CS265/CME309: Randomized Algorithms and Probabilistic Analysis
Lecture #4: Markov and Chebyshev's Inequalities, and a Sampling-Based Median Algorithm

Gregory Valiant*, updated by Mary Wootters

October 6, 2022

1 Introduction

In the next few classes, we will cover some of the core tools of probability—Markov and Chebyshev's inequalities, moment generating functions, and Chernoff bounds—and discuss several useful randomized algorithms whose analyses use these tools. Markov's inequality, Chebyshev's inequality, and Chernoff bounds, all provide bounds on the probability that a real-valued random variable deviates significantly from its expectation. In the case of Markov's inequality, all we will need to know about the random variable in question is its expectation (or a bound on its expectation) and that the random variable only takes non-negative values. Chebyshev's inequality will apply to any real-valued random variable, provided we can bound its variance. Chernoff bounds, as we will see next class, in some sense leverage the behavior of the "higher moments", $\mathbf{E}[X^t]$ for $t > 2$. For many "nice" random variables, Chernoff bounds will give stronger tail bounds than Chebyshev's inequality , which give stronger bounds than Markov's inequality. Still, we might see in class this week, there are random variables for which Markov's inequality and Chebyshev's inequalities are tight.

These tail bounds are used throughout the analysis of randomized algorithm, and are often applied to the random variable representing the runtime. (Since the runtime is non-negative, Markov's inequality will always apply.) As we have already seen, in many cases all we want from our algorithm is a constant probability of success, since we can always repeat the algorithm a small number of times to decrease the probability of failure exponentially. Hence, in such cases, even the very simple Markov's inequality can yield an acceptable bound.

2 Markov and Chebyshev

Markov's inequality applies to a real-valued random variable that only takes non-negative values, and gives a "tail bound" in terms of the expectation:

Proposition 1 (Markov's Inequality)**.** *Letting X denote a real-valued random variable that only takes non-negative values, for any $\alpha > 0$,*

$$\Pr[X \geq \alpha] \leq \frac{\mathbf{E}[X]}{\alpha}.$$

Proof. The expectation of X can we expressed as

$$\mathbf{E}[X] = \mathbf{E}[X|X \geq \alpha] \cdot \Pr[X \geq \alpha] + \mathbf{E}[X|X < \alpha] \cdot \Pr[X < \alpha].$$

Since, by assumption, $X \geq 0$, the second term is at least 0. Assuming for the sake of contradiction that $\Pr[X \geq \alpha] > \frac{\mathbf{E}[X]}{\alpha}$, the contribution of this first term alone would *exceed* $\alpha \frac{\mathbf{E}[X]}{\alpha} = \mathbf{E}[X]$, which is a contradiction. □

I often think of Markov's inequality as the following equivalent statement: For a random variable X that takes non-negative values and any $c > 0$, $\Pr[X \geq c\mathbf{E}[X]] \leq \frac{1}{c}$.

Chebyshevs inequality states that for any real-valued random variable, the probability the random variable is more than c standard deviations from its expectation is at most $1/c^2$:

Proposition 2 (Chebyshev's Inequality)**.** *Letting X denote a real-valued random variable. For any $c > 0$,*

$$\Pr[|X - \mathbf{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}] \leq \frac{1}{c^2}.$$

Proof. The proof follows by applying Markov's inequality to the random variable $Y = (X - \mathbf{E}[X])^2$. Y is a real-valued random variable, and because it is a square, it only takes non-negative values, and hence we may apply Markov's inequality. First, observe that the quantity we care about can be related to a statement about Y:

$$\Pr[|X - \mathbf{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}] = \Pr\left[Y \geq c^2\mathbf{Var}[X]\right] = \Pr\left[Y \geq c^2\mathbf{E}[Y]\right] \leq \frac{1}{c^2}.$$

In the second equality, we used the fact that $\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[Y]$, and the final inequality is from applying Markov's inequality. □

Note: Sometimes people re-write Chebyshev's inequality as: $\Pr[|X - \mathbf{E}[X]| \geq \alpha] \leq \frac{\mathbf{Var}[X]}{\alpha^2}$, which is equivalent to the statement above as can be seen by plugging in $\alpha = c\sqrt{\mathbf{Var}[X]}$.

