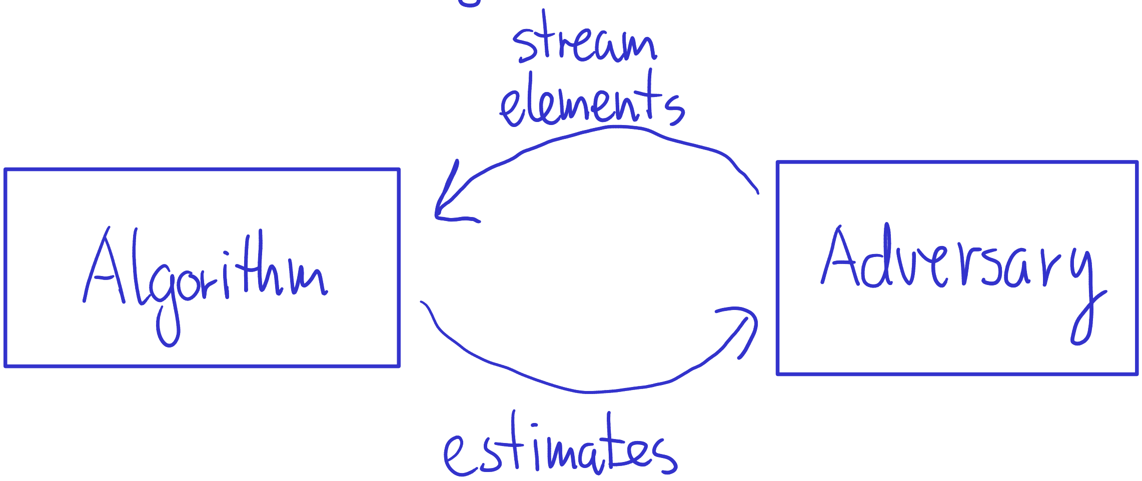
DS-563/CS-543 2024-02-15

Today: Continue adversarially robust streaming algorithms

- Quick model review
- Continue distinct elements estimation for insertion only
- Advertise a talk of interest (see Piazza)
- Distinct elements for insertion/deletion streams

next time

Model: streaming but stream is not fixed



Each round:

- Adversary sends next stream element
- Algorithm sends back the current solution/estimate

Who wins?

- Algorithm: if all solutions/estimates are good
- Adversary: if at least one solution/estimate is bad

Specific problem: distinct elements / F_0

Want $(1+\varepsilon)$-multiplicative approximation

$$(1-\varepsilon) F_0 \leq \hat{F} \leq (1+\varepsilon) F_0$$

---

non-robust algorithm = works for any fixed stream with good probability

robust algorithm = works in the adaptive model described here with good probability

Goal: non-robust $\Rightarrow$ robust

Technique 1: "Sketch / algorithm switching" (Insertion only)

A = non-robust F_0 algorithm, $(1+\varepsilon/20)$-multiplicative approximation w.p. $1-\delta/m^2$

$\varepsilon, \delta \in (0, 1/2)$ — arbitrary parameters selected by the user

m - upper bound on the stream length

Robust algorithm

estimate $\leftarrow 0$
index $\leftarrow 1$
$t = O(\varepsilon^{-1} \log m)$ independent copies $A_1, A_2, \ldots, A_t$ of A
for each stream item x:
  pass x to each A_i and process it
  if estimate from $A_{\text{index}} \geq (1 + \varepsilon/2)$ estimate:
    estimate $\leftarrow$ estimate from A index
    index $\leftarrow$ index $+1$
  output estimate

Observation 1 (space usage):
If all estimates from A_i's that we look at are good, then $O(\varepsilon^{-1} \log m)$ copies of A suffices

Proof last time: $F_0 \in \{0, 1, 2, \ldots, m\}$,
so estimate can increase at most
$3 + \log_{1+\varepsilon/2} m = O(\varepsilon^{-1} \log m)$ times

Observation 2 (correct estimates)
If all estimates from A_i's that we look at are good, then all output estimates are a good approximation.

