Regularized Submodular Maximization at Scale

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

[1]Google Zürich, [2]Snap Inc., [3]University of Haifa, [4]Yale University

Regularized Submodular Maximization

- Consider the following problem:

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

Regularized Submodular Maximization

- Consider the following problem:

utility

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

Regularized Submodular Maximization

- Consider the following problem:

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

utility: $g(\cdot)$

Cost / Penalty: $\ell(\cdot)$

Regularized Submodular Maximization

- Consider the following problem:

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

utility: $g(\cdot)$

Cost / Penalty: $\ell(\cdot)$

Submodularity

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

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

Regularized Submodular Maximization

- Consider the following problem:

utility

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

Cost / Penalty

Submodularity

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

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

Monotonicity

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

$$f(A) \leq f(B)$$

Regularized Submodular Maximization

- Consider the following problem:

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

utility: $g(\cdot)$

Cost / Penalty: $\ell(\cdot)$

Submodularity

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

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

Monotonicity

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

$$f(A) \leq f(B)$$

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

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

Regularized Submodular Maximization

- Consider the following problem:

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

utility: $g(\cdot)$

Cost / Penalty: $\ell(\cdot)$

Submodularity

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

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

Monotonicity

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

$$f(A) \leq f(B)$$

- 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!

Regularized Submodular Maximization

- Consider the following problem:

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

utility ($g(\cdot)$) — Cost / Penalty ($\ell(\cdot)$)

Submodularity

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

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

Monotonicity

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

$$f(A) \leq f(B)$$

- 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!

Regularized Submodular Maximization

- Consider the following problem:

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

Regularized Submodular Maximization

- Consider the following problem:

Non-negative Monotone

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

Regularized Submodular Maximization

- Consider the following problem:

Non-negative Monotone

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

Non-negative Modular

Regularized Submodular Maximization

- Consider the following problem:

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

Non-negative Monotone (g)

Non-negative Modular (ℓ)

- Prior Work:
  - [Sviridenko et al., 2017]:
    - Algorithm based on continuous extensions
    - Need to guess the cost of the optimal solution
  - [Feldman, 2018]:
    - Removed the guessing step
  - [Harshaw et al., 2019]:
    - Distorted-greedy: an efficient algorithm
    - Extend it to the case of weakly submodular functions

Regularized Submodular Maximization

- Consider the following problem:

Non-negative Monotone

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

Non-negative Modular

- Prior Work:
  - [Sviridenko et al., 2017]:
    - Algorithm based on continuous extensions
    - Need to guess the cost of the optimal solution
  - [Feldman, 2018]:
    - Removed the guessing step
  - [Harshaw et al., 2019]:
    - Distorted-greedy: an efficient algorithm
    - Extend it to the case of weakly submodular functions

- 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

Regularized Submodular Maximization

- Consider the following problem:

Non-negative Monotone

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

Non-negative Modular

- Prior Work:
  - [Svi
    - o
    - o
  - [Feldman, 2018]:
    - Removed the guessing step
  - [Harshaw et al., 2019]:
    - Distorted-greedy: an efficient algorithm
    - Extend it to the case of weakly submodular functions

- Many practical scenarios
  - e
  - 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"?

Regularized Submodular Maximization

- Consider the following problem:

Non-negative Monotone

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

Non-negative Modular

- Prior Work:
  - [Svi
    - o
    - o
  - [Feldman, 2018]:
    - o Removed the
  - [Harshaw et al., 20
    - o Distorted-greedy: an efficient algorithm
    - o Extend it to the case of weakly submodular functions
- Many practical scenarios
  - No random access
  - eedy policies take a few

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

Can we parallelize the greedy approach?