Remark 3. *The names "Markov's Inequality" and "Chebyshev's Inequality" are standard, though are historically inaccurate. Chebyshev was Markov's instructor, and both Markov's inequality and Chebyshev's inequality were known to Chebyshev around the time that Markov was born (1856). (Further complicating historical matters, Chebyshev's inequality was first formulated by Bienaymé, though the first proof was likely due to Chebyshev.)*

2.1 Illustrative Examples of Markov's and Chebyshev's Inequalities

Example 4. *Let X denote the number of "heads" flipped as the result of n independent tosses of a fair coin. $\mathbf{E}[X] = n/2$, and since $X \geq 0$, we may apply Markov's inequality. For example $\Pr[X \geq \frac{3n}{4}] \leq \frac{n/2}{3n/4} = \frac{2}{3}$. This is a pretty bad bound on this quantity, especially for large n.*

Using the fact that the variance of a sum of independent random variables is the sum of the variances, and the quick calculation that for a single coin toss, Y that lands heads with probability p, the variance is $\mathbf{Var}[Y] = \mathbf{E}[(Y - \mathbf{E}[Y])^2] = p(1-p)^2 + (1-p)(0-p)^2 = p(1-p)$, we have that the $\mathbf{Var}[X] = n\frac{1}{2}(1-\frac{1}{2}) = n/4$. Hence Chebyshev's inequality yields

$$\Pr[X \geq \frac{3n}{4}] < \Pr[|X - \mathbf{E}[X]| \geq \frac{n}{4}] \leq \frac{\mathbf{Var}[X]}{(n/4)^2} = \frac{4}{n}.$$

This is a much better bound than the $2/3$ probability we got from Markov's inequality, though its still far from the inverse exponential in n we might expect based on the central limit theorem. As we'll see next class, Chernoff bounds will *give the inverse exponential that reflects the actual probability that we have such a significant deviation from the expectation.*

Example 5. *Consider the "Coupon Collector" setting: each day, we get one coupon, drawn uniformly at random from a set of n types of coupons. Let X denote the number of days until we have at least one of every type. Letting X_i denote the number of days we spend waiting for the $i+1$st type of coupon (after we already have the i'th), we have $X = \sum_{i=0}^{n-1} X_i$. Once we have i coupons, the probability we get our $i+1$st on each day is $\frac{n-i}{n}$, as there are $n-i$ new types of coupons that we would be happy with. Hence X_i is distributed as a geometric random variable, with parameter $\frac{n-i}{n}$, and $\mathbf{E}[X_i] = \frac{n}{n-i}$. By linearity of expectation, $\mathbf{E}[X] = \sum_{i=0}^{n-1} \mathbf{E}[X_i] = n\sum_{j=1}^{n} \frac{1}{j} = n\log n + O(n)$.*

What is the probability that we haven't seen all n types of coupons after $2n\log n$ days? By Markov's inequality, this is at most $\frac{1}{2} + o(1)$, where the $o(1)$ term vanishes for large n and is from the $O(n)$ error term in our calculation of the expectation.

To apply Chebyshev's inequality, we leverage the fact that the variance of a geometric random variable with parameter p is $\frac{1-p}{p^2}$, and use the fact that X_i and X_j are independent for $i \neq j$, to calculate the variance of X as the sum of the variances of the X_i's:

$$\mathbf{Var}[X] = \sum_{i=0}^{n-1} \mathbf{Var}[X_i] = \sum_{i=0}^{n-1} \frac{1 - \frac{n-i}{n}}{(n-i)^2/n^2} = n\sum_{i=0}^{n-1} \frac{i}{(n-i)^2}.$$

To apply Chebyshev's inequality, we just need an upper bound on the variance, so at the risk of losing a constant factor, we can replace the numerator in the above with n, and simplify:

$$\mathbf{Var}[X] < n^2 \sum_{i=0}^{n-1} \frac{1}{(n-i)^2} = n^2 \sum_{j=1}^{n} \frac{1}{j^2} < n^2(\frac{\pi^2}{6}),$$

where we used the fact that $\sum_{i\geq 1} 1/i^2 = \pi^2/6$. Okay, so we're ready to apply Chebyshev's inequality:

$$\begin{aligned} \Pr[X \geq 2n\log n] &\leq \Pr[|X - \mathbf{E}[X]| \geq n\log n + O(n)] \leq \frac{Var[X]}{n^2\log^2 n + o(n^2\log^2 n)} \\ &< \frac{n^2(\frac{\pi}{6})}{n^2\log^2 n + o(n^2\log^2 n)} = O(1/\log^2 n). \end{aligned}$$

At the very least, this bound does go to zero as n gets large, though we could still do even better. The probability we have not seen a specific coupon by time t is $(1-1/n)^t < e^{-t/n}$. By a union bound, the probability that we haven't seen all the coupons by time t is at most $ne^{-t/n}$. If we plug in $t = 2n\log n$, we get a probability of $ne^{-2\log n} = 1/n$, which is quite a bit better than what Markov's or Chebyshev's inequality was giving us.

