CS 49/149: 21st Century Algorithms (Fall 2018): Lecture 10

Date: 16th October, 2018

Topic: Randomized Estimation Algorithms. Median-of-Average trick.

Scribe: Maryam Negahbani

Disclaimer: These notes have not gone through scrutiny and in all probability contain errors. Please email errors to maryam@cs.dartmouth.edu.

---

1 Preliminaries

Definition 1 (Discrete Random Variable)**.** Random variable X is a function defined on some discrete sample space Ω that takes any $x \in \Omega$ to a value $\mathbf{Pr}[X = x]$ with $0 \leq \mathbf{Pr}[X = x] \leq 1$ such that $\sum_{x \in \Omega} \mathbf{Pr}[X = x] = 1$.

In this lecture, we use "random variable" to refer to a discrete random variable.

Definition 2 (Independent Random Variables)**.** Two random variables X and Y are independent if $\mathbf{Pr}[X = x, Y = y] = \mathbf{Pr}[X = x]\mathbf{Pr}[Y = y]$.

Definition 3 (Expectation/Mean of a Random Variable)**.** The expected value of a random variable X is denoted by both $\mathbb{E}[X]$ and μ_x and defined as:

$$\mu_x = \mathbb{E}[x] = \sum_{x \in \Omega} x\mathbf{Pr}[X = x]$$

Fact 1 (Linearity of Expectation)**.** For any constant α and **any** two random variables X and Y:

- $\mathbb{E}[\alpha X] = \alpha\mathbb{E}[X]$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

Applying the second formula on any **finite** set of random variables, say $X_1, X_2, \ldots, X_k$ gives:

$$\mathbb{E}[\sum_{i=1}^{k} X_i] = \sum_{i=1}^{k} \mathbb{E}[X_i]$$

Note that the X_i's do not need to be independent.

Next, we define a measure for the expected deviation of a random variable from its expectation. The first thing that comes to mind is to use $X - \mathbb{E}[X]$ as X's "deviation" from its mean and then take its expectation. But the problem is that $X - \mathbb{E}[X]$ is positive for some values of $x \in \Omega$ and negative for some other ones which "cancel out" in average and cause $\mathbb{E}[X - \mathbb{E}[X]] = 0$. The next idea would be to use $|X - \mathbb{E}[X]|$ as our deviation but working with absolute values is hard. So we use $(X - \mathbb{E}[X])^2$ instead, which leads to the following definition:

Definition 4 (Variance of a Random Variable)**.** Variance of a random variable X is denoted by both $\text{Var}[X]$ and σ_X^2 and defined as:

$$\sigma_X^2 = \text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Remark: σ_X is called the standard deviation of X.

Fact 2. For any constant α, and **pair-wise independent** random variables X and Y we have:

- $\text{Var}[\alpha X] = \alpha^2 \text{Var}[X]$
- $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$

Fact 3. $\text{Var}[X] \geq 0$ so $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$.

Recall that we already knew this by Jensen's inequality since $f(x) = x^2$ is a convex function.

2 Concentration Inequalities

Next, we see two inequalities that basically say there is a "low" probability that a random variable gets a value "too far" from its expectation.

Theorem 1 (Markov's Inequality)**.** For any **non-negative** random variable X and any $t > 0$ we have:

$$\mathbf{Pr}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$$

Proof.

$$\begin{aligned}
\mathbb{E}[X] = \sum_x x\mathbf{Pr}[X = x] &= \sum_{x \geq t} x\mathbf{Pr}[X = x] + \sum_{x < t} x\mathbf{Pr}[X = x] \\
&\geq \sum_{x \geq t} x\mathbf{Pr}[X = x] \\
&\geq t \sum_{x \geq t} \mathbf{Pr}[X = x] \\
&= t\mathbf{Pr}[X \geq t]
\end{aligned}$$

The first inequality uses the fact that X is non-negative. □

Another way to look at Markov's inequality is by setting $t := \alpha\mathbb{E}[X]$ for $\alpha > 0$ which gives:

$$\mathbf{Pr}[X \geq \alpha\mathbb{E}[X]] \leq \frac{1}{\alpha}$$

Note that this upper-bound is non-trivial only if $\alpha > 1$. So for example, if you wanted to upper-bound the probability of $X \geq \mathbb{E}[X]/2$, Markov's inequality is not useful. This brings us to the next inequality:

Theorem 2 (Chebyshev's Inequality)**.** For **any** random variable X and $t > 0$ we have:

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

Proof.

