CS49/249 (Randomized Algorithms), Spring 2021 : Lecture 4

Topic: Estimation Algorithms, Chebyshev, Chernoff-Hoeffding Bound

Disclaimer: These notes have not gone through scrutiny and in all probability contain errors. Please discuss in Piazza/email errors to deeparnab@dartmouth.edu

---

- One extremely important application of randomized algorithms is in *estimation*. Imagine there is a statistic stat you are interested in. For concreteness, let us assume stat is the fraction of Hanover's population who are vegetarians. To calculate stat exactly we must ask everyone in Hanover of their dietary preference. However, the exact statistic is often not important, and what one really needs is an *estimate* est of the statistic. And polling or *random sampling* is the way to obtain such estimates. In this lecture, we will set the formal definitions of what a "good" estimate is, and what one needs to obtain one.

- For any statistic stat, an estimate est is a *random variable* often obtained via a randomized algorithm. We say est is an ***unbiased estimate*** if

$$\mathbf{Exp}[\mathsf{est}] = \mathsf{stat} \qquad \text{(Unbiased Estimate)}$$

  For the statistic fraction of vegetarians in Hanover, here is an unbiased estimate: sample an individual from the Hanover population *uniformly at random*[1] and ask them about their dietary preference. If they say vegetarian, set $\mathsf{est} = 1$, otherwise set $\mathsf{est} = 0$. If we denote the population of Hanover as n, then there are $\mathsf{stat} \cdot n$ many vegetarians, and $(1 - \mathsf{stat})n$ many non-vegetarians. Thus, we get

$$\mathbf{Exp}[\mathsf{est}] = \frac{\mathsf{stat} \cdot n}{n} \cdot 1 + \frac{(1-\mathsf{stat})n}{n} \cdot 0 = \mathsf{stat}$$

- Although the above estimate was unbiased, all of us really can feel that it was a pretty bad estimate. Indeed, it was always 0 or 1. The following definition gives a much more nuanced notion of the quality of an estimate.

  Definition 1 ((ε, δ)-estimate.)**.** An estimate est is a multiplicative (ε, δ)-estimate of a statistic stat if

$$\mathbf{Pr}[\mathsf{est} \notin (1 \pm \varepsilon) \cdot \mathsf{stat}] \leq \delta$$

  That is, with probability at least $(1-\delta)$ (which is high if δ is small), the estimate satisfies $(1-\varepsilon)\mathsf{stat} \leq \mathsf{est}(1+\varepsilon)\mathsf{stat}$.

  There is another notion of (ε, δ)-additive estimate often used when the statistic stat is bounded, as in the case of fraction of vegetarians where $\mathsf{stat} \in [0, 1]$.

  Definition 2 ((ε, δ)-additive estimate.)**.** An estimate est is an additive (ε, δ)-estimate of a statistic stat if

$$\mathbf{Pr}[\mathsf{est} \notin \mathsf{stat} \pm \varepsilon] \leq \delta$$

  That is, with probability at least $(1-\delta)$ (which is high if δ is small), the estimate satisfies $(\mathsf{stat} - \varepsilon) \leq \mathsf{est} \leq (\mathsf{stat} + \varepsilon)$.

[1] We are completely ignoring **how** one samples an individual uniformly at random from the population. We are also assuming that they answer our poll. All these are super-important considerations and are real bottlenecks to good polling. Unfortunately, it is not in the scope of this course.

- ***Variance of an estimator.*** The main result we want to establish today is to show how to obtain (ε, δ)-estimates from unbiased estimates. It is an extremely important fact which will be used many times in this course (and life).

Theorem 1. [Boosting Theorem or the Median-of-Means Theorem.]

Let $\widehat{\mathsf{est}}$ be an *unbiased* estimator of some statistic stat. Then, one can obtain an (ε, δ)-muliplicative estimate of stat using K *independent* samples of $\widehat{\mathsf{est}}$, where

$$K = \frac{C\,\mathbf{Var}[\widehat{\mathsf{est}}]}{\left(\mathbf{Exp}[\widehat{\mathsf{est}}]\right)^2} \cdot \frac{1}{\varepsilon^2} \cdot \ln\left(\frac{2}{\delta}\right)$$

where C is some constant. Consequently, one can obtain an (ε, δ)-additive estimate of stat using K' independent samples of $\widehat{\mathsf{est}}$, where $K' = \frac{C\,\mathbf{Var}[\widehat{\mathsf{est}}]}{\varepsilon^2} \cdot \ln\left(\frac{2}{\delta}\right)$

Thus, what really determines the quality of an unbiased estimate is its variance. If the variance, or more precisely the variance-to-mean-square ratio (which is the square of the standard-deviation-to-mean ratio). If that ratio is small, then one can obtain a good multiplicative estimate with not "too many" runs of the unbiased estimator.

