CS369G: Algorithmic Techniques for Big Data Spring 2015-2016

Lecture 6: Max-Stability and Misra-Gries

Prof. Moses Charikar *Scribes: Akshay Agrawal, Yifan Lu*

1 Overview

In this lecture, we bring our discussion of frequency moment estimation to a close by completing our analysis of Andoni's max-stability algorithm [1] for moments with $p > 2$. We will end the lecture by introducing a new topic, that of frequent item estimation, and presenting the classic Misra-Gries streaming algorithm [2] for the heavy hitters problem.

2 Analysis: $p > 2$ Frequency Moments via Max-Stability

Last time, we looked at a linear sketch scheme to estimate the ℓ_p norm of a vector for $p > 2$ that took $\tilde{O}(n^{1-\frac{2}{p}})$ space. Recall that the algorithm, due to Andoni [1], consists of a two-step mapping. The input $x \in \mathbb{R}^n$ is first transformed to a vector $y \in \mathbb{R}^n$ by scaling each entry x_i of x by a value $u_i^{-1/p}$, where $u_i \sim \text{Exp}(1)$. That is, $y_i = x_i u_i^{-1/p}$. We then reduce y to an m-dimensional vector z such that $z_j = \sum_{i:h(i)=j} \sigma(i) \cdot y_i$, where $h : [n] \to [m]$ and $\sigma : [n] \to \{\pm 1\}$ are hash functions. Our final estimate of $\|x\|_p$ is $\|z\|_\infty$.

For notational convenience, let $M = \|x\|_p$. We proved in the prequel that $\|y\|_\infty \in [1/2M, 2M]$ w.p. ≥ 0.75. To complete the analysis of the max-stability estimator, we now bound the probability that the infinity norm of z deviates from that of y. We do this in two steps, explained informally here and expounded upon in the following two subsections. In the first step, we show that the number of the y_i that are large is sufficiently small such that, with high probability, these large y_i do not collide when they are mapped to z. In the second step, we demonstrate that the net contributions of the remaining small coordinates of y are themselves small.

2.1 Bounding the probability of large element collisions in z

We formalize the notion of large y_i with the following definition and prove a bound on the number of such coordinates.

Definition 1. *y_i is large if $|y_i| > \dfrac{M}{c \log n}$.*

Claim 1. *For $\ell \geq 1$, there are at most ℓ^p indices i s.t. $|y_i| > \dfrac{M}{\ell}$ (in expectation).*

Proof.

$$\begin{aligned}\Pr\left[y_i > \frac{M}{l}\right] &= \Pr\left[x_i u_i^{-1/p} > \frac{M}{\ell}\right] \\ &= \Pr\left[u_i < \frac{x_i^p \ell^p}{M^p}\right] \\ &= 1 - e^{-x_i^p \ell^p / M^p} \\ &\leq \frac{x_i^p \ell^p}{M^p}\end{aligned}$$

where the last equality uses the fact that $e^x \geq 1 + x$ for all x.

Thus, if Y is number of large coordinates (and recalling that M is defined as $\|x\|_p$),

$$\mathbb{E}[Y] = \sum_i \Pr[y_i > \frac{M}{\ell}] \leq \ell^p \sum_i \frac{x_i^p}{M^p} = \ell^p$$

□

In our definition of large coordinates, taking $\ell = c \log n$; the above claim tells us that we have $O(\log^p n)$ large coordinates. By the birthday paradox, we expect to encounter collisions when the number of elements is on the order of $\sqrt{m}$. Since $\log^p n \ll O(n^{1-\frac{2}{p}} \log n) = m$ for $p > 2$, the probability that any two large coordinates in y collide when mapped to z is small, as was desired.

Exercise: Prove the last assertion.

2.2 Bounding the net contributions of small elements

Let $s = \left\{ i \,\middle|\, y_i < \frac{M}{c \log n} \right\}$ be the set of small elements and $Z'_j = \sum_{i \in s : h(i) = j} \sigma(i) \cdot y_i$ the net total they contribute to the j-th entry of z. We carry out some computations to find the variance of Z'_j and bound it with standard inequalities to show that it is (substantially) smaller than $\|x\|_p$. Clearly, $\mathbb{E}[Z'_j] = 0$. So

$$\begin{aligned}\mathrm{Var}[Z'_j] = \mathbb{E}[(Z'_j)^2] &= \mathbb{E}\left[\left(\sum_{i \in s : h(i) = j} \sigma(i) \cdot y_i\right)^2\right] \\ &= \mathbb{E}\left[\sum_{i \in s : h(i) = j} y_i^2\right] && \text{(the cross terms equal zero)} \\ &= \frac{\sum_{i \in s} y_i^2}{m} \\ &\leq \frac{\|y\|_2^2}{m}\end{aligned}$$

In order to interpret our variance in a meaningful way, we want to relate $\|y\|_2$ to $M = \|x\|_p$.