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

As $(X - \mathbb{E}[X])^2$ is a non-negative random variable, we can use Markov's inequality to bound the above probability:

$$\mathbf{Pr}[(X - \mathbb{E}[X])^2 \geq t^2] \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2} = \frac{\text{Var}[X]}{t^2}$$

□

Again, to look at Chebyshev's inequality differently, replace t by $\alpha\text{Var}[X]$ for some $\alpha > 0$:

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

Qualitatively, if you move X away from $\mathbb{E}[X]$ by some factor of α, the drop in probability is quadratic with respect to this α.

Another benefit of Chebyshev's inequality over Markov's is that in the former, the bound is proportional to $1/\alpha^2$ while in the latter, it is only $1/\alpha$ which means that Chebyshev's bound is tighter with respect to α.

3 Randomized Estimation

Suppose there is a population of M people and each person has either watched "Kill Bill" or not. You are asked to compute the ratio of people that have watched Kill Bill (number of people that have watched it to M). Of course, for computing the exact solution, one has to check each person individually and ask if they have watched Kill Bill. However, this can be infeasible if M is large. Then what if you were asked to approximate it? That is, assume the actual solution is f^* and you are supposed to find a Z such that $|Z - f^*| \leq \epsilon$ for some parameter $\epsilon > 0$. This is still not enough, because for small ϵ and large M, you still have to check almost every individual.

Now, what if you are allowed to make "bad predictions" sometimes (but not too often). That is, there is a parameter $\delta > 0$ and you could make random predictions as long as $\mathbf{Pr}[|Z - f^*| > \epsilon] \leq \delta$. This Z is a random variable and since it is used for estimation purpose here, we call it a "random estimator".

Definition 5 ((ϵ, δ)-Estimator)**.** Random variable Z is an (ϵ, δ)-estimator for f^* if:

$$\mathbf{Pr}[|Z - f^*| \geq \epsilon] \leq \delta$$

Definition 6 (Unbiased Estimator)**.** An estimator Z for f^* is unbiased if $\mathbb{E}[Z] = f^*$

If Z is an unbiased estimator, it gives a correct estimation "in average". But we still need to bound the probability of Z giving a far off estimation. Using Chebyshev's inequality, we get:

$$\mathbf{Pr}[|Z - f^*| \geq \epsilon] = \mathbf{Pr}[|Z - \mathbb{E}[Z]| \geq \epsilon] \leq \frac{\text{Var}[Z]}{\epsilon^2}$$

Thus:

Fact 4. To find an (ϵ, δ)-estimator, it suffices to find an **unbiased** estimator Z and show:

$$\frac{\text{Var}[Z]}{\epsilon^2} \leq \delta$$

The following is an example of an unbiased estimator:

TRY1

- Sample an individual uniformly at random
- $Z := \begin{cases} 1 & \text{if watched Kill Bill} \\ 0 & \text{otherwise} \end{cases}$
- Return Z

The Z returned by Try1 is in $\{0, 1\}$. These Boolean random variables are called "indicator random variables" and in this case, Z is an indicator of watching Kill Bill.

Fact 5. If $X \in \{0, 1\}$ is indicator of event S we have:

- $\mathbb{E}[X] = 1 \times \mathbf{Pr}[X = 1] + 0 \times \mathbf{Pr}[X = 0] = \mathbf{Pr}[S]$
- $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X] - \mathbb{E}[X]^2 = \mathbf{Pr}[S] - \mathbf{Pr}[S]^2 = \mathbf{Pr}[S](1 - \mathbf{Pr}[S]) \leq \frac{1}{4}$

Thus for our Z we have:

$$\mathbb{E}[Z] = \mathbf{Pr}[\text{individual has watched Kill Bill}] = \frac{\text{No. of people that watched Kill Bill}}{M} = f^*$$

Which means Z is indeed unbiased. As for the variance:

$$\text{Var}[Z] = f^*(1 - f^*) \leq \frac{1}{4}$$

Substituting the variance into the equation of fact(4) we get: $1/4 \leq \epsilon^2\delta$ which is totally useless. We would want a "knob" on the variance of our estimator, so we could adjust it relative to the parameters we get (a constant variance is not immediately helpful). Nevertheless, in the next section, we will see a general purpose technique for generating a low-variance estimator from an unbiased estimator.

