CPSC 665 : An Algorithmist's Toolkit Lecture 2 : 14 Jan 2015

Concentration Bounds

Lecturer: Sushant Sachdeva *Scribe: Cyril Zhang*

Introduction

Concentration bounds allow us to show that a random variable, under certain conditions, lies near its mean with high probability. We proved the inequalities of Markov, Chebyshev, and Chernoff, and used these to analyze a median-of-means amplification of Morris' approximate counting algorithm.

1. Markov's Inequality

We begin with our most general bound, Markov's inequality.

Theorem 1.1 (Markov's inequality). *Let X be a non-negative random variable. Then, for any $t > 0$,*

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

Proof. Let ρ be the probability density function of X, so that $\Pr[a \leq X \leq b] = \int_a^b \rho(x)dx$.

$$\mathbb{E}[X] = \int_0^\infty x\rho(x)dx \geq \int_t^\infty x\rho(x)dx \geq t \cdot \int_t^\infty \rho(x)dx = t \cdot \Pr[X \geq t].$$

□

A way to think about what this proof is doing: X dominates the scaled indicator variable $T = t \cdot \mathbb{1}_{X \geq t}$, so we have $\mathbb{E}[X] \geq \mathbb{E}[T] = t \cdot \Pr[X \geq t]$.

Markov's inequality gives a rather weak bound when applied directly; the random variables we care about are usually much more highly concentrated. Let's look at a toy example: flip 100 coins. What's the probability that at least 70 of them come up heads? Markov's inequality tells us that it's no greater than 5/7. As we'll see, we can do much better.

2. Chebyshev's Inequality

Theorem 2.1 (Chebyshev's inequality). *Let X be any random variable. Then, for any $t > 0$,*

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

Proof. Use Markov's inequality on the positive random variable $(X - \mu)^2$, whose expected value is precisely the variance of X:

$$\Pr[|X - \mu| \geq t] = \Pr[|X - \mu|^2 \geq t^2] = \Pr[(X - \mu)^2 \geq t^2]$$
$$\leq \frac{\mathbb{E}[(X - \mu)^2]}{t^2} = \frac{\mathrm{Var}[X]}{t^2}.$$

□

Back to our toy example of 100 coins. The variance of a Bernoulli random variable with parameter p is $(1-p)(-p)^2 + p(1-p)^2 = p(1-p)$. The variance of a sum of independent random variables is the sum of variances. So, with 100 coins, the variance of the number of heads is $100 \cdot \frac{1}{4} = 25$. Chebyshev's inequality with $t = 20$ gives us a bound of $25/400 = 1/16$. Note that we're being a little crude, since the two-tailed bound when we need the one. Much better than Markov's 5/7. But this is not the best we can do; the real answer is around 1.6×10^{-4}.

3. Chernoff Bounds

As we flip more and more coins, we should expect that the number of heads $X^{(n)}$ gets more and more concentrated around the mean. We can ask: how large does n have to be for the probability that you see more than $(1+\delta)\mathbb{E}[X^{(n)}]$ heads fall below ϵ?

Markov's inequality says:

$$\Pr\left[X^{(n)} \geq (1+\delta)\mathbb{E}[X^{(n)}]\right] \leq \frac{1}{1+\delta}.$$

Oops. This doesn't even depend on n. How about Chebyshev?

$$\Pr\left[X^{(n)} \geq (1+\delta)\mathbb{E}[X^{(n)}]\right] \leq \Pr\left[|X^{(n)} - \mu| \geq \delta \cdot \mu\right] \leq \frac{\mathrm{Var}[X^{(n)}]}{\delta^2\mu^2}$$
$$= \frac{n/4}{\delta^2(n/2)^2} = \frac{1}{\delta^2 n}.$$

So, n needs to be at least $\frac{1}{\delta^2\epsilon}$. This is still much weaker concentration than the true behavior of a sum of many independent coin flips. Chernoff bounds state a tighter result.

Say you have n i.i.d. Bernoulli random variables $\{X_i\}$, each with parameter $\mathbb{E}[X_i] = p$, so that $\mathbb{E}[X^{(n)}] = np := \mu$. We wish to bound $\Pr[X^{(n)} \geq (1+\delta)\mu]$.