Claim 2. $\|y\|_2^2 \leq n^{1-\frac{2}{p}}\|x\|_p^2$

Proof. Consider $p = 2$. $\mathbb{E}[y_i^2] = x_i^2\mathbb{E}[u_i^{-2/p}] = O(x_i^2)$, since the expectation of an exponential random variable is constant. It follows that $\mathbb{E}[\|y\|_2^2] = O(\|x\|_2^2)$.

We now invoke Hölder's inequality, which states that, given $f, g \in \mathbb{R}^n$, $\sum_i f_i g_i \leq \|f\|_a\|g\|_b$, where $1/a + 1/b = 1$, $f_i = x_i^2$, and $g_i = 1$. Choosing $a = p/2$ and $b = 1/(1-2/p)$ so that $\sum_i f_i g_i = \|x\|_2^2$, Hölder's tells us that $\|x\|_2^2 \leq n^{1-2/p}\|x\|_p^2$. □

The above claim implies that $\mathrm{Var}[Z_j'] \leq \dfrac{n^{1-\frac{2}{p}} \cdot M^2}{m}$. To bound the probability that the sum of the Z_j''s is small, we will need an inequality stronger than Chebyshev; in particular, we will use Bernstein's inequality, which derives from the machinery of Chernoff bounds.

Theorem 1 (Bernstein)**.** *Given $x_1, \ldots, x_n$ (independent) with $\mathbb{E}[x_i] = 0$ and $|x_i| \leq Q$,*

$$\Pr\left[\sum_{x_i} > t\right] \leq \exp\left(\frac{-t^2/2}{\sum_i \mathbb{E}[x_i^2] + \frac{1}{3}Qt}\right).$$

We can apply Bernstein to bound the probability that $|Z_j'|$ exceeds M:

$$\begin{aligned}\Pr[|Z_j'| > \alpha M] &\leq \exp\left(\frac{(\alpha M)^2}{\frac{n^{1-\frac{2}{p}}M^2}{m} + \frac{1}{3}\frac{M}{c\log n}(\alpha M)}\right) \\ &\leq \exp(-c'\log n)\end{aligned}$$

To bound the probability that the sum of the Z_j''s exceed M, we can simply apply the union bound over all indices j. (Note that we can control c' as desired.)

2.3 Discussion

- Our proof gives us a constant approximation. Can we get a $(1+\epsilon)$-approximation? We can do it by making m suitably large and using multiple copies of the process and taking the median.
- We need n exponential random variables. How do we do sample from them in a streaming fashion? Nisan's construction [3] works, but other methods work as well (tabulation based hashing, for example).
- The sketch itself is linear, even though the final estimator is not.

3 Frequent Items/Heavy Hitters

We'll now discuss the problem of identifying elements that occur frequently in a data stream. Let's start with a simple instantiation of this problem: if we know that there is one item that occurs at least half the time, how do we find it with one pass?

Here's a possible construction: We keep one element and one counter, initialized to zero. For each element, if the counter is zero we replace the stored element with the current one and we set the counter to one. If we see the element again, we increment the counter. Otherwise, we decrement the counter.

Why does this work? For the sake of analysis, we reformulate the algorithm such that instead of storing an element and a counter, we have a stack. An increment of the counter corresponds to pushing the current element onto the stack. A decrement corresponds to popping from the stack. It is simple to show that this formulation is functionally equivalent.

Every time you pop off the stack, there are two elements of interest: the one you popped and the one you used to cause the pop. The pop cannot happen more than $m/2$ times. This means the element that occurs more than half the time cannot be popped and matched that many times, so it must be at the top of the stack in the end.

The Misra-Gries algorithm [2] generalizes this to the *heavy hitters* problem, in which we must report elements i such that $f_i \geq \frac{m}{k}$ in a stream of m elements: We have k bins each with an element and a counter. When we see a new element, j, if j belongs to some bin, we increase the counter for that bin. Otherwise, we add $(j, 1)$ to list of bins. Finally, if there exist k bins already, decrease all counters by 1 and remove any bin with counter 0. If $\hat{f}_i$ is the frequency reported by the algorithm of i. $\hat{f}_i = 0$ for all i not in memory and $f_i - \frac{m}{k} \leq \hat{f}_i \leq f_i$. If $m = \|f\|_1$ then the error is $\frac{\|f\|_1}{k}$ with $k = \frac{1}{\epsilon}$.

While Misra-Gries is, of course, a streaming algorithm, it does not produce a sketch and its output is therefore not composable. In the next lecture, we will look at a data structure that can produce a sketch for heavy hitters.

References

[1] A. Andoni, High frequency moments via max-stability, 2013.

[2] J. Misra & D. Gries, Finding repeated elements, *Cornell University Technical Report*, 1982.

[3] N. Nisan, Pseudorandom generators for space-bounded computation" *Combinatorica*, 12.4 (1992): 449-461.