18.304 April 29, 2013

The Second Moment Method

Pratiksha Thaker

1 Overview

The second moment method is the study of the use of variance, and in particular Chebyshev's inequality. It can be used to show that when the expected value of a nonnegative random variable is large compared to its variance, it takes on the value zero with probability approaching zero. Or, said differently, it can be used to show that a random variable has a positive probability of being positive. In this lecture, we will define variance and Chebyshev's inequality and then look at two applications: prime factors and distinct sums. The lecture is primarily derived from chapter 4 of [AS08].

2 Variance & Chebyshev's Inequality

Definition 1 (Variance)

$$\mathrm{Var}[X] = E[(X - E[X])^2].$$

Notation: $\sigma^2 = \mathrm{Var}[X]$, $\sigma = \sqrt{\mathrm{Var}[X]}$.

If we know the variance or an estimate of it, we can use Chebyshev's inequality. First we state Markov's inequality without proof, because we will need it to prove Chebyshev's inequality.

Theorem 2 (Markov's inequality) *X a nonnegative random variable, $t > 0$.*

$$P(X \geq t) \leq E[X]/t.$$

Theorem 3 (Chebyshev's inequality)

$$\Pr[|X - E[X]| \geq t\sigma] \leq \frac{1}{t^2}.$$

A useful variant is

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

Proof

$$\Pr[|X - E[X]| \geq t\sigma] = \Pr[(X - E[X])^2 \geq t^2\sigma^2]$$

Then using Markov's inequality for $(X - E[X])^2$, we get

$$\begin{aligned}\Pr[(X - E[X])^2 \geq t^2\sigma^2] &\leq \frac{E[(X - E[X])^2]}{t^2\sigma^2} \\ &= \frac{\sigma^2}{t^2\sigma^2} \\ &= \frac{1}{t^2}\end{aligned}$$

■

Definition 4 (Covariance)

$$\mathrm{Cov}[X, Y] = E[(X - E[X]) \cdot (Y - E[Y])] = E[XY] - E[X] \cdot E[Y]$$

It is useful to note the variance for a sum of random variables, i.e. $X = X_1 + X_2 + \ldots + X_n$:

$$\mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + \sum_{i \neq j} \mathrm{Cov}[X_i, X_j].$$

In many cases, we can upper-bound the variance of X_i, so to find an upper bound on the variance of the sum, we then only need a way of upper-bounding the covariance. (Note that we usually don't care about lower-bounding the variance.)

3 Number Theory

For an integer n, let $\nu(n)$ be the number of primes that divide n. Then we can use Chebyshev's inequality to prove a result which says that "most" integers have approximately $\ln \ln n$ prime factors.

Theorem 5 *Fix any $c > 1/2$. The number of integers $x \in \{1, \ldots, n\}$ with*

$$|\nu(x) - \ln \ln n| > c\sqrt{\ln \ln n} + 10$$

is at most $(1/c^2 + o(1)) \cdot n$.

Proof Choose x uniformly at random from $\{1, \ldots, n\}$. For a prime p, let

$$X_p = \begin{cases} 1 & \text{if } p|x, \\ 0 & \text{otherwise.} \end{cases}$$

Choose $M = n^{1/10}$, and set $X = X(x) = \sum_{p \leq M} X_p$ (that is, the total number of primes less than M). X can have at most 10 prime factors greater than $n^{1/10}$ (by prime number thm?), so we have $\nu(x) - 10 \leq X(x) \leq \nu(x)$, and because of the way we formulated the claim originally, it is then sufficient to prove the claim for $X(x)$. We have

$$\mathrm{E}[X_p] = \frac{\lfloor n/p \rfloor}{n}$$

Then since $y - 1 < \lfloor y \rfloor \leq y$,

$$\mathrm{E}[X_p] = 1/p + O(1/n).$$

Then using linearity of expectation,

$$\mathrm{E}[X] = \sum_{p \leq M} (1/p + O(1/n)).$$

Now we use a fact from number theory which can be derived from Stirling's formula:

$$\sum_{p \leq x} 1/p = \ln \ln x + O(1)$$

so

$$\mathrm{E}[X] = \ln \ln n + O(1).$$

In order to apply Chebyshev's inequality, we'll now upper-bound $\mathrm{Var}[X]$ using this formula:

$$\mathrm{Var}[X] = \sum_{p \le M} \mathrm{Var}[X_p] + \sum_{1 \le p \ne q \le M} \mathrm{Cov}[X_p, X_q].$$