Pick some $\lambda > 0$. We'll obtain a family of inequalities parameterized by λ. Then, apply $x \mapsto e^{\lambda x}$ to both sides:

$$\Pr[X^{(n)} \geq (1+\delta)\mu] = \Pr\left[\exp\left(\lambda X^{(n)}\right) \geq \exp\left(\lambda(1+\delta)\mu\right)\right].$$

Apply Markov's inequality:

$$\leq \frac{\mathbb{E}\left[\exp(\lambda X^{(n)})\right]}{\exp\left(\lambda(1+\delta)\mu\right)}.$$

First, we bound the numerator. Since the X_i's are independent, expectation is multiplicative:

$$\mathbb{E}\left[\exp(\lambda X^{(n)})\right] = \mathbb{E}\left[\prod_{i=1}^{n}\exp(\lambda X_i)\right] = \prod_{i=1}^{n}\mathbb{E}\left[\exp(\lambda X_i)\right].$$

Each $\mathbb{E}[\exp(\lambda X_i)]$ is identical and easy to evaluate:

$$\mathbb{E}[\exp(\lambda X_i)] = p \cdot e^{\lambda} + (1-p)\cdot e^0 = 1 + (e^\lambda - 1)p \leq \exp\left((e^\lambda - 1)p\right),$$

where the last inequality follows from $1 + x \leq e^x$, which we showed last class using convexity. So the entire numerator is bounded by

$$\exp\left(n \cdot (e^\lambda - 1)p\right) = \exp\left(\mu(e^\lambda - 1)\right).$$

So we have so far

$$\Pr\left[X^{(n)} \geq (1+\delta)\mu\right] \leq \frac{\exp\left(\mu(e^\lambda - 1)\right)}{\exp\left(\lambda(1+\delta)\mu\right)},$$

for all λ. In particular, it's true for $\lambda = \ln(1+\delta)$. This gives us the strongest statement of the Chernoff bound:

Theorem 3.1 (Messy Chernoff upper bound).

$$\Pr\left[X^{(n)} \geq (1+\delta)\mu\right] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu.$$

This is not the form we usually want, but it's the strongest bound we can get by this method. Here's a lemma that gives us a cleaner form:

HW 1: Show that for $\delta \in [0,1]$, $(1+\delta)^{1+\delta} \geq \exp\left(\delta + \frac{\delta^2}{3}\right)$. Easy calculus.

This gives us, for $0 \leq \delta \leq 1$,

$$\Pr\left[X^{(n)} \geq (1+\delta)\mu\right] \leq \exp\left(\frac{-\delta^2\mu}{3}\right).$$

HW 2: Prove the Chernoff lower bound: for $\delta \in [0,1]$, $\Pr[X^{(n)} \leq (1-\delta)\mu] \leq \exp\left(-\frac{\delta^2\mu}{2}\right)$. Same strategy as the upper bound.

Altogether, we have:

Theorem 3.2 (**Chernoff bound).** *Let $\{X_i\}$ be i.i.d. Bernoulli random variables with parameter p, $X^{(n)} = \sum_{i=1}^n X_i$, and $\mu = \mathbb{E}[X^{(n)}] = np$. Then, for any $\delta \in [0,1]$,*

$$\Pr\left[X^{(n)} \geq (1+\delta)\mu\right] \leq \exp\left(-\frac{\delta^2\mu}{3}\right).$$

$$\Pr\left[X^{(n)} \leq (1-\delta)\mu\right] \leq \exp\left(-\frac{\delta^2\mu}{2}\right).$$

To guarantee a tail probability less than $\epsilon = \exp\left(-\frac{\delta^2 np}{3}\right)$, we need $n \geq \frac{6}{\delta^2}\ln\left(\frac{1}{\epsilon}\right)$. Back to our example with 70 heads in 100 coin tosses, Chernoff gives us around 3%, not much of an improvement from Chebyshev. But as n grows larger, the Chernoff bound gets even stronger.

Some remarks:

- The δ^2 in the exponent is tight with the Gaussian distribution up to a constant, which is what a sum of i.i.d. Bernoulli variables converges to, by the central limit theorem.
- In probability theory, we call $\mathbb{E}[e^{\lambda X}]$ the **moment generating function** of X. It's a power series where the coefficient of λ^k is the k-th moment of X, scaled by $1/k!$. You might also recognize it as the **Laplace transform**.

The same kind of technique allows us to prove Chernoff-like bounds under more general conditions. First, if we have a sum of any (independent) random variables between 0 and 1, the same tail bounds hold. The Bernoulli variable case can thus be thought of as the "worst case" of this result.