At this point, you might be wondering, why would we ever want to use Markov's or Chebyshev's inequalities? In the previous two examples, they don't seem very good. The reason is that they are very general—and in some cases, they can be tight.

Markov's inequality can be a good idea when the random variables in question don't have (easily or nicely) bounded variance. Here's a simple example of when this can happen.

Example 6. *Let X be a random variable so that $\Pr[X = k] = c/k^3$ for $k = 1, 2, 3, \ldots$, where $c = \left(\sum_{k=1}^{\infty} 1/k^3\right)^{-1}$ is a normalizing constant. Then we have*

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} \frac{ck}{k^3} = c\sum_{k=1}^{\infty} k^{-2} = \frac{c\pi^2}{6},$$

while

$$\mathbf{E}[X^2] = \sum_{k=1}^{\infty} \frac{ck^2}{k^3} = c\sum_{k=1}^{\infty} k^{-1},$$

which diverges. Thus, $\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}X)^2$ is unbounded. In this case, Chebyshev's inequality is out of the question, while Markov's inequality at least tells us something:

$$\Pr[X \geq t] \leq \frac{c\pi^2}{6t}.$$

One place where Chebyshev's inequality can be a good idea is when you are dealing with a sum of *pairwise independent* random variables. We say that $X_1, \ldots, X_n$ are pairwise independent if X_i and X_j are independent for all $i \neq j$. In this case, it's very easy to apply Chebyshev's inequality to $\sum_i X_i$, since

$$\mathbf{Var}\left[\sum_i X_i\right] = \sum_i \mathbf{Var}[X_i].$$

Example 7. *Consider the family of hash functions $\mathcal{H} = \{h_{\mathbf{a},b} : \mathbf{a} \in \mathbb{Z}_p^k, b \in \mathbb{Z}_p$, where p is a prime, and where $h_{\mathbf{a},b} : \mathbb{Z}_p^k \to \mathbb{Z}_p$ is given by $h_{\mathbf{a},b}(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} + b \mod p$. (You may have seen universal hash families similar to this one in CS161). Given vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from the universe Z_p^k, we will choose a random hash function $h \in \mathcal{H}$ (that is, we will choose a random $\mathbf{a}$ and b), and we will hash these vectors into p buckets using h. The hope is that about $1/p$ of the items end up in each bucket. We can use Chebyshev's inequality to bound the probability that this doesn't happen.*

Fix $y \in \mathbb{Z}_p$, and let X_i be the indicator function that is 1 if $h(\mathbf{x}_i) = y$. You can check that $\mathbf{E}[X_i] = 1/p$ and $\mathbf{Var}[X_i] = \frac{1}{p}\left(1 - \frac{1}{p}\right)$.

Claim 8. *The X_i's are pairwise independent.*

Proof. To do this, we'll compute the probability that $h(\mathbf{x}_i) = y$ and $h(\mathbf{x}_j) = y$. We'd like this to be $(1/p)^2$. We can write these equations as matrices by:

$$\begin{bmatrix} - & - & \mathbf{x}_i & - & - \\ - & - & \mathbf{x}_j & - & - \end{bmatrix} \cdot \begin{bmatrix} | \\ \mathbf{a} \\ | \end{bmatrix} + \begin{bmatrix} b \\ b \end{bmatrix} = \begin{bmatrix} y \\ y \end{bmatrix}.$$

Suppose first that neither of $\mathbf{x}_i, \mathbf{x}_j$ are zero. Since they are distinct, the first $2 \times k$ matrix is full rank, which means that if we choose a random vector $\mathbf{x} \in \mathbb{Z}_p^k$, the matrix vector product

$$\begin{bmatrix} - & - & \mathbf{x}_i & - & - \\ - & - & \mathbf{x}_j & - & - \end{bmatrix} \cdot \begin{bmatrix} | \\ \mathbf{a} \\ | \end{bmatrix}$$

is uniformly distributed in $\mathbb{Z}_p^2$. (Here, we are using the fact that linear algebra "works" over $\mathbb{Z}_p$, which implicitly uses the fact that p is prime...you don't need to understand the details for this class, but feel free to work them out as a fun exercise!). Adding the vector $(b, b)^T$ doesn't change the fact that our output is uniformly random, so the left hand side is uniformly random. Thus, in this case, the probability that we get (y, y) as output is $1/p^2$.

