COMS E6998-9: Algorithms for Massive Data (Spring'19) Feb 5, 2019

Lecture 5: High freq. moments, precision sampling, dim. reduction

Instructor: *Alex Andoni* Scribes: *Dave Epstein, Jaewan Bahk*

1 Improving CountMin

Definition 1. *CountMin for detecting heavy hitters.*

$$\forall j \in [L], k \in [w] : \; H_j[k] = \sum_{i:h_j(i)=k} f_i$$

where $f \in \mathbb{R}^n$ is a frequency vector, $h_j : [n] \to [w]$ is a random hash function. We estimate $\hat{f}_i = \min_j H_j[h_j(i)]$. To find ϕ-heavy-hitters (ϕ-HH), we return any item with $\hat{f}_i > \phi||f||_1$.

That is, we pick items that hash to buckets with a value large enough to qualify as HH. However, we ask ourselves if we can do better than this CountMin in the following setting for finding ϕ-HH where most frequencies are far smaller than that of the one ϕ-HH:

$$\phi = n^{-0.1}; f = (1, 1, \ldots, 2\phi n, 1, \ldots, 1)$$

CountMin uses space $O(\frac{1}{\phi\epsilon} \ln n) = O(\frac{\ln n}{\epsilon} n^{0.1})$. So, can we do better in this regime? Consider the following sketch for a counter:

Definition 2. *Alternate counter.*

$$\forall j \in [l], k \in [w] : \; H_j[k] = \sum_{i:h_j(i)=k} r_{ji} f_i$$

where $f \in \mathbb{R}^n$ is a frequency vector, $h_j : [n] \to [w]$ is a random hash function, and $r_{ji} =_r \{\pm 1\}$.

Intuitively, we want smaller values of f_i to cancel each other out, or for additional items that are *not* the heavy-hitter but hash to the same bucket as it to have a minimal effect. Indeed, with analysis similar to that of the length of a random walk, we expect the sum of all low-frequency elements hashing to the same bucket $|\sum_{i:h_j(i)=k, f_i=1} r_{ji}| \approx \sqrt{n/w}$, a significant improvement over the n/k we'd expect without the introduction of the random sign.

In this counting sketch, we can no longer rely on the naive median or minimum of values in the hash tables due to the introduction of the possibly negative r.v., so our new estimator becomes $\hat{f}_i = \text{median}_j \; r_{ji} H_j[h_j(i)]$.

2 Approximating Frequency Moments F_p where $p \notin \{0, 1, 2, \infty\}$

We have devised algorithms for F_0 (count distinct, Flajolet-Martin and bottom-k), F_1 (count, Morris), F_2 (using Tug-of-War), and now for F_∞ using heavy hitters. We observe that the heavy hitters algorithm

is harder in terms of runtime complexity ($\Theta(n)$) than the zeroth through second frequency moments ($O_\epsilon(\ln n)$). Later on in the class, we will focus on approximating frequency moments for $p \in [0, 2]$, which will be doable also in $O_\epsilon(\ln n)$ time, but in this lecture we study approximating the moments for some fixed $p > 2$.

Theorem 3. *For $p > 2$, there exists a sketch that uses space $O(n^{1-\frac{2}{p}} \ln n)$ for a constant ($O(1)$) approximation for F_p. This sketch is even an optimal result for linear sketches with constant approximation.*

2.1 Precision sampling

To prove this theorem, we introduce a new tool called *precision sampling*. To illustrate its usefulness, we consider a toy example.

Example 4. *Given numbers $a_1, \ldots, a_n \in [0, 1]$ we wish to reason about $\sum_i a_i$. Say that we want to distingiush between two cases: (1) $\sum a_i \leq 0.1$, (2) $\sum a_i > 10$. Note that 0.1 and 10 are relatively small compared to the maximal value that the sum can take (which is n).*

How many samples a_i do we need to be able to distinguish between these two cases[1]?

Observation 5. *Without "precision sampling" (not yet defined), we need roughly n samples to distinguish between the two cases. An adversary can construct the following two groups of numbers: one where $a_i = 0\ \forall i$ and one where $a_i = 0\ \forall i \notin S, a_i = 1\ \forall i \in S$, where $S = i_1, \ldots, i_{11}$. These two groups are identical except for 11 items, but belong to different cases. Hence, we must sample roughly n numbers to decisively discern between the cases.*

This is obviously an unsatisfying solution - we want to be able to answer this question without having to basically sample every number in the group. Precision sampling presents an alternative framework for accessing samples. In this model, we suppose we can access an *estimate* of a_i up to some precision $u_i > 0$. Then, algorithms interact with the precision sampling framework as such:

1. **Nature** chooses $a_1, \ldots, a_n$
2. The **algorithm** sets some uncertainties $u_1, \ldots, u_n$
3. **Nature** returns $\hat{a_i}$ s.t. $|a_i - \hat{a_i}| \leq u_i$
4. The **algorithm** uses $\{(\hat{a_i}, u_i)\}_{i=1}^n$ to decide on the problem it is tackling

For example, we might use $u_i = \frac{1}{n}\ \forall i$ which leads to $\sum \hat{a_i} \in [\sum a_i \pm 1]$. More trivially, we could even set $u_i = 0$. Obviously we need to constrain u_i for the framework to differ in any way from standard sampling. Ideally, we would like to avoid requiring fine precision, and be able to solve problems like distinguishing between the two cases of $\sum a_i$ presented above while only coarsely sampling elements.