Theorem 3.3 (**Hoeffding's inequality with identical means).** *Let $\{X_i\}_{i=1}^n$ be independent random variables such that $0 \leq X_i \leq 1$ and $\mathbb{E}[X_i] = p$. Let $X^{(n)} = \sum_{i=1}^n X_i$, so that $\mathbb{E}[X^{(n)}] = np := \mu$. Then, the inequalities in Theorem 3.2 hold.*

Proof. Identical up to the point where we compute $\mathbb{E}[\exp(\lambda X_i)] = 1 + (e^\lambda - 1)p \leq \exp\Big((e^\lambda - 1)p\Big)$. It will suffice to show $\mathbb{E}[\exp(\lambda X_i)] \leq 1 + (e^\lambda - 1)p$, so that the rest of the proof follows identically. Since $e^{\lambda x}$ is convex (as a function of x), it lies below any secant line. If we pick the secant line that runs through $(0,1)$ and $(1, e^\lambda)$, we find $e^{\lambda x} \leq 1 + (e^\lambda - 1)x$ on the support of X_i. Taking expectations on both sides of $e^{\lambda X_i} \leq 1 + (e^\lambda - 1)X_i$ gives us the desired inequality. $\square$

What if each variable has a different mean? Then, the bounds still hold.

Theorem 3.4 (**Hoeffding's inequality with arbitrary means).** *Let $\{X_i\}_{i=1}^n$ be independent random variables such that $0 \leq X_i \leq 1$ and $\mathbb{E}[X_i] = p_i$. Let $X^{(n)} = \sum_{i=1}^n X_i$, so that $\mathbb{E}[X^{(n)}] = \sum_i p_i := \mu$. Then, the inequalities in Theorem 3.2 hold.*

Proof. In the Chernoff proof, we got a bound on the numerator:

$$\prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda X_i)\right] \leq \exp\left((e^\lambda - 1)\mu\right).$$

We still have $\mathbb{E}[\exp(\lambda X_i)] \leq 1 + (e_\lambda - 1)p_i$. By the AM-GM inequality ($\prod a_i \leq \left(\frac{\sum a_i}{n}\right)^n$),

$$\begin{aligned}\prod_{i=1}^{n}\left(1 + (e^\lambda - 1)p_i\right) &\leq \left(1 + (e^\lambda - 1)\frac{\sum p_i}{n}\right)^n \\ &\leq \exp\left(n(e^\lambda - 1)\left(\frac{\sum p_i}{n}\right)\right) = \exp\left((e^\lambda - 1)\mu\right).\end{aligned}$$

So, again, our proof can proceed identically. □

4. Approximate counting

4.1. **A sampling algorithm.** Suppose you want to count a very large number of objects– so large that you don't have enough memory to store the number. Robert Morris (1978) only had one-byte counters, and needed to count a two-byte number of objects. His approach was to estimate the log of the count by randomly rejecting most of the objects.

Formally, you see a stream of objects, and you want to estimate the number of objects, without having to store huge integers. The algorithm is as follows:

- Initialize $X_0 := 0$.
- Every time you see an object, $X_{i+1} := \begin{cases} X_i + 1 & \text{w.p. } \frac{1}{2^X_i} \\ X_i & \text{otherwise.} \end{cases}$
- Output $2^{X_n} - 1$.

Notice: that the process increments the estimator $2^{X_i} - 1$ in expectation:

$$\begin{aligned}\mathbb{E}[2^{X_n}|X_{n-1}] &= \frac{1}{2^{X_{n-1}}} \cdot 2^{X_{n-1}+1} + \left(1 - \frac{1}{2^{X_{n-1}}}\right) \cdot 2^{X_{n-1}} \\ &= 2^{X_{n-1}} + 1\end{aligned}$$

By induction, this gives us $\mathbb{E}[2^{X_n} - 1] = n$.