Since X_p is a Bernoulli random variable, we have

$$\mathrm{Var}[X_p] = 1/p \cdot (1 - 1/p) + O(1/n)$$

so using the upper bound formula

$$\sum_{p \le M} \mathrm{Var}[X_p] = \sum_{p \le M} (1/o + O(1/n)) = \ln \ln n + O(1).$$

Now if p and q are distinct primes, then $X_p X_q = 1$ occurs only if $pq|x$. So

$$\begin{aligned}
\mathrm{Cov}[X_p, X_q] &= \mathrm{E}[X_p X_q] - \mathrm{E}[X_p] \cdot \mathrm{E}[X_q] \\
&= \frac{\lfloor n/(pq) \rfloor}{n} - \frac{\lfloor n/p \rfloor}{n} \cdot \frac{\lfloor n/q \rfloor}{n} \\
&\le \frac{n/(pq)}{n} - \frac{n/p - 1}{n} \cdot \frac{n/q - 1}{n} \\
&= 1/(pq) - (1/p - 1/n) \cdot (1/q - 1/n) \\
&\le (1/n) \cdot (1/p + 1/q).
\end{aligned}$$

Thus

$$\sum_{p \ne q} \mathrm{Cov}[X_p, X_q] \le 1/n \sum_{p \ne q} (1/p + 1/q) \le (2M)/n \sum_{p} (1/p)$$

and recall that $M = n^{1/10}$, so

$$\sum_{p \ne q} \mathrm{Cov}[X_p, X_q] \le O(n^{-9/10} \ln \ln n) = o(n^{-8/9}).$$

Using our earlier formula for $\mathrm{Var}[X]$, we get

$$\mathrm{Var}[X] = \ln \ln n + O(1).$$

Now, we can use the second variant of Chebyshev's inequality that we discussed earlier:

$$\Pr[|X - \mathrm{E}[X]| > t] < \frac{\mathrm{Var}[X]}{t^2}.$$

If we choose $t = c\sqrt{\ln \ln n}$, we get

$$\begin{aligned}
\Pr[|X - \mathrm{E}[X]| > c\sqrt{\ln \ln n}] &< \frac{\mathrm{Var}[X]}{(c\sqrt{\ln \ln n})^2 x} \\
&\le \frac{\ln \ln n + O(1)}{c^2 \ln \ln n} \\
&= 1/c^2 + o(1)
\end{aligned}$$

for any $c > 1/2$. Since $|X - \nu| \le 10$, we finally conclude that

$$\begin{aligned}
&\frac{|\{x \in \{1, \ldots, n\} : |\nu(x) - \ln \ln n| \le c\sqrt{\ln \ln n} + O(1)\}|}{n} \\
&= \Pr[|\nu(x) - \ln \ln n| > c\sqrt{\ln \ln n} + O(1)] \le \frac{1}{c^2} + o(1).
\end{aligned}$$

■

4 Distinct Sums

A set of positive integers $\{x_1, \ldots, x_k\}$ is said to have distinct sums if all sums

$$\sum_{i \in S} x_i, S \subseteq \{1, \ldots, k\}$$

are distinct. (All sums of all subsets of the integer set are distinct.) Define

$$f(n) = \max\{k \in \mathbb{N} : \text{there exists } \{x_1, \ldots, x_k\} \text{ with distict sums}\}.$$

We will use Chebyshev's inequality to bound $f(n)$.