Guided by this intuition, we quantify a *cost* c_p on $u_1, \ldots, u_n$ parametrized by a parameter $p > 1$:

$$c_p = \frac{1}{n} \sum_{i=1}^{n} (\frac{1}{u_i})^{1/p}$$

[1]We make no observations about the underlying distribution of the variables a_i. In this class, we consider worst-case runtimes, so in the setting where we would make distribution assumptions, we must assume the existence of an adversary that intends to fool our algorithm. An example of such an adversarial distribution is presented as part of the argument that $\sim n$ samples are needed to distinguish between the two cases presented above.

We now move to formalizing the precision sampling framework.

Lemma 6. *Precision Sampling Lemma: we can solve goals like the two-case distinction described in the above example, with* $\mathbb{E}_{u_i}[c_p] = O(1)$.

To guide the proofs that will follow, we note that this lemma requires that the u_i must be random variables, and that the final estimator we use not will "look like" $\sum \hat{a_i}$ as we might expect.

Proof. Let $u_i \sim Exp(\lambda = 1)$ be the random uncertainties[2]. Then, our estimator for $\sum a_i$ is $E = \max_i \frac{\hat{a_i}}{u_i}$. To analyze this estimator, we consider the quantity $\frac{\hat{a_i}}{u_i}$. Trivially, $\frac{\hat{a_i}}{u_i} \in [\frac{a_i \pm u_i}{u_i}] = [\frac{a_i}{u_i} \pm 1]$. Therefore, $E \in [\max(\frac{a_i}{u_i} \pm 1)] = [\max \frac{a_i}{u_i}] \pm 1 = \frac{1}{\min \frac{u_i}{a_i}} \pm 1$.[3] We take a brief detour to analyze the quantity $\min u_i/a_i$.

Fact 7. $\min \frac{u_i}{a_i}$ *is distributed as* $\frac{u}{\sum a_i}$ *where* $u \sim Exp(1)$ *(due to the memoryless property of the exponential distribution).*

Now, we have that $E \sim \frac{\sum a_i}{u} \pm 1$ where $u \sim Exp$. Since $u \in [0.5, 2]$ w.p. $\geq e^{-0.5} - e^{-2}$, $E \in [\frac{\sum a_i}{2} - 1, \frac{\sum a_i}{2} + 1]$ with the same probability. Returning to our toy example, such an estimator E provides enough granularity to distinguish between the two cases, since in case (1) where $\sum a_i \leq 0.1$, $E_{max} = 1.05$, and in case (2) where the sum is > 10, $E_{min} = 4$. Since there is no overlap between the values that E can take on in these two cases, we can use it to answer the original question posed.

It remains to prove that c_p is indeed small (more specifically, $\mathbb{E}[c_p] = O(1)$), which was the motivation for this more complex framework in the first place.

Proof. $\mathbb{E}[c_p] = \mathbb{E}[\frac{1}{n}\sum(\frac{1}{u})^{1/p}] = \mathbb{E}[(\frac{1}{u})^{1/p}]$. The last equality follows since the u_i have identical expectations, and n such u_is are summed up and divided by n. This expectation is equal to $\int_0^\infty (\frac{1}{u})^{1/p}\ e^{-u}\ du = \underbrace{\int_0^1 (\frac{1}{u})^{1/p}\ e^{-u}\ du}_{\text{(Integral 1)}} + \underbrace{\int_1^\infty (\frac{1}{u})^{1/p}\ e^{-u}\ du}_{\text{(Integral 2)}}$. We split the integral into two parts since in the first integral, the fraction dominates, and in the second, the exponential does. We now evaluate both parts of the integral:

1. For $u \in [0, 1]$, $e^{-u} \leq 1$. Since $(\frac{1}{u})^{1/p}$ converges for $p > 1$, the integral evaluates to $O_p(1)$ (constant as a function of p).
2. For $u \geq 1$, $(\frac{1}{u})^{1/p} \leq 1$. Since $\int_0^\infty e^{-u} = 1$ (it's the pdf of a distribution), the entire integral is also $O(1)$.

Adding both parts together, the original integral is $O(1)$ as well, so $\mathbb{E}[c_p] = O(1)$. □

2.2 F_p sketching

Now, we return to prove the original theorem on the existence of $O(n^{1-\frac{2}{p}} \ln n)$-space sketches for $O(1)$-approximations to F_p, $p > 2$. The proof is slightly complicated so some details are elided.

Proof. We hold a hash table H of width w and define:

$$\forall k \in [w],\ H[k] = \sum_{i:h(i)=k} \frac{r_i f_i}{u_i^{1/p}}$$

[2]This distribution is on $u > 0$ where $pdf(u) = e^{-u}$. For conciseness, we will denote it as Exp in the future.