3.1 The "Average" Trick

TRY2 (Input: Parameter k)

- Run Try1 k times to obtain $Z_1, Z_2, \ldots, Z_k$
- $Y := \frac{1}{k} \sum_i Z_i$
- Return Y

By fact(1):

$$\mathbb{E}[Y] = \frac{1}{k}\sum_i \mathbb{E}[Z_i] = f^*$$

And by fact(2):

$$\text{Var}[Y] = \frac{1}{k^2}\sum_i \text{Var}[Z_i] \leq \frac{1}{4k} \leq \frac{1}{k}$$

So one thing to remember about the average trick is that:

Given an unbiased estimator, the average trick reduces its variance by a factor of $1/k$.

This k is the knob we were looking for. After we substitute $\text{Var}[Y]$ in fact(4) we find the value of k to be:

$$k = \frac{1}{\epsilon^2\delta}$$

This Y is our first (ϵ, δ)-estimator. Interestingly, Y is independent of M. That is, no matter if your population is small or large, you only need k many samples to give a good estimator. This seems to be counter intuitive but where is the contradiction coming from? Note that, the situation would have been different if we wanted to estimate the **number** of people that have watched Kill Bill (instead of their ratio). In that case, you would want to find Y' such that for given ϵ' and δ:

$$\mathbf{Pr}[|Y' - f^*M| \geq \epsilon'] \leq \delta \Rightarrow \mathbf{Pr}[|\frac{Y'}{M} - f^*| \geq \frac{\epsilon'}{M}] \leq \delta$$

So if we just substitute Y'/M with Y and replace ϵ'/M by ϵ we get our previous problem. But now, the ϵ depends on M and so is k:

$$k = \frac{1}{\epsilon^2\delta} = \frac{M^2}{\epsilon'^2\delta}$$

In this case, the number of samples taken by Try2 has a quadratic dependence on the size of the population.

3.2 The "Median-of-Average" Trick

In this section, we will see another general purpose method for reducing the dependency of the number of samples on δ from $1/\delta$ to $\log(1/\delta)$.

In the last section, we found Y such that $\mathbf{Pr}[|Y - f^*| \geq \epsilon] \leq \delta$ which basically means:

$$\mathbf{Pr}[Y \geq f^* + \epsilon \text{ or } y \leq f^* - \epsilon] \leq \delta$$

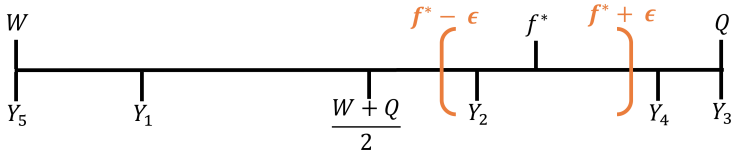
But what if we only care about bounding $\mathbf{Pr}[Y > f^* + \epsilon]$? By repeating Try2 with $k = 2/\epsilon^2$ for t times independently:

$$\mathbf{Pr}[Y_i \geq f^* + \epsilon] \leq \frac{1}{2}, \ \ \forall i \in \{1, 2, \ldots, t\}$$

What is the probability that **all of** $Y_1, Y_2 \ldots, Y_t$ turn out to be bigger than $f^* + \epsilon$? Next, we prove that this probability is small, which means $W := \min_i Y_i$ should be a good estimator.

$$\begin{aligned}
\mathbf{Pr}[W \geq f^* + \epsilon] &= \mathbf{Pr}[\min_i Y_i \geq f^* + \epsilon] \\
&= \mathbf{Pr}[\text{All of } Y_i\text{'s are} \geq f^* + \epsilon] \\
&= \prod_{i=1}^{t} \mathbf{Pr}[Y_i \geq f^* + \epsilon] \\
&\leq \frac{1}{2^t}
\end{aligned}$$

Thus, to get $\mathbf{Pr}[W \geq f^* + \epsilon] \leq \delta$ we just need to set $t = \log(1/\delta)$. Similarly, for the same $Y_1, Y_2 \ldots, Y_t$ we can show that if $Q := \max_i Y_i$, then $\mathbf{Pr}[Q \leq f^* - \epsilon] \leq \delta$. So it seems that the good (ϵ, δ)-estimator for f^* lies somewhere between W and Q. The first idea that comes to mind is to use $(W + Q)/2$ as our estimator. But even when both W and Q are good (i.e. $W \leq f^* + \epsilon$ and $Q \geq f^* - \epsilon$), $(W + Q)/2$ might be far from f^* as shown in the following diagram:



However, the picture suggests that the median might be a good estimator. Because for the median to be outside of the $f^* \pm \epsilon$ interval, it must be that at least half of the Y_i's are out of the interval, and the probability of this event is very low.