THRESHOLD-STREAMING

[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

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

THRESHOLD-STREAMING

[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

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

For $r = 1$ we have $h(r) = \phi^{-2} \approx 0.382$

THRESHOLD-STREAMING

[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

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

For $r = 1$ we have $h(r) = \phi^{-2} \approx 0.382$

$$\beta_S = \frac{g(S) - \ell(S)}{\ell(S)}$$

$$r = r_{OPT} = \frac{\beta_{OPT}}{2\sqrt{1 + 2\beta_{OPT}}}$$

[Kazemi, Minaee, Feldman, Karbasi]

THRESHOLD-STREAMING produces a solution such that

$$g(S) - \ell(S) \geq \left(\frac{1 + \beta_{OPT} - \sqrt{1 + 2\beta_{OPT}}}{2\beta_{OPT}} - \varepsilon' \right) \cdot (g(OPT) - \ell(OPT))$$

THRESHOLD-STREAMING

[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

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

For $r = 1$ we have $h(r) = \phi^{-2} \approx 0.382$

$$\beta_S = \frac{g(S) - \ell(S)}{\ell(S)}$$

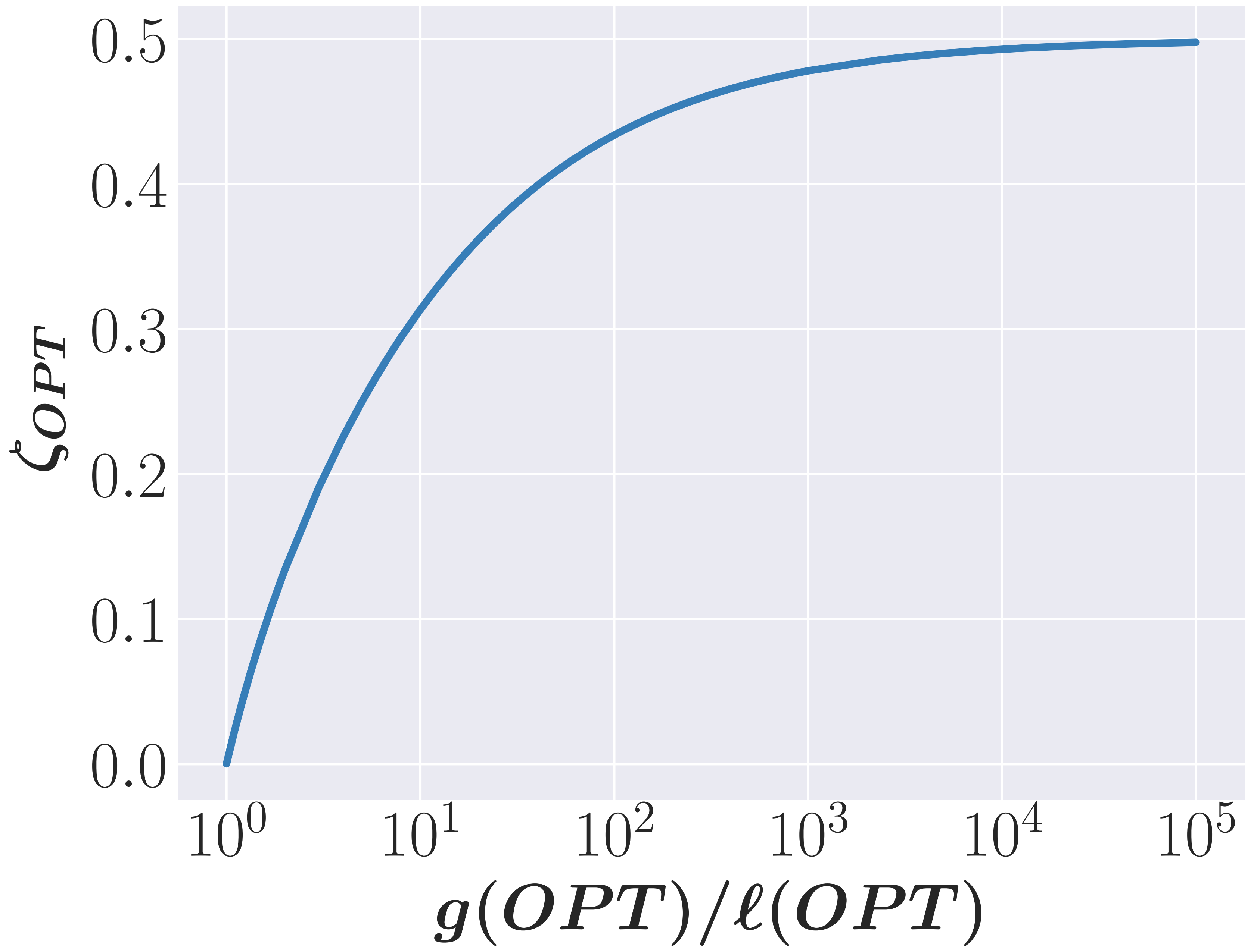
$$r = r_{OPT} = \frac{\beta_{OPT}}{2\sqrt{1 + 2\beta_{OPT}}}$$

