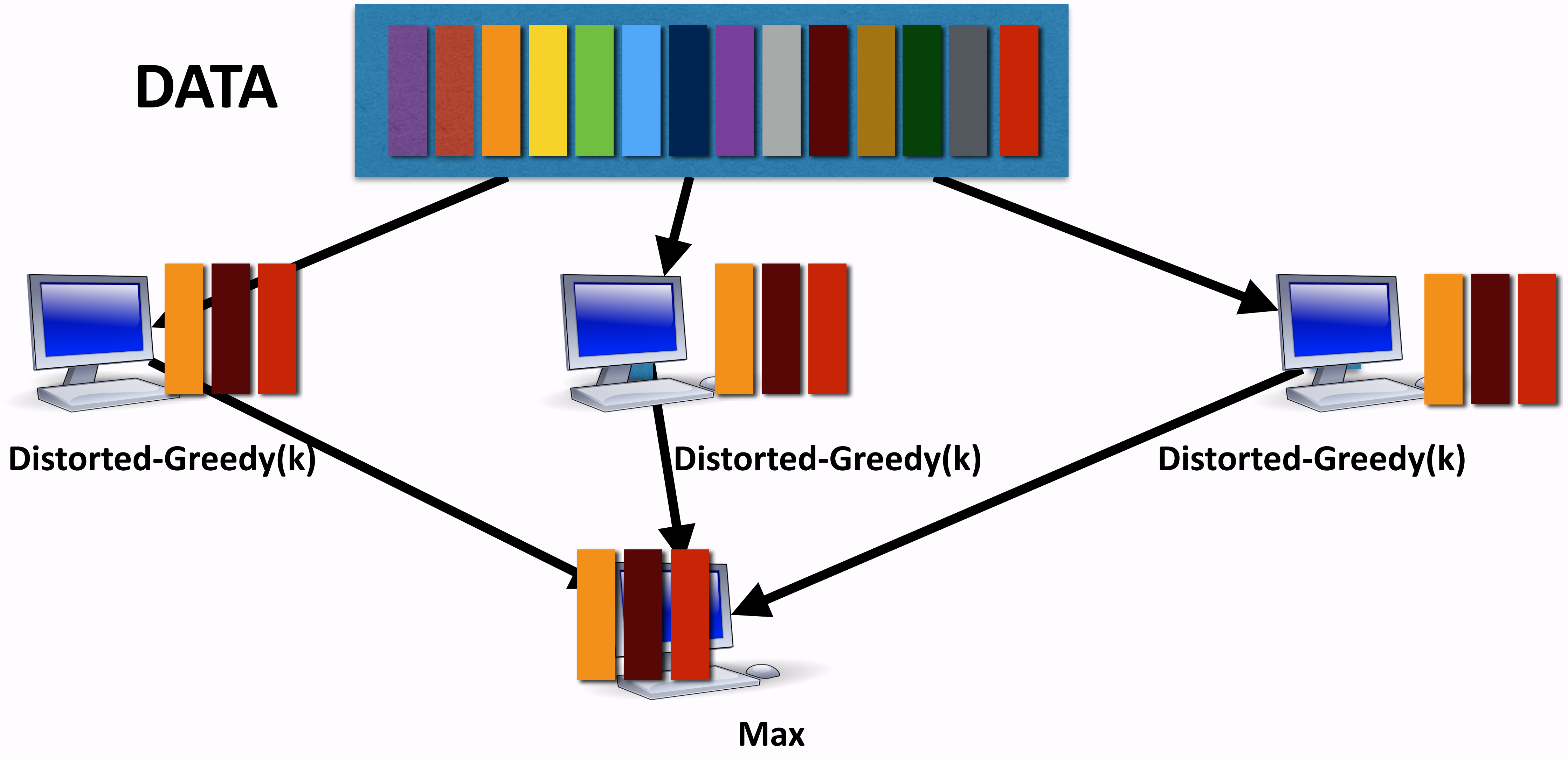
Max

[Kazemi, Minaee, Feldman, Karbasi]

For every $\epsilon, r > 0$, THRESHOLD-STREAMING produces a set $S \subseteq \mathcal{N}$ of size at most k such that

$$g(S) - \ell(S) \ge \max_{T \subseteq \mathcal{N}, |T| \le k} [h(r) - \varepsilon) \cdot g(T) - r \cdot \ell(T)]$$

- Multi-Stage Distributed Algorithm



[Kazemi, Minaee, Feldman, Karbasi]

MultiStage-DISTRIBUTED-GREEDY returns a set $D \subseteq \mathcal{N}$ of size at most k after $O(1/\varepsilon)$ iterations such that

$$\frac{\mathbb{E}[g(D) - \ell(D)]}{1 - \varepsilon} \ge (1 - e^{-1}) \cdot g(OPT) - \ell(OPT)$$

Does not require to keep multiple copies of the data.
It improves the state-of-the-art for monotone-submodular functions.

Mode Finding for SLC Distributions

- A distribution is strongly log-concave if its generating polynomial is SLC
  - Strong negative dependence among sampling items
  - Many examples of SLC distributions:
    - Determinantal point processes
    - The uniform distribution on the independent sets of a matroid
  - SLC does not imply log-submodularity [Gotovos, 2019]

- γ-additively weak submodularity for SLC functions [Robinson et al. (2019)]:

A set function $\rho : 2^{\mathcal{N}} \to \mathbb{R}$ is γ -additively weak submodular if for any $S \subseteq \mathcal{N}$ and $u, v \in \mathcal{N} \setminus S$ with $u \neq v$, we have:

$$\rho(S) + \rho(S \cup \{u, v\}) \leq \gamma + \rho(S \cup \{u\}) + \rho(S \cup \{v\})$$

For a γ-additively weak submodular function ρ, the following function is submodular:

$$\Lambda(S) \triangleq \rho(S) - \frac{\gamma}{2} \cdot |S| \cdot (|S| - 1)$$

$$\ell(S) = \sum_{u \in S} \ell_u, \text{ where } \ell_u \triangleq \max\{\Lambda(\mathcal{N} \backslash u) - \Lambda(\mathcal{N}), 0\} = \max\{\rho(\mathcal{N} \backslash u) - \rho(\mathcal{N}) + \gamma \cdot (|\mathcal{N}| - 1), 0\}$$

$g(S) \triangleq \Lambda(S) + \ell(S)$ is monotone and submodular.

Use our algorithms to maximize $\Lambda(\cdot)$

[Kazemi, Minaee, Feldman, Karbasi]

THRESHOLD-STREAMING returns a solution R such that

$$\rho(R) \geq (h(r) - \varepsilon)\rho(OPT) - (\alpha(r) - r - 1 + \varepsilon)\ell(OPT) - \frac{\gamma[(h(r) - \varepsilon)l(l-1) - |R| \cdot (|R| - 1)]}{2}$$

[Kazemi, Minaee, Feldman, Karbasi]

MultiStage-DISTRIBUTED-GREEDY returns a solution R such that

$$\mathbb{E}[\rho(R)] \geq (1 - \varepsilon)\left[\left(1 - e^{-1}\right)\rho(\mathrm{OPT}) - e^{-1}\ell(\mathrm{OPT})\right] - \frac{\gamma\left[\left(1 - e^{-1}\right)l(l-1) - \mathbb{E}[|R|(|R| - 1)]\right]}{2}$$