Advanced Algorithms

Lecture 8: Concentration Bounds

Lecturer: Nathan Klein

1 Concentration Bounds

Having a good grasp on probability is key to algorithm design for two reasons:

1. Algorithms can often be shown to work on *random* inputs. In the next lecture, we will show that for the Max Cut problem (which is APX-Hard), there is an algorithm that gets within a $1 - o(1)$ factor of optimal with high probability on random graphs. To prove statements of this form we will need a few probabilistic tools.

2. Often, using randomness inside an algorithm can help simplify its analysis (as we saw for Max Cut) or, in some cases, even allow us to find polynomial time algorithms where no deterministic counterpart is known. In particular, for some problems such as polynomial identity testing, there is a polynomial time randomized algorithm that works with high probability on *every* input, and no deterministic polynomial time algorithm is known. We will explore this a little, but if you're interested in learning more check out CS 537 – Randomized Algorithms.

Beyond algorithms, probability is also an incredibly important tool for understanding data. We will use what we learn today in the lectures on data, as we will generally think of a dataset as a collection of samples from some unknown distribution we want to compute some statistics on.

Today we'll focus on understanding **concentration bounds**. Roughly speaking, this says that if a random variable X is the sum of many independent or near-independent factors, it is unlikely to deviate much from its expectation.

1.1 Probability Warmup

Variance represents how much a random variable deviates from its mean: a high variance means it's often far away from it, and low variance means it's usually quite close. Formally,

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

Lemma 1.1. *Let X be a Bernoulli random variable such that $X = 1$ with probability p and $X = 0$ otherwise. Then,* $\mathrm{Var}(X) = p(1-p)$.

Proof. Let's compute $\mathbb{E}[X]$ first. This is

$$\mathbb{E}[X] = \mathbb{P}[X = 0] \cdot 0 + \mathbb{P}[X = 1] \cdot 1 = p$$

So,

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$$ □

1.2 Polling

An important application of concentration bounds is to understand the number of people you need to ask in a poll to get a good estimate. Let's say I want to understand what fraction of Americans support a particular candidate in an election. Typically, this is done by polling: you call people up and you ask them. There are some huge complications that come with polling in practice, but for now, let's say that whoever you call picks up and tells the truth. So, let's think of polling as a random process. In each of n calls, we'll pick a uniformly random voter, and let the random variable X_i be 1 if the person prefers that candidate and 0 otherwise. Then, $\mathbb{E}[X]$ is exactly the probability a random person will vote for your candidate.

We all have an intuitive sense that from only a small number of samples, we can get a good sense of what $\mathbb{E}[X]$ is. And we also all know something is wrong with Kirk in this scene from Gilmore Girls:

> **Kirk:** I took it upon myself to poll the town, and I think you're gonna be pretty happy with the results.
>
> **Sookie:** We are?
>
> **Kirk:** Jackson is solidly in the lead.
>
> **Sookie:** Already?
>
> **Lorelai:** We just started bugging people.
>
> **Kirk:** Well, I modeled my poll after the Gallup poll. The Gallup poll uses a sample of 1,005 voters to represent the 280 million people of the United States. Using that logic, the correct sampling size of the town of Stars Hollow would be 0.002. Rounding that up means one person needs to be polled, so I picked me.
>
> **Lorelai:** You polled yourself?
>
> **Kirk:** I was right there. Seemed like a perfect opportunity.

But what, exactly, is his error? Why does the Gallup poll only need 1,000 voters and why does Kirk need more than 1?

1.3 The Law of Large Numbers

The beginning of the story on the concentration of measure is the Law of Large Numbers.

Theorem 1.2 (Law of Large Numbers)**.** *Let $X_1, \dots, X_n$ be a sequence of independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$ finite. Then for any $\epsilon > 0$, where $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean, $\lim_{n\to\infty} \mathbb{P}[|\hat{\mu} - \mu| \geq \epsilon] = 0$. In other words,* ***the sample mean always converges to the true mean****.*

But the question is: how fast? The Gallup poll says that for 280 million people, 1,000 samples is enough. Kirk says that for 10,000 people,[1] 1 sample is enough. There's even some sort of logic to Kirk's thought: in a town of 100 people, it's not even possible to poll 1,000 people, so you might hope that the smaller the town, the fewer people you need to poll.

