Adaptively Robust Resettable Streaming

Edith Cohen[1, 3], Elena Gribelyuk[2, 1], Jelani Nelson[4, 1], Uri Stemmer[3, 1]

[1] Google Research, [2] Princeton University, [3] Tel Aviv University, [4] UC Berkeley

ICML International Conference On Machine Learning

Resettable Streaming Model

- **Input:** updates to an underlying frequency vector $x \in \mathbb{Z}^n$, which arrive sequentially one at a time (*worst-case*, fixed in advance).
  - Insert(i, Δ): $x_i \leftarrow x_i + \Delta$
  - Reset(i): $x_i \leftarrow 0$
- **Output:** At the end of the stream, A outputs an approximation of a given function of the stream.
- **Goal:** A should use space *sublinear* in the length T of the update sequence and universe size n.
- **Motivation:** right to be forgotten, machine unlearning

Frequency Moments

We consider the F_p **moment estimation** problem for $p \in [0,1]$:

- Given a stream of T updates to coordinates $i \in [n]$, let x_i denote the frequency of element i.

$$F_p = \sum_{i \in [n]} |x_i|^p$$

- We let $F^{(t)}$ denote the value of the relevant statistic on the *prefix* of the stream until time t. Let $\varepsilon \in (0,1)$ be a fixed accuracy parameter.
- By reduction from *set-disjointness*, any algorithm which achieves $(1+\varepsilon)$ relative error (w.p. $2/3$) requires $\Omega(n)$ space.
- **Goal:** At each step $t \in T$, output an estimate $\hat{F}^{(t)}$ such that $\left|\hat{F}^{(t)} - F^{(t)}\right| \leq \varepsilon \cdot \max_{t' \leq t} F^{(t')}$ (*prefix-max* error guarantee)

Adversarially Robust Streaming

- **Input:** updates (Insert or Reset), which arrive sequentially and *adversarially.*
- **Output:** At each time t, A receives an update u_t, updates its internal state, and returns a *current estimate* r_t , which is recorded by the *adversary*.

"Future updates may depend on previous estimates"

Previous Approaches

- **Sketch Switching** [BJWY20]: $\tilde{O}(T)$ space
- **Differential Privacy Robustification** [HKMMS20]: $\tilde{O}(\sqrt{T})$ space
- All known algorithms rely on persistent randomness, which can be learned by the adaptive adversary.

Question: can we design adaptively robust sketches for resettable streaming that return a *prefix-max* estimate to fundamental statistics using $\text{poly}(\frac{1}{\varepsilon}, \log\frac{1}{\delta}, \log T)$ bits of space?

Main Contribution

Theorem (Robust resettable cardinality, sum, and Bernstein; informal): For any $\varepsilon, \delta \in (0,1)$ and a resettable adaptive stream with T_{Inc} insertions, there exist sketches for cardinality, sum, and Bernstein statistics of size $k = \text{poly}\left(\frac{1}{\varepsilon}, \log\left(\frac{T_{Inc}}{\delta}\right)\right)$ bits. With probability at least $1-\delta$, the sketch maintains an estimate $\widehat{F^{(t)}}$ satisfying $\left|\hat{F}^{(t)} - F^{(t)}\right| \leq \varepsilon \cdot \max_{t' \leq t} F^{(t')}$.

- As a corollary of our construction, our sketch for distinct elements estimation (F_0) is also event-level differentially private.

Overview of our Approach

We proceed by robustifying the following classical algorithm for distinct elements (F_0) [GLH06]:

- Let $p \in (0,1)$ be a fixed sampling probability, let $S = \emptyset$. For each update in the stream:
  - If Insert(i):
    - $S \leftarrow S \setminus \{i\}$
    - Sample $b \sim$ Bernoulli(p).
    - If $b = 1$, $S \leftarrow S \cup \{i\}$.
  - If Reset(i):
    - Set $S \leftarrow S \setminus \{i\}$.
  - Return $\widehat{N_t} = \frac{|S|}{p}$
- A modification of the above algorithm (by adaptively adjusting the sampling rate) admits a streaming algorithm for resettable F_0 satisfying the *prefix-max* guarantee.

Robustification for F_0 (High-Level Idea):

- We compose the standard fixed-rate sketch & estimator with the *Binary Tree Mechanism*: the input to the mechanism is a sequence of updates to the *sample-size*, and the output is a sequence of noisy estimates of the prefix-sum ($\tilde{S}_t$). The estimator returns $\widetilde{N_t} = \frac{\tilde{S}_t}{p}$ (protect key membership in sample).

Robustification for F_1 (High-Level Idea):

- We examine the F_1 (sum) algorithm of [CCD12] and carefully identify the sensitive privacy units for robustness.
- Decompose the estimator into a *protected* (low-sensitivity) component and a *deterministic* component, which can be known to the adversary; we then feed updates to the *protected* part of the estimator to the *Binary Tree Mechanism*.
- Technical challenge: Unlike our robust F_0 algorithm, each private unit contributes to many distinct updates to the private portion of the estimator.