First, a lower bound. Note that $1, 2, 4, \ldots, 2^{k-1}$ with $2^{k-1} \leq n$ has distinct sums. This holds for any $2^{k-1} \leq n$ so $k \leq \lg n + 1$, giving

$$f(n) \geq \lfloor \lg n \rfloor + 1.$$

Now, a first try for an upper bound. Note that the value of each sum for any subset is an integer between 0 and kn. Moreover, there are 2^k possible k-integer subsets which will all have to have distinct sums. This implies that

$$2^k \leq kn + 1.$$

The function $\frac{2^x - 1}{x}$ is increasing in $(0, \infty)$, so we have $f(n) \leq x(n)$ where $x(n)$ is the positive solution to $2^x = xn + 1$. Now we'll try to derive an asymptotic formula. Note that

$$\begin{aligned} 2^x \approx xn &\implies x = \lg x + \lg n + o(1) \\ x \approx \lg n &\implies \lg x = \lg \lg n + o(1). \end{aligned}$$

Then we get

$$x = \lg n + \lg \lg n + o(1)$$

by combining the inequalities.

Can we improve the upper bound? Yes, though only slightly (at least when only using the second moment method). Suppose $\{x_1, \ldots, x_k\} \subseteq \{1, \ldots, n\}$ have distinct sums. Define

$$X = \epsilon_1 x_1 + \ldots + \epsilon_k x_k$$

where $\epsilon_i \in \{0, \frac{1}{2}\}$ each with probability 1/2. So X is a random sum, and $\mathrm{E}[X] = \frac{1}{2} \sum_{i=1}^{n} x_i$, and

$$\mathrm{Var}[X] = \sum_{i=1}^{k} \mathrm{Var}[\epsilon_i x_i] + \sum_{i \neq j} \mathrm{Cov}[\epsilon_i x_i, \epsilon_j x_j].$$

Then using the independence of ϵ_i and ϵ_j, and the fact that $\mathrm{Var}[\epsilon_i x_i] = (1/2)(x_i - x_i/2)^2 + (1/2)(0 - x_i/2)^2 = x_i^2/4$, we get

$$\mathrm{Var}[X] = \sum_{i=1}^{k} \mathrm{Var}[\epsilon_i x_i] = \frac{1}{4} \sum_{i=1}^{k} x_i^2 \leq \frac{1}{4} k n^2.$$

Now we use the first version of Chebyshev's inequality:

$$\Pr[|X - \mathrm{E}[X]| \geq t\sigma] \leq \frac{1}{\sigma^2}$$

and since $\sigma \leq \frac{1}{2}n\sqrt{k}$,

$$\Pr[|X - \mathrm{E}[X]| \geq (t/2)n\sqrt{k}] \leq \frac{1}{t^2},$$

and equivalently

$$\Pr[|X - \mathrm{E}[X]| < (t/2)n\sqrt{k}] \geq 1 - \frac{1}{t^2}.$$

This gives us a lower bound.

But we said that $\{x_1, \ldots, x_k\}$ have distict sums, so for any integer $0 \leq i \leq kn$, $\Pr[X = i] \in \{0, 2^{-k}\}$. Then for any subset $T \subseteq \{0, \ldots, kn\}$,

$$\Pr[X \in T] = \Pr[\vee_{i \in T}(X = i)] \leq \sum_{i \in T} \Pr[X = i] \leq \sum_{i \in T} 2^{-k} = |T|2^{-k}.$$

If we choose $T = \{i \in \mathbb{N} : |i - \mathrm{E}[X]| < (t/2)n\sqrt{k}\}$, then

$$\Pr[|X - \mathrm{E}[X]| < (t/2)n\sqrt{k}] = \Pr[X \in T] \leq 2^{-k}(tn\sqrt{k} + 1).$$

This gives us an upper bound. We will now obtain a contradiction where the lower bound we showed earlier exceeds this upper bound. Rearranging gives

$$1 - \frac{1}{t^2} \geq \frac{tn\sqrt{k} + 1}{2^k} \implies \frac{2^k(1 - 1/(t^2)) - 1}{t\sqrt{k}} \leq n.$$

This implies that for any $t > 0$,

$$f(n) \leq \lg n + (1/2)\lg\lg n + C$$

where C is some constant depending on t, and it happens that the optimal choice of t is $\sqrt{3}$.

References

[AS08] N. Alon and J. H. Spencer. *The probabilistic method.* Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons Inc., Hoboken, NJ, third edition, 2008. With an appendix on the life and work of Paul Erdős.