Weirdly, that's not really true. Of course, it is true that polling everyone is always enough. But the tools we'll use will not depend on the population size at all!

[1] Apparently the population of Stars Hollow, although this means Kirk's math is off by a factor of 10 or so.

1.4 Markov's Inequality

The most basic tool in our arsenal is Markov's inequality.

Theorem 1.3 (Markov's Inequality)**.** *Let X be a non-negative random variable with $\mathbb{E}[X] = \mu$. Then,*

$$\mathbb{P}[X \geq k] \leq \frac{\mu}{k}$$

or equivalently

$$\mathbb{P}[X \geq k \cdot \mu] \leq \frac{1}{k}$$

Proof. Since X is non-negative, for any $c \in \mathbb{R}$ we have:

$$\mathbb{E}[X] \geq c \cdot \mathbb{P}[X \geq c]$$

Rearranging gives the result. □

What's nice about Markov is it does require any assumptions on X besides non-negativity. Nothing is needed about its variance and it is not required to be the sum of independent random variables, like in the law of large numbers.

Lemma 1.4. *Let A be a randomized algorithm which halts within t steps in expectation. Then, with probability at least 99%, it halts within $100t$ steps.*

Proof. Let X be a random variable indicating the number of steps A takes to halt. Then, $\mathbb{E}[X] = t$. So:

$$\mathbb{P}[X \geq 100 \cdot t] \leq \frac{1}{100}$$ □

Markov is nice for a lot of quick computations like this, and you'll see more on your homework. It's also useful in proving stronger concentration inequalities, as we'll see in a second. But it's not very good for polling. Let's try to see it in action and see what goes wrong. Say we have independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ like in the law of large numbers, and let $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Now say we let n be large, like 1,000. Suppose that $\mu = \frac{1}{2}$. Then, we might hope that we are close to μ with good probability. Using Markov, what's the chance we get $\hat{\mu} \geq \frac{3}{4}$?

We know that $\mathbb{E}[\hat{\mu}] = \frac{1}{2}$, so plugging it in,

$$\mathbb{P}\left[\hat{\mu} \geq \frac{3}{4}\right] \leq \frac{\mathbb{E}[\hat{\mu}]}{\frac{3}{4}} = \frac{1}{2} \cdot \frac{4}{3} = \frac{2}{3}$$

That's pretty terrible. That says that with probability $\frac{2}{3}$, our error can be as high as 25%, which on a poll is useless. Furthermore, notice that this bound *does not depend on the number of people polled*. So we should know something is wrong here.

1.5 Chebyshev's Inequality

The next useful bound we'll learn is Chebyshev. This works for any random variable (as opposed to Markov, which requires non-negativity) and says:

Theorem 1.5 (Chebyshev's Inequality)**.** *For any random variable X with $\mathbb{E}[X] = \mu$,*

$$\mathbb{P}[|X - \mu| \geq k] \leq \frac{\text{Var}(X)}{k^2}$$

or equivalently,

$$\mathbb{P}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

where $\sigma = \sqrt{\text{Var}(X)}$ is the standard deviation.

Proof.

$$\mathbb{P}[|X - \mu| \geq k] = \mathbb{P}\left[(x - \mu)^2 \geq k^2\right] \leq \frac{\mathbb{E}\left[(x - \mu)^2\right]}{k^2} = \frac{\text{Var}(X)}{k^2}$$

where we applied Markov. □

Chebyshev can be used to prove the law of large numbers assuming finite variance: that if $X_1, \ldots, X_n$ is a sequence of independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$ finite and $\text{Var}(X_i)$ finite, then for any ϵ, $\lim_{n\to\infty} \mathbb{P}[|\hat{\mu} - \mu| \geq \epsilon] = 0$, where $\hat{\mu} = \sum_{i=1}^{n} X_i$.