Proof:

- Consider setting estimate to some value. Why is it going to be a $(1+\varepsilon)$-multiplicative approximation until increased?
- Initially, it is a $(1+\varepsilon/20)$-multiplicative approximation.
- So initially, we may overestimate by a factor of $(1+\varepsilon/20)$, but because F_0 can only increase (due to insertion-only stream), we will never overestimate by more.
- How much can we underestimate? We know that $\underbrace{F_0(1-\varepsilon/20)}_{\text{max underestimate}} < (1+\varepsilon/2)\,\text{estimate}$ holds initially + later if estimate not increased

$$\text{Hence, estimate} > \frac{1-\varepsilon/20}{1+\varepsilon/2} F_0 \geq (1-\varepsilon/20)(1-\varepsilon/2) F_0$$
$$\geq \left(1 - \frac{\varepsilon}{20} - \frac{\varepsilon}{2}\right) F_0 \geq (1-\varepsilon) F_0$$

- So estimate is a $(1+\varepsilon)$-multiplicative approximation ∎

Remains to show:

All estimates from A_i's that we look at are good (i.e., $(1+\varepsilon/20)$-mult. approx.) with probability $\geq 1-\delta$

Proof:

1. Without loss of generality: Assume the Adversary is deterministic. If a randomized Adversary can make one of the estimates to be bad with probability $> \delta$, then by an averaging argument, the randomness in the Adversary can be fixed to achieve the same with probability $> \delta$.

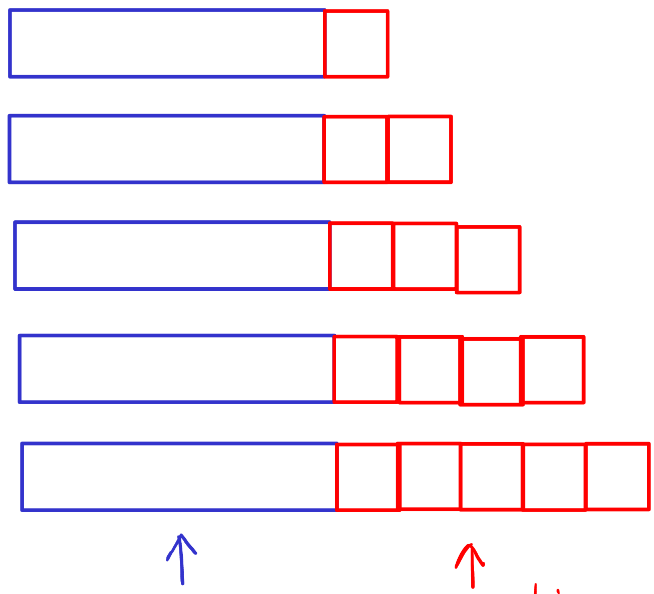
2. Proof by induction on k:
   The estimates we look at from A_i s.t. $i \in \{1, \ldots, k\}$ are all good w.p. $\geq 1 - \frac{k\delta}{m}$.
   - Base case of $k = 0$: trivially true
   - Inductive step: Suppose true for k.
     Want to show for $k+1$.

   Upper bound the probability that an estimate from A_{k+1} is bad.

   After estimate is set to an estimate from A_k (or 0 if $k=0$), the Adversary is given this value until it gets an estimate from A_{k+1}, at which point we stop using A_{k+1}.

Thought experiment: Keep giving the estimate from $\mathcal{A}_k$ to the Adversary

This produces up to m fixed streams, extensions of the stream so far



previous stream

consecutive extensions by the deterministic Adversary while still fed the last estimate from $\mathcal{A}_k$

It suffices that $\mathcal{A}_{k+1}$ gives good estimates on all these streams.

By the Union Bound, the probability $\mathcal{A}_{k+1}$ gives a bad estimate on one of them is at most $m \cdot \frac{\delta}{m^2} = \frac{\delta}{m}$.

Again by the Union Bound, the probability that we get bad estimate from one of the first $k+1$ $\mathcal{A}_i$'s is at most

$$\frac{k\delta}{m} + \frac{\delta}{m} = \frac{(k+1)\delta}{m}.$$

End of inductive step & proof.

3. Since the number of A_i's used is bounded by m, all estimates from A_i's that we look at are good with probability at least

$$1 - \frac{m \cdot \delta}{m} = 1 - \delta$$

$\blacksquare$

Hence it suffices to keep $O(\varepsilon^{-1} \log m)$ copies of A. To get A, which works with probability $1 - \frac{\delta}{m^2}$, from the F_0 algorithm in class, take the median of $O(\log(m/\delta))$ estimates. (See HW1 Q3 for details.)

Hence a $\mathrm{polylog}(m)$ factor overhead suffices for fixed ε & δ.