CS 395T: Sublinear Algorithms Fall 2016

Lecture 4 — September 6, 2016

Prof. Eric Price *Scribes: Cody Freitag, William Swartworth*

1 Overview

We previously discussed Markov's Inequality, Chebyshev's Inequality, and Chernoff Bounds, which progressively require more assumptions for tighter concentration at the tails.

Markov's Inequality

Let $X \geq 0$ and $t > 0$. Then,

$$\Pr[X \geq t] \leq \frac{\mathrm{E}[X]}{t}.$$

Chebyshev's Inequality

Let $\mu := \mathrm{E}[X]$, $\sigma^2 := \mathrm{Var}[X] = \mathrm{E}[(X-\mu)^2]$, and $t > 0$. Then,

$$\Pr[|X - \mu| \geq t] \leq \frac{\sigma^2}{t^2}.$$

Chernoff Bounds

Let $X = \sum_{i=1}^{n} X_i$ such that $X_i \in [0,1]$, fully independent, and $\mu := \mathrm{E}[X]$. Then,

$$\Pr[|X - \mu| > t] \leq e^{-C\frac{t^2}{\sigma^2}}$$

for some constant C.

Today, we will look at how to improve the failure probability for the distinct elements problem from last class. Along the way, we will begin to develop some tools that will allow for tighter analysis of the LogLog algorithm.

2 Improving Distinct Elements

Recall the LogLog algorithm from last class:

1. For $i = 1$ to m, in parallel:
   (a) Pick a hash function $h_i : [n] \to [0,1]$.
   (b) Record $Y_i = \min_{x \in S} h_i(x)$.
2. Compute $\hat{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$.
3. Output $1/\hat{Y} - 1$.

Note n is the size of the universe, S is the stream, and k is the true number of distinct elements in the stream. We showed last time that $\mathrm{Var}(Y_i) \lesssim 1/k^2$. Using Chebyshev's Inequality, this implies that $m = O(1/\epsilon^2)$ trials suffice to get $|\hat{Y} - \mu| \leq \epsilon/k$ with probability at least $3/4$. What if instead we want $|\hat{Y} - \mu| \leq \epsilon/k$ with at least $1-\delta$ probability? We have that $\mathrm{Var}(\hat{Y}) \lesssim \frac{1}{n^2 m}$, so Chebyshev's tells us $m = O(\frac{1}{\epsilon^2 \delta})$ trials suffice.

We want to improve the dependency of m on δ from $1/\delta$ to $\log(1/\delta)$, which we can go about in two different ways. Either we can find a different algorithm with a better dependence on δ, or we can show that our current algorithm suffices.

2.1 Median of Means

We want to find a different algorithm that calculates $\tilde{Y}$ such that $|\tilde{Y} - \mu| \leq \epsilon/k$ using fewer samples. We will solve this using the "median of means" approach. So we will split our trials in to R groups, $S_1, \ldots, S_R$, and set

$$\tilde{Y} = \operatorname*{median}_{i \in [R]} \left(\operatorname*{mean}_{j \in S_i} Y_j \right).$$

Let E_i be the indicator of the event $|(\mathrm{mean}_{j \in S_i} Y_j) - \mu| \leq \epsilon/k$. If at least $R/2$ of the E_i events occur, then $|\tilde{Y} - \mu| \leq \epsilon/k$.

Each $E_i \in \{0, 1\}$ is independent, and if $|S_i| \geq O(1/\epsilon^2)$, then $\Pr[E_i] \leq 1/4$. So using a Chernoff bound, we get that

$$\Pr\left[\sum_i E_i \geq R/2 \right] \leq e^{-\Omega(R)}.$$

This means that if $R \geq O(\log(1/\delta))$, then $|\tilde{Y} - \mu| \leq \epsilon/k$ with probability $1 - \delta$. In total,

$$m = O\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta} \right)$$

trials suffice.

In fact, we didn't need to take the median of the means in order to use $O\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$ trials. We will now develop some more tools to eventually show why the taking the mean of $O\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$ trials works.