Proof. Since $X_1, \ldots, X_n$ are pairwise independent,

$$\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n}\text{Var}(X_i)$$

Now applying Chebyshev,

$$\mathbb{P}[|\hat{\mu} - \mu| \geq \epsilon] \leq \frac{\text{Var}(X_i)}{\epsilon^2 n}$$

and the limit of this as $n \to \infty$ is 0 as desired. □

Let's try polling again. We can bound $\text{Var}(X_i) = p(1-p) \leq \frac{1}{4}$, so $\text{Var}(\hat{\mu}) = \frac{1}{n^2} \cdot \frac{n}{4} = \frac{1}{4n}$. Therefore,

$$\mathbb{P}[|\hat{\mu} - \mu| \geq \epsilon] \leq \frac{1}{4n\epsilon^2}$$

Let's see if this is reasonable. Let's aim for a polling error of around 5%, or $\epsilon = 0.05$, using the Gallup poll of 1,000 people. Then this is $\frac{1}{4{,}000 \cdot 0.05^2} = \frac{1}{10}$. That's certainly something, but being more than 5% off with that large of a probability is still not great.

1.6 Hoeffding's Inequality

So, finally: Hoeffding. This has a very strong requirement on the random variable X: it must be the sum of *mutually independent* random variables $X_1, \ldots, X_n$. Luckily, that's the case for polling.

Theorem 1.6 (Hoeffding's Inequality for Binary Random Variables)**.** *Let $X_1, \ldots, X_n$ be mutually independent binary random variables and $X = \sum_{i=1}^{n} X_i$, $\mathbb{E}[X] = \mu$. Then,*

$$\mathbb{P}[X - \mu \geq t] \leq e^{\frac{-2t^2}{n}}$$

and so,

$$\mathbb{P}[|X - \mu| \geq t] \leq 2e^{\frac{-2t^2}{n}}$$

Hoeffding can be proved using Markov's inequality as well and the mutual independence assumption. Now let's see Hoeffding in action for our polling problem. Let $n = 1000$. If we aim for 5% error, then we want $t = 50$. Then this says:

$$\mathbb{P}[|X - \mu| \geq 50] \leq 2e^{\frac{-2 \cdot 50^2}{1000}} = 2e^{-5} \approx 0.013$$

So, we are within 5% error with probability around 99%. That's more like it. I find the following rule of thumb helpful to know:

Lemma 1.7 (Sampling Rule of Thumb)**.** *To get an estimate of μ within ϵ additive error with probability 99%, choose $n = \frac{3}{\epsilon^2}$ samples.*

Proof.

$$\mathbb{P}[|X - \mu| \geq \epsilon n] \leq 2e^{\frac{-2\epsilon^2 n^2}{n}} = 2e^{-2\epsilon^2(3/\epsilon^2)} = 2e^{-6} < 0.01$$ □

There is also a more general version of Hoeffding you may find useful:

Theorem 1.8 (Hoeffding's Inequality for General Random Variables)**.** *Let $X_1, \ldots, X_n$ be mutually independent binary random variables so that $a_i \leq X_i \leq b_i$ with probability 1. Let $X = \sum_{i=1}^{n} X_i$, $\mathbb{E}[X] = \mu$. Then,*

$$\mathbb{P}[X - \mu \geq t] \leq e^{\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

and so,

$$\mathbb{P}[|X - \mu| \geq t] \leq 2e^{\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

I also recommend taking a look at Chernoff bounds to get a better sense of the full landscape of concentration bounds. But for this course, Hoeffding will suffice.

1.7 Why Polling is Really Hard

Given these tools, you may say: why are we so bad at predicting the outcome of elections in the real world? There are a number of complicating factors:

1. People lie when you poll them.

2. People often don't answer the phone. Now, suddenly, the μ you are estimating is the probability someone prefers a particular candidate conditioned on the fact that they pick up the phone for an unknown number. That might be different from the true probability someone votes for your candidate.

3. Finally, the probability someone votes may actually depend on which candidate they prefer. And it is hard to predict whether someone will vote or not. So, even knowing the probability someone would vote for a particular candidate *if they end up voting* may be useless!

This shows the importance of getting **very good data**. As long as you are sure the data is reliable, you can estimate the mean with pretty good accuracy in as little as 1000 samples and incredible accuracy with 10,000. But often getting that data is next to impossible. Pollsters just aren't that certain, and they usually have way more than 10,000 data points. So what's going on here? It's the three complicating factors listed above (and probably some others I haven't thought of).