[Kazemi, Minaee, Feldman, Karbasi]

THRESHOLD-STREAMING produces a solution such that

$$g(S) - \ell(S) \geq \left(\frac{1 + \beta_{OPT} - \sqrt{1 + 2\beta_{OPT}}}{2\beta_{OPT}} - \varepsilon' \right) \cdot (g(OPT) - \ell(OPT))$$

Approximation Factor as a Function of $g(OPT)/\ell(OPT)$



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

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

Multi-Stage Distributed Algorithm

[Kazemi, Minaee, Feldman, Karbasi]

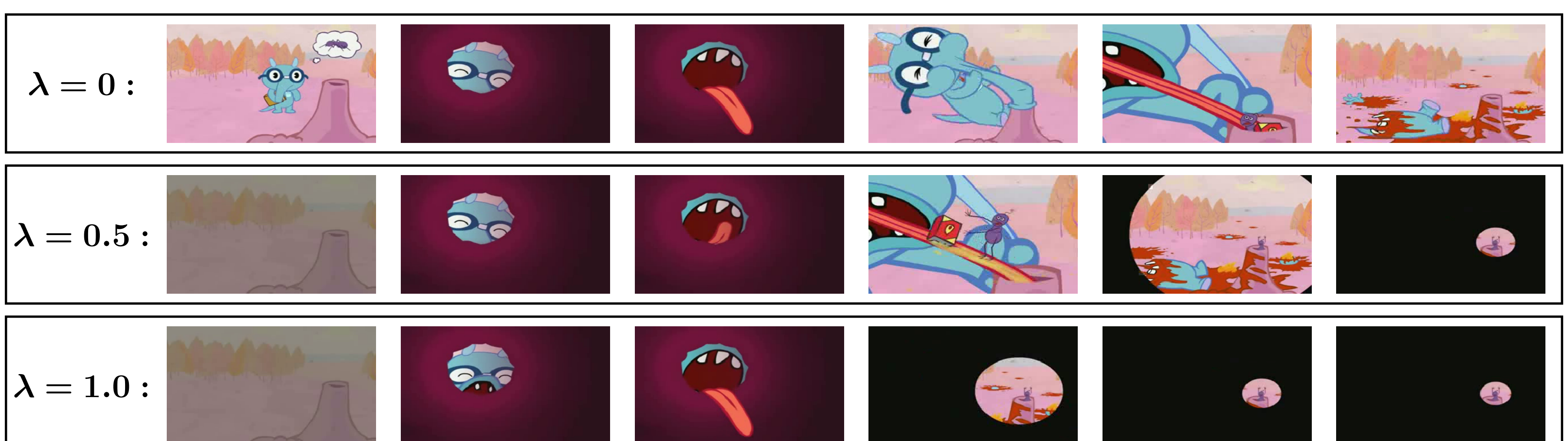
MultiStage-DISTRIBUTED-GREEDY returns a set $D \subseteq \mathcal{N}$ of size at most k after $O(1/\varepsilon)$ iterations such that

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

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

Applications

- Mode Finding for SLC Distributions
  - Strong negative dependence among sampling items
  - Many examples of SLC distributions:
    - Determinantal point processes
    - The uniform distribution on the independent sets of a matroid
- Vertex cover of social networks
- Data summarization
  - Video, location and text summarization



Thank

You !