# Beyond 1/2-Approximation for Submodular Maximization on Massive Data Streams

Ashkan Norouzi-Fard, Jakub Tarnawski, Slobodan Mitrović, Amir Zandieh, **Aida Mousavifar**, Ola Svensson

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

## Submodularity

- A set function $f: 2^V \to \mathbb{R}$ with diminishing return property

*ground set* $V = \{$  $\}$

$f($  $)$ = cost, utility, …

$$\forall\, A \subseteq B \subseteq V, e \notin B$$

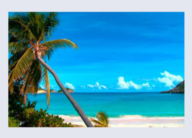

$$f(A \cup \{e\}) - f(A) \;\geq\; f(B \cup \{e\}) - f(B)$$

## Problem

- Extract a small, representative subset from a big data set

$$S^* = \arg \max_{|S| \leq k} f(S)$$

- We assume $f$ is submodular, monotone, and $f(\emptyset) = 0$



## Related Works

- Greedy:
  - Add $e = \arg\max f(e|S)$
  - $k$-passes
  - $f(S) \geq \mathrm{OPT}(1 - 1/e)$

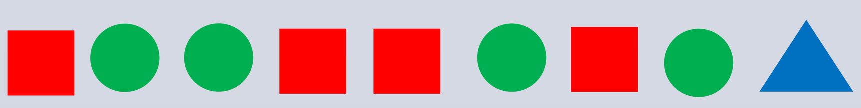
- SIEVE –STREAMING:
  - Add e if $f(e|S) \geq \frac{\mathrm{OPT}/2 - f(S)}{k - |S|}$
  - 1-pass
  - $f(S) \geq (0.5 - \epsilon)\,\mathrm{OPT}$

## Beyond 0.5 Ratio

**Theorem:** Any algorithm for streaming submodular maximization that only queries the value of the submodular function on feasible sets (sets of cardinality at most $k$) and is $>$ 0.5-approximation must use $\Omega(n/k)$ memory.

- Reduction from INDEX problem

$$f(S) = |S \cap U| + \begin{cases} \min(k, |S \cap V|) & w \notin S \\ k & w \in S \end{cases}$$

$\mathrm{OPT} = 2k - 1$ 

## Random Streams

- In many real-world scenarios the data arrives in random order.

**Theorem:** There exists an algorithm (SALSA) such that, for any stream of elements that arrive in random order, the value of the solution returned by SALSA is $\geq$ OPT $(0.5 + \epsilon)$ in expectation and uses $O(k \log k)$ memory.

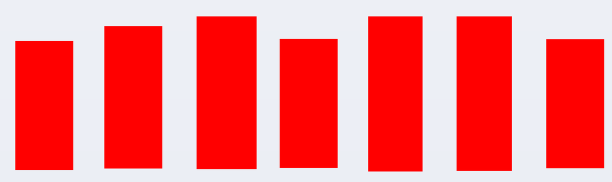
## SALSA Algorithm

- Adaptive thresholds

| $\beta\%$ | $(1-\beta)\%$ |
|---|---|
| $\geq^? T_1 \cdot \frac{OPT}{k}$ | $\geq^? T_2 \cdot \frac{OPT}{k}$ |

- Structure of OPT

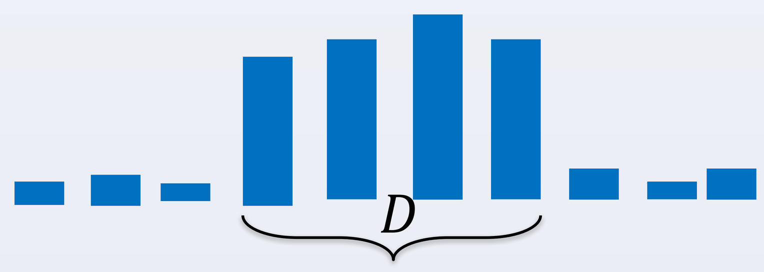
Balanced



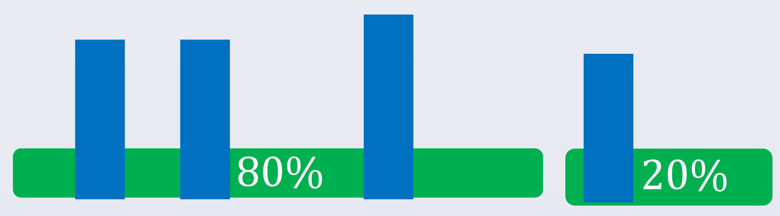
| 10% | 90% |
|---|---|
| $\geq^? (0.5+\epsilon) \cdot \frac{OPT}{k}$ | $\geq^? T_2 \cdot \frac{OPT}{k}$ |