Let's go back to the example of stat being the fraction of vegetarians in a population. We saw an unbiased estimate $\widehat{\mathsf{est}}$. What is it's variance? A calculation gives that

$$\mathbf{Var}[\widehat{\mathsf{est}}] = \mathsf{stat}(1 - \mathsf{stat})$$

Thus, if one obtains $K = O\left(\frac{(1-\mathsf{stat})\ln(1/\delta)}{\mathsf{stat}\cdot\varepsilon^2}\right)$ many independent samples, or in other words polls that many people independently uniformly at random, then one can come with a very good estimate. As you can see, the number grows as stat becomes smaller: this is to be expected, if the number of vegetarians are small, we will need to sample more to detect them. On the other hand, if we only want to estimate the fraction of vegetarians to within *additive* 1%, say (so $\varepsilon = 0.01$), then the number of samples we need is $O(\mathsf{stat}(1 - \mathsf{stat}) \cdot \ln(1/\delta))$. Note, and this is something often counter-intuitive to many, this ***doesn't*** depend on the population size. The same *number* of people need be sampled[2] even from New York City.

- ***First obtain an $(\varepsilon, \frac{1}{3})$-estimate using Chebyshev.*** We will prove Theorem 1 in two steps. Both steps will also introduce tools which are much more important than the theorem itself.

The first idea is to take a bunch of unbiased estimates and creating a more refined estimate by taking the mean. Here is the algorithm.

```
1: procedure BETTER-ESTIMATOR(s):▷ Assumes access to an unbiased estimate est.
2:     Sample s independent unbiased estimates est_0, est_1, ..., est_s.
3:     Return est' := (1/s) Σ_{i=1}^{s} est_i.
```

Claim 1. Let $s \geq \frac{3\,\mathbf{Var}[\widehat{\mathsf{est}}]}{\varepsilon^2\left(\mathbf{Exp}[\widehat{\mathsf{est}}]\right)^2}$. Then, est' returned by BETTER-ESTIMATOR(s) satisfies

$$\mathbf{Pr}[\mathsf{est}' \notin (1 \pm \varepsilon)\mathsf{stat}] \leq \frac{1}{3}$$

[2] assuming both cities have the similar fraction of vegetarians.

To prove this, we will use *Chebyshev inequality* which is the most general purpose deviation inequality.

Theorem 2. (Chebyshev's Inequality)

Let X be any random variable. Then for any $t > 0$, we have

$$\mathbf{Pr}[|X - \mathbf{Exp}[X]| \geq t] \leq \frac{\mathbf{Var}[X]}{t^2}$$

Proof. We first note that

$$\mathbf{Pr}[|X - \mathbf{Exp}[X]| \geq t] = \mathbf{Pr}[(X - \mathbf{Exp}[X])^2 \geq t^2]$$

Then we notice that $D := (X - \mathbf{Exp}[X])^2$ is a non-negative random variable, and therefore we can apply Markov's inequality on it to get

$$\mathbf{Pr}[|X - \mathbf{Exp}[X]| \geq t] = \mathbf{Pr}[D \geq t^2] \leq \frac{\mathbf{Exp}[D]}{t^2} = \frac{\mathbf{Var}[X]}{t^2}$$

□

Proof of Claim 1.

- We first calculate the variance of est'. We use two simple facts: $\mathbf{Var}[cX] = c^2\,\mathbf{Var}[X]$, and the linearity of variance[3] for *independent* random variables $X_1, \ldots, X_t$, $\mathbf{Var}[\sum_{i=1}^t X_i] = \sum_{i=1}^t \mathbf{Var}[X_i]$. Using these, we obtain

$$\mathbf{Var}[\mathsf{est}'] = \mathbf{Var}\left[\frac{1}{s}\sum_{i=1}^{s} \widehat{\mathsf{est}}_i\right] = \frac{\mathbf{Var}[\widehat{\mathsf{est}}]}{s}$$

  Also, recall that $\mathbf{Exp}[\mathsf{est}'] = \frac{1}{s}\sum_{i=1}^{s} \mathbf{Exp}[\widehat{\mathsf{est}}_i] = \mathsf{stat}$ since each $\widehat{\mathsf{est}}_i$ is an unbiased estimate of stat. Thus, taking the mean of s independent samples is still unbiased but the variance decays. Useful nugget of information.