3 Higher Moments

Suppose X has mean zero, and variance $\sigma^2 = \mathrm{E}[X^2]$. We can apply Markov's Inequality as we did to prove Chebyshev to get a concentration inequality for X. Namely we have,

$$\Pr[|X| \geq t] = \Pr[X^2 \geq t^2] \leq \frac{\mathrm{E}[X^2]}{t^2}.$$

Now suppose that $X = \sum_{i=1}^n X_i$, for n independent, mean-zero $X_i \in [-1, 1]$. Note that $\mathrm{E}[X_i^k] \leq 1$ for all values of k, and when $k = 1$, $\mathrm{E}[X_i] = 0$ since each X_i has mean zero. To apply our second

moment bound we would like to estimate $E[X^2]$. Using only pairwise independence, we have

$$\mathrm{E}\left[X^2\right] = \mathrm{E}\left[\sum_i X_i^2 + \sum_{i \neq j} X_i X_j\right] = \sum_i \mathrm{E}\left[X_i^2\right] \leq n.$$

From our second-moment bound, this implies that $\Pr[|X| \geq t\sqrt{n}] \leq 1/t^2$.

What if we look at the fourth moment instead? We can apply Markov's Inequality as before, so

$$\Pr[|X| \geq t] = \Pr[X^4 \geq t^4] \leq \frac{\mathrm{E}[X^4]}{t^4}.$$

We can bound $E[X^4]$ now using 4-wise independence to get

$$\mathrm{E}\left[\left(\sum_i X_i\right)^4\right] = \sum_i \mathrm{E}[X_i^4] + \sum_{i \neq j} 6\,\mathrm{E}[X_i^2]\,\mathrm{E}[X_j^2] \leq n + 6n(n-1) \leq 6n^2$$

This implies that $\Pr[|X| \geq t\sqrt{n}] \leq 6/t^4$. Notice that for large t this beats our second moment bound, while for small t the second moment wins.

In fact this method extends to all (even) higher moments. The same approach with a slightly more careful analysis shows that

$$\Pr[|X| \geq t\sqrt{n}] \leq \left(\frac{\sqrt{k}}{t}\right)^k$$

for all even k. Furthermore, we only use the fact that the the X_i's are k-wise independent. So this is a useful method to concentrate $\sum X_i$ without full independence that the Chernoff Bound requires.

If our X_i are fully-independent, then for a given value of t we can optimize this bound over k. Setting $k = t^2/e$ implies that

$$\Pr[|X| \geq t\sqrt{n}] \leq e^{-t^2/(2e)}.$$

In fact, Chernoff claims that for $X_i \in [0, 1]$,

$$\Pr[|y - \mu| \geq \epsilon\mu] \leq 2e^{-\epsilon^2/(2+\epsilon)\mu}.$$

Instead, for $X_i \in [-1, 1]$ we get

$$\Pr[|y - \mu| \geq \epsilon\mu] \leq 2e^{-\epsilon^2\mu^2/(2en)},$$

which is only worse by a μ/n factor in the exponent.

4 Moment Generating Function

While higher moments can be useful for concentrating random variables, they quickly become cumbersome to work with. Moment generating functions encapsulate information about all of the moments within a single, much simpler object.

The Moment Generating Function of a mean-zero random variable X is

$$\mathrm{MGF}(X) = \mathrm{E}\left[e^{\lambda X}\right] = \mathrm{E}\left[1 + \lambda X + \frac{\lambda^2 X^2}{2!} + \cdots\right].$$

Following the method of the last section, Markov's Inequality tells us that

$$\Pr[X > t] = \Pr[e^{\lambda X} > e^{\lambda t}] \leq \frac{\mathrm{E}[e^{\lambda X}]}{e^{\lambda t}}$$

for all $\lambda > 0$. (And we get a similar lower tail for $\lambda < 0$.)