- $f(S_L)$ is large $\geq \mathrm{OPT}(0.5 - \epsilon)$
  Decrease $T_2$
- $S_L$ contains $\geq \frac{k}{100}$ elements of OPT
  Keep $T_2 = T_1$

Dense




$$|D| \leq \frac{k}{100}$$
$$f(D) \geq \frac{OPT}{10}$$



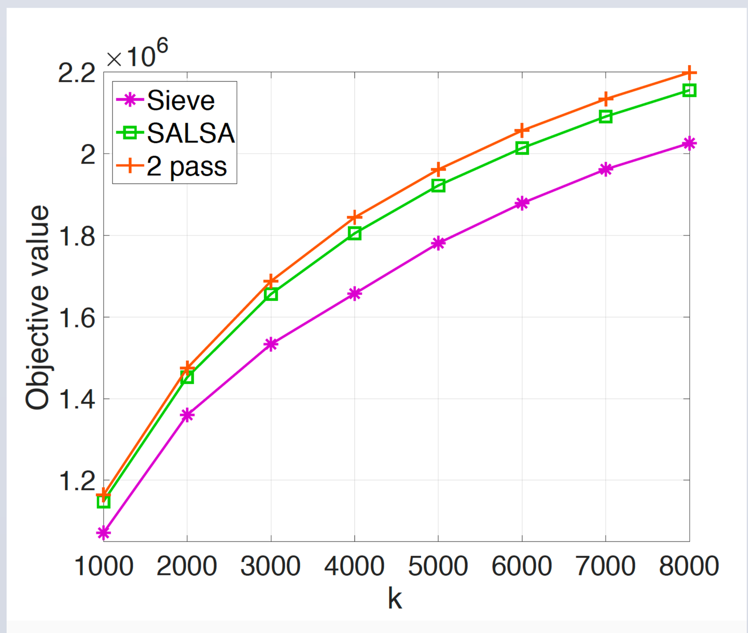

$\geq^? 2\frac{OPT}{k}$ $\qquad$ $\geq^? \frac{OPT}{10k}$

## Multiple-Pass

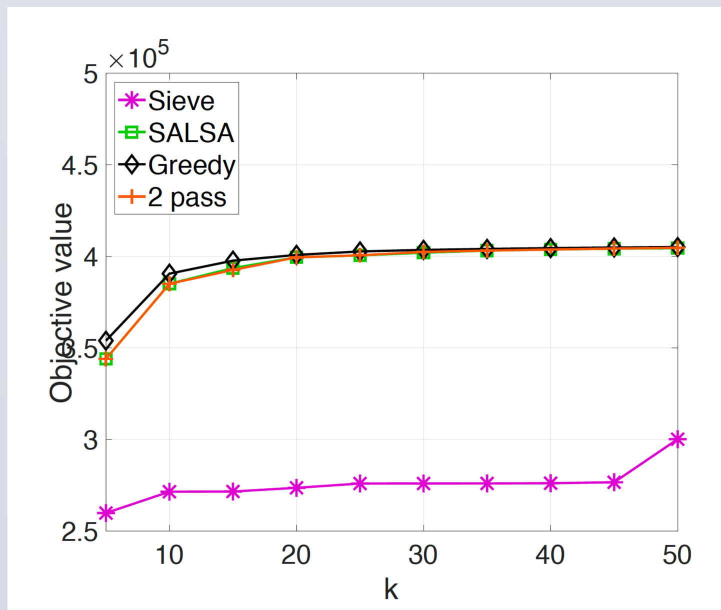
**Theorem:** There exists a $(1 - 1/e - \epsilon)$–approximation algorithm that uses $O(1/\epsilon)$ passes for the streaming submodular maximization. It uses $O(k \log k/\epsilon)$ memory.

- $p$-pass
  - Add e if $f(e|S) \geq \left(\frac{p}{p+1}\right)^i \cdot \frac{OPT}{k}$
  - $p$-pass
  - $f(S) \geq \mathrm{OPT} \cdot \left(1 - \left(\frac{p}{p+1}\right)^i\right)$

## Experiments




Maximum Coverage




Exemplar-based Clustering

## Open Problems

- What is the best achievable bound in random streams?
- Hardness result under no assumption ?

## References

1. Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An analysis of approximations for maximizing submodular set functionsi. Mathematical Programming , 14(1):265–294, 1978.
2. Badanidiyuru, A., Mirzasoleiman, B., Karbasi, A., and Krause, A. Streaming submodular maximization: Massive data summarization on the fly. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, pp. 671–680, New York, NY, USA, 2014. ACM.

RESEARCH POSTER PRESENTATION DESIGN © 2019
www.PosterPresentations.com