To get an idea of the concentration, we first compute the variance:

$$\begin{aligned}\mathrm{Var}[2^{X_n} - 1] = \mathrm{Var}[2^{X_n}] &= \mathbb{E}[(2^{X_n})^2] - \mathbb{E}[2^{X_n}]^2 \\ &= \mathbb{E}[2^{2X_n}] - (n+1)^2.\end{aligned}$$

And,

$$\mathbb{E}[2^{2X_n}|X_{n-1}] = \left(1 - \frac{1}{2^{X_{n-1}}}\right) \cdot 2^{2X_{n-1}} + \frac{1}{2^{X_{n-1}}} \cdot 2^{2(X_{n-1}+1)} = 2^{2X_{n-1}} + 3 \cdot 2^{X_{n-1}}.$$

So, by the law of total expectation,

$$\mathbb{E}[2^{2X_n}] = \mathbb{E}[2^{2X_{n-1}}] + 3 \cdot \mathbb{E}[2^{X_{n-1}}].$$

Induction gives

$$\mathbb{E}[2^{2X_n}] = \frac{3}{2}n(n+1) + 1,$$

So we have that the variance of the estimator is

$$\frac{3}{2}n(n+1) + 1 - (n+1)^2 = \frac{1}{2}n^2 - \frac{1}{2}n$$

$$\leq \frac{1}{2}n^2.$$

Used alone, this algorithm produces a very erratic estimator. Let's combine multiple copies of it to get something better-behaved.

4.2. **Mean of copies.** The most natural thing to try is to take a mean of several trials. Suppose we run t independent copies of the algorithm $\{X^{(i)}\}_{i=1}^{t}$, and let $Z = \frac{1}{t}\sum X^{(i)}$ be our new estimator. Then $\mathbb{E}[Z] = n$, and $\mathrm{Var}[Z] \leq t \cdot \frac{\frac{1}{2}n^2}{t^2} = \frac{n^2}{2t}$. Chebyshev's inequality gives

$$\Pr\Big[|Z - n| \geq \delta n\Big] \leq \frac{\mathrm{Var}[Z]}{\delta^2 n^2} \leq \frac{1}{2\delta^2 t}.$$

So, as we increase the number of copies, the variance decreases as $1/t$. We can't use Chernoff bounds here to get exponentially decreasing failure probability, since the summands are not bounded by a small range. We'll need a different strategy.

4.3. **Median of means.** Choose some target failure probability p. Let's use a mean of $t = \frac{3}{2\delta^2}$ copies of the counter as a subroutine, giving us a failure probability of

$$\Pr\Big[|Z - n| > \delta n\Big] \leq \frac{1}{3}.$$

Now, let's use this subroutine r times, getting independent estimators $Z^{(1)}, Z^{(2)}, \ldots, Z^{(r)}$. Consider what happens when we take their median. Call an estimator "wrong" if it lies outside the range $[(1-\delta)n, (1+\delta)n]$. Note that if the median of $\{Z^{(i)}\}$ is wrong, then at least $r/2$ of the $Z^{(i)}$'s are wrong.

Let $Y^{(i)}$ be the indicator variable for the event that $Z^{(i)}$ is wrong. Then, $W = \sum_{i=1}^{r} Y^{(i)}$ is the number of wrong $Z^{(i)}$'s. Denote $p \stackrel{\text{def}}{=} \mathbf{E}[Y^{(i)}]$. Thus, $\mathbb{E}[W] = pr$, and by our choice of t, we have $p \leq \frac{1}{3}$. Now, use the Chernoff bound on W to bound the probability that it's large enough for the median to be wrong.

$$\begin{aligned} \Pr\left[W \geq \frac{r}{2}\right] &= \Pr\left[W \geq \frac{1}{2p} \cdot pr\right] = \Pr\left[W \geq \left(1 + \frac{1}{2p} - 1\right)\mathbb{E}[W]\right] \\ &\leq \exp\left(-\left(\frac{1}{2p} - 1\right)^2 \frac{\mathbb{E}[W]}{3}\right) = \exp\left(-\frac{(1-2p)^2}{12p} r\right). \end{aligned}$$

Since this is a decreasing function of p and r, we see that it suffices to pick $r \geq O(\ln \frac{1}{\epsilon})$ to have $\Pr\left[|\mathrm{median}\{Z^{(i)}\} - n| > \delta n\right] \leq \epsilon$, and hence the failure probability falls exponentially.

Notes

The approximate counter algorithm used in this lecture can be found in the original paper by Morris [Mor78]. It is a short, simple, and fun read. The presentation in this lecture is largely from lecture notes on this topic from Jelani Nelson's course at Harvard [WL13].

References

[Mor78] Robert Morris. Counting large numbers of events in small registers. *Communications of the ACM*, 21(10):840–842, 1978.

[WL13] Andrew Wang and Andrew Liu. Lecture notes for CS 229r : Algorithms for Big Data by Jelani Nelson, Fall 2013.