- Next, we apply Chebyshev on est' to get

$$\begin{aligned}\mathbf{Pr}[\mathsf{est}' \notin (1 \pm \varepsilon)\mathsf{stat}] &= \mathbf{Pr}\left[|\mathsf{est}' - \mathbf{Exp}[\mathsf{est}']| \geq \varepsilon\,\mathbf{Exp}[\widehat{\mathsf{est}}]\right] \\ &\leq \frac{\mathbf{Var}[\mathsf{est}']}{\varepsilon^2\left(\mathbf{Exp}[\widehat{\mathsf{est}}]\right)^2} = \frac{\mathbf{Var}[\widehat{\mathsf{est}}]}{s\varepsilon^2\left(\mathbf{Exp}[\widehat{\mathsf{est}}]\right)^2} \leq \frac{1}{3}\end{aligned}$$

□

- ***The Chernoff Bound.*** Claim 1 is great, but the "error probability" is still $\frac{1}{3}$ which we would like to ramp down to δ. If you stare at the proof of the claim, you probably notice that if instead of using $s \geq 3 \cdot \left(\frac{\mathbf{Var}}{\varepsilon^2\,\mathbf{Exp}^2}\right)$ samples, we used $\frac{1}{\delta} \cdot \left(\frac{\mathbf{Var}}{\varepsilon^2\,\mathbf{Exp}^2}\right)$ many we would have our (ε, δ)-estimate. However, the next technique shows how we can actually have a muuuuuuuch better dependence on δ. This is really important if we want our confidence levels really high.

[3] I guess this is also an important fact which you should recall from CS30/Discrete Probability.

To explain this best, we introduce one of the *most used* theorems in TCS: the Chernoff bound, or the Azuma-Bernstein-Chernoff-Hoeffding-Stein... bound[4]. Indeed, we will dedicate the rest of this lecture to understanding this bound, and continue with the proof of Theorem 1 in the next class.

Theorem 3 (Chernoff Bounds)**.** Let $X_1, X_2, \ldots, X_n$ be ***independent*** *Bernoulli* random variables with each $X_i \in \{0, 1\}$. Let $X = \sum_{i=1}^{n} X_i$. Then, for any $\varepsilon \in (0, 1)$,

$$\mathbf{Pr}[X \leq (1-\varepsilon)\,\mathbf{Exp}[X]] \leq e^{-\frac{\varepsilon^2\,\mathbf{Exp}[X]}{2}} \tag{LT}$$

and

$$\mathbf{Pr}[X \geq (1+\varepsilon)\,\mathbf{Exp}[X]] \leq e^{-\frac{\varepsilon^2\,\mathbf{Exp}[X]}{3}} \tag{UT1}$$

For the "upper tail", that is for "larger" deviations, we have when $1 \leq t \leq 4$, we have the following (changing ε to t so as to underscore that the deviation is big)

$$\mathbf{Pr}[X \geq (1+t)\,\mathbf{Exp}[X]] \leq e^{-\frac{t^2\,\mathbf{Exp}[X]}{4}} \tag{UT2}$$

and for $t > 4$ (really large), we have

$$\mathbf{Pr}[X \geq (1+t)\,\mathbf{Exp}[X]] \leq e^{-\frac{t \ln t\,\mathbf{Exp}[X]}{2}} \tag{UT3}$$

Remark: Important*: Equations (UT1) to (UT3) hold with* all $\mathbf{Exp}[X]$ *occurences replaced by* any *upper bound* $\mathbf{Exp}[X] \leq U$.

Remark: *Note the asymmetry in the denominators in the exponent. For most applications this is not important. What are qualitatively more important : (a) things are in the* exponent*, (b) the expectation of the sum shows up in the exponent irrespective of the number of terms, (c) the X_i's are independent but not necessarily identical, and (d) the dependence on ε in the exponent is quadratic.*

It is instructive to compare with what Chebyshev gives us. To apply Chebyshev, set $t = \varepsilon\,\mathbf{Exp}[X]$ to get $\mathbf{Pr}[X \geq (1+\varepsilon)\,\mathbf{Exp}[X]] \leq \frac{\mathbf{Var}[X]}{\varepsilon^2\,\mathbf{Exp}^2[X]}$. Unless $\mathbf{Var}[X] \leq \varepsilon^2\,\mathbf{Exp}^2[X]$, the RHS is a trivial bound. The Chernoff-bound always gives something non-trivial.

- ***Some consequences.***
  - *Number of heads.* Suppose we toss n fair coins and let X be the number of heads we observe. We know $\mathbf{Exp}[X] = \frac{n}{2}$. But what really can we say about the "range" of X? In particular, what is the probability see t more heads than expected, that is, $\mathbf{Pr}[X \geq \frac{n}{2} + t]$? Or rather for what value of t does this get below probability δ?