In the case of a Gaussian $p(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-t^2/(2\sigma^2)}$, we can express the moment generating function explicitly:

$$\begin{aligned}
\mathrm{E}[e^{\lambda X}] &= \int_{-\infty}^{\infty} p(t) e^{\lambda t} dt \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}t^2 + \lambda t} dt \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t+\lambda\sigma^2)^2 + \frac{\sigma^2\lambda^2}{2}} dt \quad \text{(By completing the square.)} \\
&= e^{\frac{\sigma^2\lambda^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t+\lambda\sigma^2)^2} dt \\
&= e^{\frac{\sigma^2\lambda^2}{2}} \int_{-\infty}^{\infty} p(t - \lambda\sigma^2) dt \\
&= e^{\frac{\sigma^2\lambda^2}{2}} \quad \text{(Since the integral of a gaussian is 1).}
\end{aligned}$$

So we've shown that for X gaussian with mean 0,

$$\mathrm{MGF}(X) = e^{\frac{\lambda^2\sigma^2}{2}}$$

Therefore,

$$\Pr[X \geq t] \leq e^{-\frac{t^2}{2\sigma^2}}$$

from setting $\lambda = t/\sigma^2$.

Combining this with the lower tail, we get the bound

$$\Pr[|X| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

5 Subgaussian Random Variables

What if we want to use the MGF to bound a non-gaussian random variable? It's possible we can compute $\mathrm{MGF}(X)$ exactly, but it could be the case that our random variable is too complicated. We define the notion of a *subgaussian* random variable as one whose moment generating function is bounded by that of a gaussian.

Definition 5.1. *A mean-zero random variable X is subgaussian with parameter σ^2 if for all λ,*

$$\mathrm{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}}.$$

In general, one can show that X is subgaussian (up to constant factors in σ) if any of the following criteria hold:

1. MGF Bound: $\mathrm{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for all λ,
2. Tail Bound: $\Pr[|X| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$ for all $t > 0$, or
3. Moment Bound: $\mathrm{E}[|X|^k] \leq \sigma^k k^{\frac{k}{2}}$ for all $k > 0$

Remark 5.2. *A (possibly non-mean-zero) random variable X is subgaussian with parameter σ^2 if $X - \mathrm{E}[X]$ is subgaussian with parameter σ^2.*

We now give some useful lemmas about subgaussian random variables.

Lemma 5.3. *Suppose X is a bounded random variable, i.e. $X \in [a, b]$ for $a, b \in \mathbb{R}$. Then X is subgaussian with parameter $\sigma^2 = \left(\frac{b-a}{2}\right)^2$.*

Lemma 5.4. *If X and Y are subgaussian with parameters s_1^2 and s_2^2, respectively, and independent, then $X + Y$ is subgaussian with parameter $\sigma^2 = s_1^2 + s_2^2$.*

Proof. We'll directly apply the MGF bound to show this.

$$\begin{aligned}
\mathrm{E}[e^{\lambda(x+y)}] &= \mathrm{E}[e^{\lambda x}e^{\lambda y}] \\
&= \mathrm{E}[e^{\lambda x}]\,\mathrm{E}[e^{\lambda y}] \\
&\leq e^{\frac{\lambda^2 s_1^2}{2}} e^{\frac{\lambda^2 s_2^2}{2}} \\
&= e^{\frac{\lambda^2(s_1^2+s_2^2)}{2}}
\end{aligned}$$

□

5.1 Proof of the Chernoff Bound

With only the previous two lemmas, we have the machinery in place to prove the following version of the Chernoff Bound.

Theorem 5.5. *Let $X = \sum_{i=1}^n X_i$ such that $X_i \in [0, 1]$, independent, and $\mu := \mathrm{E}[X]$. Then,*

$$\Pr[|X - \mu| > t] \leq e^{-\Omega\left(\frac{t^2}{\sigma^2}\right)}$$

Proof. Since each X_i is bounded by $[0, 1]$, X_i is subgaussian with parameter $\sigma^2 = \left(\frac{1-0}{2}\right)^2 = \frac{1}{4}$. Since all X_i are independent, $X = \sum_{i=1}^n X_i$ is subgaussian with parameter $\frac{n}{4}$. This means

$$\Pr[|X - \mu|] > t] \leq 2e^{-2t^2/n} = e^{-\Omega\left(\frac{t^2}{n}\right)}.$$

□

Note that by a similar analysis, if some of the X_i are gaussian or on larger ranges, the same theorem holds.