Now consider the case where $\mathbf{x}_i$ (say) is equal to zero. In this case, $\mathbf{x}_j \neq 0$, since they are distinct. Then the matrix vector product

$$\begin{bmatrix} - & - & \mathbf{x}_i & - & - \\ - & - & \mathbf{x}_j & - & - \end{bmatrix} \cdot \begin{bmatrix} | \\ \mathbf{a} \\ | \end{bmatrix}$$

is equal to $(0, z)^T$, where z is a uniformly random element of $\mathbb{Z}_p$. After adding $(b, b)^T$, we get $(b, b + z)^T$. Since b and z are independent, and b is uniform, this vector is again uniform in $\mathbb{F}_p^2$, and so the probability that it is equal to (y, y) is again $1/p^2$. □

Thus, we can apply Chebyshev's inequality to bound the probability that any given bucket y has too many elements in it to bound the probability that any given bucket y has too many elements in it.

$$\Pr\left[\sum_i X_i \geq \left(\frac{1}{p} + \epsilon\right) n\right] \leq \frac{\mathbf{Var}\left[\sum_i X_i\right]}{(\epsilon n)^2} = \frac{\sum_i \mathbf{Var}[X_i]}{(\epsilon n)^2} = \frac{(1/p)(1 - 1/p)}{\epsilon^2 n}.$$

Union bounding over all p buckets shows that

$$\Pr\left[\exists y \in \mathbb{Z}_p \textit{ so that there are more than } \left(\frac{1}{p} + \epsilon\right) n \textit{ elements in bucket } y\right] \leq \frac{1 - 1/p}{\epsilon^2 n} \leq \frac{1}{\epsilon^2 n}.$$

Thus, as $n \to \infty$, it's very likely that about a $1/p$ fraction of the $\mathbf{x}_i$'s will land in any bucket.

Note: At this point, we are done with the material covered in the recorded mini-lectures. Section 3 gives an application of these tools that we will work through in class. The notes below are meant for reference after Class 4.

3 Sampling-Based Median Algorithm

We now describe a sampling based algorithm for computing the median of a set of n numbers. In CS161 you might have seen a different randomized algorithm for computing the median (a recursive algorithm that resembles quick-sort with a random pivot), and also a deterministic $O(n)$ comparison

algorithm. Our sampling-based algorithm will be extremely simple, and will require only $\frac{3}{2}n + o(n)$ pairwise comparisons—a better constant factor than the other algorithms, and shockingly close to the trivial lower-bound of n comparisons. For simplicity, we describe the algorithm in the case that all the numbers are distinct, and n is odd, though a simple modification will work beyond this setting.

The idea that really enables the following sampling based median algorithm is the following: we can take a relatively small sample of the elements such that 1) the sample is small enough that we can thoroughly analyze/inspect the sample, without spending much time, and 2) the information we get from the sample lets us fine-tune how we interact with the full set of numbers, allowing us to efficiently pluck out a smallish set of candidate medians that we will then closely inspect. (And, as you might guess, there are a number of other algorithmic problems for which analogs of this sampling-based approach are successful.)

Algorithm 9. SAMPLING-BASED MEDIAN

Given list S of n distinct numbers:

1. *Sample a list $R = r_1, \ldots, r_{n^{3/4}}$ by independently drawing r_i uniformly at random from the set S.*
2. *Sort list R. Henceforth, we assume that $r_1 \leq r_2 \leq \ldots r_{n^{3/4}}$.*
3. *Define $a = r_{n^{3/4}/2-\sqrt{n}}$ and $b = r_{n^{3/4}/2+\sqrt{n}}$.*
4. *We now form a list of candidate medians: for each element $x \in S$, compare it to a and b, and form the list $T = \{x \in S : a \leq x \leq b\}$, counting the number of elements that are less than a and the number that are greater than b. Let $N_{<a}$ denote the number of elements of S that are less than a, and $N_{>b}$ denote the number that are greater than b.*
5. *If median$(S) \in T$ (i.e. $N_{<a}, N_{>b} \leq n/2$) and $|T| < 4n^{3/4}$, sort the list T, and return the ith smallest element of T, where $i = (n+1)/2 - N_{<a}$, otherwise return ''FAIL''.*

