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

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

Submodularity

- A set function $f: 2^V \rightarrow \mathbb{R}$ with diminishing return property
- ground set $V = \{\text{food, utility, ...}\}$
- $f(A \cup \{e\}) - f(A) \geq f(B \cup \{e\}) - f(B)$

Problem

- Extract a small, representative subset from a big data set $S^* = \arg\max_{S \subseteq V} f(S)$
- We assume $f$ is submodular, monotone, and $f(\emptyset) = 0$

Related Works

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

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
  - $f(S) = |S \cap U| + \left\{ \min\{k, |S \cap V|\} \right\}_{w \in S}^{w \notin S}$
  - $OPT = 2k - 1$

Random Streams

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

SALSA Algorithm

Adaptive thresholds

- $\beta_p = (1 - T_1 - \epsilon) OPT$ if $f(S) \geq \beta_p$
- $(1 - \beta_p)$ otherwise

Structure of OPT

Balanced

- $f(S) \geq \beta_p$ if $S$ is large
- $S_k$ contains at least $k$ elements of OPT
- Keep $T_2 = T_1$

Dense

- $f(S) \geq (0.5 + \epsilon) OPT$
- Decrease $T_2$

Multiple-Pass

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

Experiments

- Maximum Coverage
- Exemplar-based Clustering

Open Problems

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

References