TRY2.9 (Input: Parameter t)

- Run Try2 with $k = 2/\epsilon^2$ for $2t + 1$ times independently to obtain $Y_1, Y_2, \ldots, Y_{2t+1}$
- $\hat{f} := \text{median}(Y_1, Y_2, \ldots, Y_{2t+1})$
- Return $\hat{f}$

Now we upper-bound $\mathbf{Pr}[\hat{f} \geq f^* + \epsilon]$:

$$\begin{aligned}
\mathbf{Pr}[\hat{f} \geq f^* + \epsilon] &= \mathbf{Pr}[\text{“some” } t+1 \text{ many of } Y_i\text{'s are} \geq f^* + \epsilon] \\
&= \mathbf{Pr}[\exists S \subset \{1, 2, \ldots, 2t+1\} \text{ s.t. } |S| = t+1 \text{ and } Y_i \geq f^* + \epsilon \; \forall i \in S]
\end{aligned}$$

For any $S \subset \{1, 2, \ldots, 2t+1\}$ of size exactly $t+1$, define E_S to be the event that $Y_i \geq f^* + \epsilon$ for all $i \in S$. Since $\mathbf{Pr}[Y_i \geq f^* + \epsilon] \leq \frac{1}{2}$ for all $i \in S$, we have $\mathbf{Pr}[E_S] \leq \frac{1}{2^{t+1}}$. Assume there are N many of these S's. Then the existence event from the last equality above can be described as the event that E_{S_i} happens for some $i \in \{1, 2, \ldots, N\}$. By the union bound:

$$\mathbf{Pr}[\hat{f} \geq f^* + \epsilon] = \mathbf{Pr}[E_{S_1} \cup E_{S_2} \cup \cdots \cup E_{S_N}] \leq \sum_{i=1}^{N} \mathbf{Pr}[E_{S_i}] \leq \frac{N}{2^{t+1}}$$

Similarly, it can be shown that:

$$\mathbf{Pr}[\hat{f} \leq f^* - \epsilon] \leq \frac{N}{2^{t+1}}$$

So overall, the probability that $\hat{f}$ is out of the $f^* \pm \epsilon$ interval is:

$$\begin{aligned}\mathbf{Pr}[|\hat{f} - f^*| \geq \epsilon] &= \mathbf{Pr}[\hat{f} \geq f^* + \epsilon \text{ or } \hat{f} \leq f^* - \epsilon] \\ &\leq \mathbf{Pr}[\hat{f} \geq f^* + \epsilon] + \mathbf{Pr}[\hat{f} \leq f^* - \epsilon] \\ &\leq \frac{N}{2^t}\end{aligned}$$

But there is a tiny problem here. $N = \binom{2t+1}{t+1} \approx 4^t$ so our upper-bound becomes trivial. There is an easy fix for this problem: instead of setting $k = 2/\epsilon^2$ let $k = 10/\epsilon^2$.

TRY3 (Input: Parameter t)

- Run Try2 with $\mathbf{k = 10/\epsilon^2}$ for $2t + 1$ times independently to obtain $Y_1, Y_2, \ldots, Y_{2t+1}$
- $\hat{f} := \text{median}(Y_1, Y_2, \ldots, Y_{2t+1})$
- Return $\hat{f}$

Now we repeat the previous analysis with this new k:

$$\mathbf{Pr}[Y_i \geq f^* + \epsilon] \leq \frac{1}{10}, \ \ \forall i \in \{1, 2, \ldots, 2t+1\}$$

$$\mathbf{Pr}[\hat{f} \geq f^* + \epsilon] \leq \sum_{i=1}^{N} \mathbf{Pr}[E_{S_i}] \leq \frac{N}{10^{t+1}}$$

$$\mathbf{Pr}[|\hat{f} - f^*| \geq \epsilon] \leq \frac{N}{10^t} \approx \frac{4^t}{10^t} = \left(\frac{2}{5}\right)^t$$

To get the right hand side to be δ we set t to be:

$$t = \log_{\frac{5}{2}} \frac{1}{\delta} = O(\log \frac{1}{\delta})$$

Putting it all together, the algorithm is: Sample $kt = \frac{10}{\epsilon^2} \log_{\frac{5}{2}} \frac{1}{\delta}$ many individuals, divide them into t batches of size k each, compute the average of each batch and return their median. This median-of-average trick basically says:

One can construct an (ϵ, δ)-estimator given $O(V \frac{1}{\epsilon^2} \log \frac{1}{\delta})$ instances of an unbiased estimator with variance V.