[3]$\max x = \max \frac{1}{1/x} = \frac{1}{\min 1/x}$

Where $h : [n] \to [w], r_i =_r \{\pm 1\}, u_i \sim Exp$. This combines the randomized sign idea from earlier with the precision sampling regime. Note that the u_i must be independent and that achieving this proof with limited randomness is nontrivial. Note also that we use one hash table here, not L, and that this sketch is a linear one (see previous lecture).

Our estimator for F_p is then $E = \max_k |H[k]|^p$. The intuition behind the max operator is similar to that of precision sampling, taking advantage of the properties of the exponential distribution. We will now prove that E is an $O(1)$ approximation to F_p.

Proof. Let $y_i = \frac{f_i}{u_i^{1/p}}$ and $z_k = H[k]$ ($y \in \mathbb{R}^n, z \in \mathbb{R}^k$). We can then visualize this process as taking the original length n frequency vector f, scaling each value by $\frac{1}{u_i^{1/p}}$ (using precision sampling and yielding y), perturbing each y_i's sign randomly (the tug-of-war idea), and bucketing them by hash (from CountMin) resulting in a length k vector, z.

Claim 8.

$$||y||_\infty \sim (\frac{F_p}{u})^{1/p}$$

Where $u \sim Exp$.

Proof. $||y||_\infty = max_i|y_i| = (max_i|y_i|^p)^{1/p} = (max_i \frac{|f_i|^p}{u_i})^{1/p} \sim (\frac{\sum |f_i|^p}{u})^{1/p} = (\frac{F_p}{u})^{1/p}$. □

Corollary 9.

$$||y||_\infty^{1/p} \in [0.5, 2]F_p \ w.p. \geq e^{-0.5} - e^{-2}$$

We want $||y||_\infty^{1/p}$ to be relatively large w.r.t. the norm of the vector y so that we can extract it using HH. It turns out that the correct norm to examine here is the 2-norm.

Claim 10.

$$||y||_2^2 \leq O(1)||f||_2^2 \ w.p. \ 99\%$$

Proof. $\mathbb{E}[||y||_2^2] = \mathbb{E}[\sum y_i^2] = \mathbb{E}[\sum(\frac{f_i^2}{u_i^{2/p}})] = \sum \frac{f_i^2}{\mathbb{E}[u_i^{2/p}]}$. The denominator of each sum term's fraction is $O(1)$ if $p > 2$, so the whole expectation is $\leq ||f||_2^2 \cdot O(1)$. Then, by Markov, $||y||_2^2 \leq O(1) \cdot ||f||_2^2$ w.p. 99%. □

Continuing with our quest to find F_p using HH, we relate $||f||_2^2$ to the p-norm of f.

Corollary 11.

$$||y||_2^2 \leq O(F_p^{2/p} \cdot n^{1-2/p}) \ w.p. \ 99\%$$

Proof. This follows from Holder's inequality. Details omitted.

Now, all that remains is to analyze z (recalling that $z_k = H_k$). We observe that the maximum element in y will be a HH in z with $\phi \approx n^{1-2/p}$. We now analyze $||z||_\infty$ (which is $E^{1/p}$ where E is our estimate for F_p). Fix some k. We are interested in estimating z_k, but for convenience will estimate z_k^2:

$$\mathbb{E}[z_k^2] = \mathbb{E}_{r_i,h}[(\sum_{i:h(i)=k} r_i y_i)^2] \underset{\text{by T.o.W.}}{=} \mathbb{E}_h[\sum_{i:h(i)=k} y_i^2] = \frac{\sum y_i^2}{w} = \frac{||y||_2^2}{w}$$

Combining results, we have that $\mathbb{E}[z_k^2] \leq \frac{1}{w} O(F_p^{2/p} \cdot n^{1-2/p})$ w.p. 99%. Setting $w >> n^{1-2/p} \ln n$,[4] we have $\mathbb{E}[z_k^2] << F_p^{2/p}$. By Markov, we have that $z_k^2 \leq O(1)F_p^{2/p}$ w.p. 99% which implies that $|z_k| \leq O(1)F_p^{1/p}$ with the same probability.

[4] Reasoning behind factor of $\ln n$ not discussed in lecture.

To conclude, we have that in z, $z_{h(\arg\max_i y_i)}$ is relatively large and is $\approx F_p^{1/p}$ by Corollary 9. We also have that the average bucket is $\leq O(1)F_p^{1/p}$ from the above analysis. So the bucket where the maximum element of y is hashed to is relatively large compared to other buckets, and is $\approx F_p^{1/p}$, justifying our original estimator $E = \max_k |H[k]|^p$.

Claim 12. *For all k, $z_k \leq O(1)F_p^{1/p}$, and $\exists k$ s.t. $z_k = \Omega(F_p^{1/p})$.*

Claim not proved. However, from this claim, it follows directly that we can use our estimator to predict F_p with a constant-factor approximation! □