[4]There are many such inequalities which bound how far the sum of random variables can deviate from its expectations. Entire books have been written on this (pdf copies of many of these are freely available).

This random variable (number of heads in n fair coin tosses) has a name : it is called the *Binomial Distribution* or *Binomial Random Variable*. And the exact answer to the above question is:

$$\mathbf{Pr}[X \geq \frac{n}{2} + t] = \frac{1}{2^n} \sum_{j \geq \frac{n}{2}+t} \binom{n}{j}$$

and you can sweat hard and figure it out. But it still doesn't answer after what t will the probability be at most δ. Is this $t = \Theta(1)$ or $\Theta(\log n)$ or $\Theta(n)$ or what? We now see that the Chernoff bound will lead to the answer.

To this end, define n independent random variables where $X_i = 1$ if the ith coin toss if head, and $X_i = 0$ otherwise. Next, observe $X = \sum_{i=1}^{n} X_i$ and that $\mathbf{Exp}[X] = \frac{n}{2}$. Apply Theorem 3 with ε such that $\varepsilon \, \mathbf{Exp}[X] = t$, that is, $\varepsilon = \frac{2t}{n}$.

$$\mathbf{Pr}[X \geq \frac{n}{2} + t] \leq e^{-\frac{4t^2}{n^2} \cdot \frac{n}{6}} \quad = \quad e^{-2t^2/3n}$$

Therefore, if $t \geq \sqrt{\frac{3n}{2} \cdot \ln\left(\frac{1}{\delta}\right)}$, the RHS probability is $\leq \delta$. The answer is $\Theta(\sqrt{n})$. To put in some numbers, the chances of observing $\frac{n}{2} + 20\sqrt{n}$ is less than 10^{-100}.

Exercise: *Repeat a similar calculation to figure for what t we get $\mathbf{Pr}[X \leq \frac{n}{2} - t] \leq \delta$.*

Remark: *By the way, the variance of X is $\frac{n}{4}$ and thus the standard deviation of X is $\frac{\sqrt{n}}{2}$. Indeed, in this case the Chernoff bound tells us that the probability of seeing the number of heads more than c std-deviations away drops like $e^{-O(c^2)}$. Chebyshev on the other hand says that the probability drops like $\frac{1}{c^2}$.*

– *The Fraction of Vegetarians.* For the simple estimation problem of fraction of vegetarians in Hanover, the Chernoff bound already solves the problem without needing Theorem 1. Indeed, let us pick k samples $\widehat{\mathsf{est}}_1, \ldots, \widehat{\mathsf{est}}_k$ and return $\mathsf{est} := \frac{\sum_{i=1}^{k} \widehat{\mathsf{est}}_i}{k}$. Let $X_i := \widehat{\mathsf{est}}_i$ and let $X = \sum_{i=1}^{k} X_i$. We get $\mathbf{Exp}[X] = \mathsf{stat} \cdot k$, and thus,

$$\mathbf{Pr}[\mathsf{est} \geq (1+\varepsilon)\mathsf{stat}] \quad = \quad \mathbf{Pr}[X \geq (1+\varepsilon)\mathsf{stat} \cdot n] \leq e^{-\frac{\varepsilon^2}{3} \cdot \mathsf{stat} \cdot k}$$

Therefore, if we choose k such that $k \cdot \mathsf{stat} \cdot \frac{\varepsilon^2}{3} \geq \ln(2/\delta)$, we would get that $\mathbf{Pr}[\mathsf{est} \geq (1+\varepsilon)\mathsf{stat}] \leq \frac{\delta}{2}$. The application of the other Chernoff bound would give $\mathbf{Pr}[\mathsf{est} \leq (1-\varepsilon)\mathsf{stat}] \leq \frac{\delta}{2}$. Which would imply $\mathbf{Pr}[\mathsf{est} \notin (1 \pm \varepsilon)\mathsf{stat}] \leq \delta$, that is, est is an (ε, δ)-multiplicative estimate. This gives us $k \geq \frac{3\ln(2/\delta)}{\mathsf{stat} \cdot \varepsilon^2}$ samples are sufficient (as we had found before, up to constants).

Learning Tidbits:

- Algorithm Design: *When you design an unbiased estimate $\widehat{\mathsf{est}}$, it is the $\mathbf{Var}[\widehat{\mathsf{est}}]/\mathbf{Exp}^2[\widehat{\mathsf{est}}]$ which defines the quality. If this is "small", then with that many samples times $\frac{1}{\varepsilon^2}\ln(1/\delta)$, one can get an (ε, δ)-estimate* est.
- Analysis: ***Chernoff Bounds**! (Hard to overstate its importance)*