Theorem 1. *If the algorithm does not output FAIL, then it correctly outputs the median. The probability the algorithm returns FAIL is at most $O(1/n^{1/4})$ (and hence we can repeat until success without any significant increase in expected runtime) and the algorithm performs at most $2n + o(n)$ pairwise comparisons.*

Before proving the above theorem, I wanted to note that this runtime can actually be improved to $3n/2 + o(n)$ expected pairwise comparisons if in Step 4, we first compare each number to a, and then only compare it to b if the number was greater than a. This takes one or two lines of reasoning similar to the reasoning that will occur in the proof below.

Proof. The first statement in the theorem is true by construction. Step 2 and 5 require sorting two lists each of size at most $4n^{3/4}$, which requires $O(n^{3/4} \log n) = o(n)$ pairwise comparisons. Step 4 trivially requires $\leq 2n$ pairwise comparisons.

The meat of the proof is bounding the probability of failure. Let m denote the true median of set S. The algorithm succeeds provided the following three conditions hold, where the first two conditions together guarantee that the true median lies between a and b, and hence will be in the set T.

1. $|\{r_i \in R : r_i < m\}| < \frac{n^{3/4}}{2} + \sqrt{n}$.
2. $|\{r_i \in R : r_i > m\}| < \frac{n^{3/4}}{2} + \sqrt{n}$.
3. $|T| \leq 4n^{3/4}$.

By symmetry, the probability of the first two conditions are equal to each other. To bound the probability that the first condition is not met, let X_i denote the $0/1$ random variable that is 1 if the ith element selected to be in set R is less than the median, m. Condition 1 holds unless $\sum X_i \geq n^{3/4}/2 + \sqrt{n}$. Since $E[X_i] \leq 1/2$, this probability is at most the probability that a fair coin flipped $n^{3/4}$ times lands heads more than $n^{3/4}/2 + \sqrt{n}$ times. Since $\mathbf{Var}[\sum X_i] \leq n^{3/4}/4$, by Chebyshev's inequality we have that this probability is at most

$$\Pr\left[\left|\sum_{i=1}^{n^{3/4}} X_i - n^{3/4}/2\right| \geq \sqrt{n}\right] \leq \frac{\mathbf{Var}[\sum X_i]}{\sqrt{n}^2} \leq \frac{n^{3/4}/4}{n} = O(1/n^{1/4}).$$

We now bound the probability that the third condition is not satisfied. The third condition is satisfied if a is not one of the $n/2-2n^{3/4}$ smallest elements of S, and if b is not one of the $n/2-2n^{3/4}$ largest elements of S. [If both of those conditions are true then there will be at most $4n^{3/4}$ elements of S between a and b.] By symmetry, these two probabilities are equal, so we will just focus on bounding the probability that a *is* one of the smallest $n/2 - 2n^{3/4}$ elements of S. The probability of this is the probability that set R ends up with more than $n^{3/4}/2 - \sqrt{n}$ elements from the smallest $n/2 - 2n^{3/4}$ elements of S. Letting X_i denote the event that the ith element selected to be in R is in this smallest batch of elements of S, we have that $\Pr[X_i = 1] = (n/2 - 2n^{3/4})/n = \frac{1}{2} - \frac{2}{n^{1/4}}$. Hence $\mathbf{E}[\sum_{i=1}^{n^{3/4}} X_i] = n^{3/4}/2 - 2\sqrt{n}$, and the probability that we end up with too many such small elements in R is at most:

$$\begin{aligned}
\Pr[a \text{ too small}] &= \Pr\left[\sum_{i=1}^{n^{3/4}} X_i \geq n^{3/4}/2 - \sqrt{n}\right] \\
&\leq \Pr\left[\left|\sum X_i - \mathbf{E}[\sum X_i]\right| \geq \sqrt{n}\right] \\
&\leq \frac{n^{3/4}/4}{\sqrt{n}^2} \\
&= O(1/n^{1/4}),
\end{aligned}$$

where the last inequality is from applying Chebyshev's inequality. □