CS 388R: Randomized Algorithms, Fall 2021 September 1, 2021

Lecture 2: Concentration Inequalities

Prof. Eric Price *Scribe: Ryan Chhong, Michael Jaber*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Overview

In the last lecture, we gave a brief introduction to randomized algorithms as well as presented a simple randomized algorithm of Karger for computing min-cuts in a graph.

In this lecture, we will investigate concentration inequalities as they will be essential to our analysis of randomized algorithms in future lectures. In particular, we will discuss Chernoff bounds as well as various heuristic methods for understanding probabilities.

2 An Excercise on Expectation

We will start by flipping a fair coin 1000 times. We will repeat this experiment a large amount of times. Let X represent the number of heads.

The Gaussian Distribution is defined as the following.

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The Central Limit Theorem states that if we examine $X = X_0 + X_1 + ...$, where all X_i are independent, we will converge to an approximately Gaussian/Normal distribution.

$$\mu = E[X] = \sum_i E[X_i] = \frac{n}{2} = 500$$
$$\sigma^2 = Var(X) = \sum_i Var[X_i] = \sum_i (E[(X_i - E[X_i])^2]) = \sum_i \frac{1}{4} = \frac{n}{4} = 250$$
$$\sigma = \sqrt{\sigma^2} \approx 16$$

Furthermore, we define the Gaussian Tail Bound as the following. Given that $X \sim \mathcal{N}(\mu,\sigma^2)$

$$Pr[X > \mu + t] < e^{-\frac{t^2}{2\sigma^2}}$$

Based on the bounds established, we can now provide estimates for the following probabilities

$$Pr[X = 500] \geq \frac{.68}{32} \approx 2\%$$
$$Pr[X > 510] \approx 30\%$$
$$Pr[X > 600] < e^{-18}$$
$$Pr[X > 900] < e^{-260}$$

The last two bounds were found using the Gaussian Tail Bound. For the first bound, note that around 68% of the area under the curve is within 1 standard deviation. There are 32 values that occur within one standard deviation, so the average is .68/32. Since the event $X = 500$ occurs with the highest probability, it must be at least the average. The second bound is approximated by noticing that if we use the previous assumption that every event within one standard deviation occurs with equal probability, then the 10 events corresponding to $X = 500$, $X = 501$, ..., $X = 509$ account for about $10 \cdot 2\% = 20\%$ of the mass. Since $X \geq 500$ occurs with probability 1/2, then the the probability that $X > 510$ should be about $50\% - 20\% = 30\%$.

For the third and fourth bounds, we use the Central Limit Theorem to assume that this distribution is very close to a Gaussian, then apply the known tail bounds for Gaussians. Note that going from within k standard deviations to ck standard deviations results in a decay of $1/c^2$ in the exponent. These tails are getting small very fast as we approach $X = 1000$.

3 Chernoff Bounds

The Chernoff Bound can be though as a formal Central Limit Theorem-type tail bound for finite n. Consider $X = X_1 + X_2 + ... + X_n$, for independent random variables $X_i \in [0, 1]$. There are two forms of this bound. Depending on the situation, only one of these bounds might give something useful.

3.1 Additive Chernoff Bounds

$$Pr[X > \mu + t] \leq e^{\frac{-2t^2}{n}}$$
$$Pr[X < \mu - t] \leq e^{\frac{-2t^2}{n}}$$
$$Pr[|X - \mu| > t] \leq 2e^{\frac{-2t^2}{n}}$$

3.2 Multiplicative Chernoff Bound

$$Pr[X > (1+\epsilon)\mu] \le e^{-\frac{min\{\epsilon,\epsilon^2\}}{3}\mu}$$
$$Pr[X > (1+\epsilon)\mu] \le e^{-\frac{\epsilon^2}{2+\epsilon}\mu}$$
$$Pr[X > (1+\epsilon)\mu] \le e^{-((1+\epsilon)log(1+\epsilon)-\epsilon)\mu} \qquad \text{(Bennett's Inequality)}$$
$$Pr[X < (1+\epsilon)\mu] \le e^{-\frac{\epsilon^2}{2}\mu}$$
$$Pr[|X-\mu| > \epsilon\mu] \le 2e^{-\frac{min\{\epsilon,\epsilon^2\}}{3}\mu}$$

4 Example: Dice

Suppose we had a 2000-sided die that we rolled 1000000 times and we wished to count the number of 1s that came up. Let us examine the estimated probabilities using the Central Limit Theorem.

$$p = \frac{1}{2000}$$
$$E[X] = 1000000(\frac{1}{2000}) = 500$$
$$\sigma^2 = Var(X) = \sum_i Var(X_i) = np(1-p) \approx np = 500$$

$$Pr[X = 500] \ge 1 - 2\%$$
$$Pr[X > 510] \approx 40\%$$
$$Pr[X > 600] < e^{-8}$$
$$Pr[X > 900] < e^{-130}$$

However, what's more interesting is when we apply the Gaussian Tail Bound, we find that we get $Pr[X > 600] < e^{-\frac{1}{50}}$. This is an extremely useless bound, as we find that the probability of this is less than around 98%. The issue with the Central Limit Theorem/Gaussian Bounds is now shown: σ^2 is treated as being $\sqrt{n}$, when in reality, it should be np.

5 Application: Correctness Amplification

Suppose that we had a randomized algorithm that succeeded around 60% of the time, which takes time $T(n)$ to run. We will use Chernoff bounds to show that we can modify this algorithm to have success probability $1-\delta$. We will do so at the cost of running the algorithm multiple times. Assume that there exists only 1 correct answer.

To generate such an algorithm, we will simply take the randomized algorithm we have and repeat it k times and output the majority answer. Therefore, our algorithm is correct when a majority of the runs of our algorithm were successful. Let X_i denote the random variable that is 1 when the i^{th} run of the algorithm is incorrect, and 0 otherwise. Therefore, the majority answer is incorrect when $\sum X_i > \frac{k}{2}$. THerefore, we can bound our error probability by bounding $Pr[\sum X_i > \frac{k}{2}]$. Let $X = \sum X_i$.

Now consider $E[X]$. By linearity of expectation, $E[X]$ is at most $\frac{4}{10}k$. Applying Chernoff's bound gives

$$Pr[X \geq \frac{k}{2}] \leq Pr[X \geq \mu + \frac{k}{10}] \leq e^{\frac{\frac{-2k^2}{100}}{k}} = e^{-\frac{k}{50}}$$

Setting $k = 50\log(\frac{1}{\delta})$ gives us an error probability of δ, so we succeed with probability $1 - \delta$