Median of Means Proof, More General Chernoff Bounds[1]

- We will complete the proof of the following theorem using the Chernoff bound.

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

Recall, last lecture from an unbiased estimator $\widehat{\mathsf{est}}$, we obtained an estimator est' which was $(\varepsilon, \frac{1}{3})$-approximate. That is,

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

To obtain est', we used $s \geq \frac{3\,\mathbf{Var}[\widehat{\mathsf{est}}]}{\varepsilon^2\left(\mathbf{Exp}[\widehat{\mathsf{est}}]\right)^2}$ independent samples of $\widehat{\mathsf{est}}$.

- ***Boosting using the median.*** To obtain the better estimate, we take a bunch of samples from est', and return the *median* of these estimates.

```
1: procedure MEDIAN-OF-MEANS: ▷ Assumes access to an unbiased estimate est^.
2:     for i = 1 to c independently do: ▷ c is an integer ≥ 36 ln(2/δ).
3:         Run BETTER-ESTIMATOR(s) with s ≥ 3 Var[est^] / (ε^2 (Exp[est^])^2) to obtain est'_i.
4:     Return est ← median of (est'_1, ..., est'_c).
```

Observation 1. The total number of independent samples of $\widehat{\mathsf{est}}$ used is $c \cdot s \geq \frac{108\,\mathbf{Var}[\widehat{\mathsf{est}}]\ln(2/\delta)}{\varepsilon^2\left(\mathbf{Exp}[\widehat{\mathsf{est}}]\right)^2}$

We claim that est is an (ε, δ)-estimate. To prove this, let us define some "bad events". Let X_i be the *indicator* random variable of the event that the ith estimate est'_i is an over-estimate That is,

$$X_i = \begin{cases} 1 & \text{if } \mathsf{est}'_i \geq (1+\varepsilon)\mathsf{stat} \\ 0 & \text{otherwise} \end{cases}$$

Note that X_i's are mutually independent because all the estimates are independent. (1) also gives us $\mathbf{Exp}[X_i] \leq \frac{1}{3}$. Define $X = \sum_{i=1}^{c} X_i$, and thus $\mathbf{Exp}[X] \leq \frac{c}{3}$. Next comes the main observation.

[1] *Lecture notes by Deeparnab Chakrabarty. Last modified : 31st March, 2021*
These have not gone through scrutiny and may contain errors. If you find any, or have any other comments, please email me at deeparnab@dartmouth.edu. Highly appreciated!

Claim 1. If $X < \frac{c}{2}$, then est is *not* an over-estimate, implying, $\mathbf{Pr}[\mathsf{est} \geq (1+\varepsilon)\mathsf{stat}] \leq \mathbf{Pr}[X \geq \frac{c}{2}]$.

Proof. If less than half of the est'_i's are over-estimates, then the median is *not* going to be an over-estimate. □

Now the Chernoff bound gives us (using the upper bound[2] $\frac{c}{3} \geq \mathbf{Exp}[X]$)

$$\mathbf{Pr}\left[X \geq \frac{(1+\varepsilon)c}{3}\right] \leq e^{-\frac{\varepsilon^2 c}{9}} \underbrace{\Rightarrow}_{\varepsilon=\frac{1}{2}} \mathbf{Pr}\left[X \geq \frac{c}{2}\right] \leq e^{-\frac{c}{36}} \underbrace{\leq}_{\text{since } c \geq 36\ln(2/\delta)} \frac{\delta}{2}$$

Thus, we get $\mathbf{Pr}[\mathsf{est} \geq (1+\varepsilon)\mathsf{stat}] \leq \frac{\delta}{2}$ A similar argument (with "underestimate bad-events") gives $\mathbf{Pr}[\mathsf{est} < (1-\varepsilon)\mathsf{stat}] \leq \frac{\delta}{2}$. Therefore, $\mathbf{Pr}[\mathsf{est} \notin (1\pm\varepsilon)\mathsf{stat}] \leq \delta$. Which completes the proof.

Exercise: *Do this argument.*

- ***The General Chernoff Bound.*** The Chernoff bound statement stated last lecture is a corollary to the general statement which looks a little complicated, but over time you just get used to it. Here it is.

Theorem 2 (General Chernoff: the upper tail)**.** 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 $t > 0$, we have

$$\mathbf{Pr}[X \geq (1+t)\,\mathbf{Exp}[X]] \quad \leq e^{-\mathbf{Exp}[X]\cdot g(t)} \tag{2}$$

where $g(t) := (1+t)\ln(1+t) - t$.

When $t \in (0,1)$, then one can show that $g(t) \geq \frac{t^2}{3}$, which in turn implies (UT1) from the last lecture. Similarly, when $t \in [1,4]$, then one can show that $g(t) \geq \frac{t^2}{4}$ implying (UT2), and when $t > 4$, then one can show that $g(t) \geq \frac{t\ln t}{2}$ implying (UT3). All this can also be proved using calculus as well. There is a similar statement for the lower tail.

Theorem 3 (General Chernoff: the lower tail)**.** 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 $0 < t < 1$, we have

$$\mathbf{Pr}[X \geq (1-t)\,\mathbf{Exp}[X]] \quad \leq e^{-\mathbf{Exp}[X]\cdot h(t)} \tag{3}$$

where $h(t) := (1-t)\ln(1-t) + t$.

One obtains (LT) by showing $h(t) \geq \frac{t^2}{2}$

- *Proof Strategy for Theorem 2 and Theorem 3.* The exact proof of the theorems are not as important as the three main ideas behind it. I would recommend you all to understand these ideas and try to fill the details (which requires some calculussing) yourself. If you fail, then refer to any of the bazillion proofs either in the textbook or on the web. However, this is not the only way to prove these: see here for five different ways.

[2]see remark following the statement in the previous notes

a. (Idea 1). Observe that for any parameter $\gamma > 0$, the event $\{X \geq (1+t)\,\mathbf{Exp}[X]\}$ is the same as $\{e^{\gamma X} \geq e^{\gamma(1+t)\,\mathbf{Exp}[X]}\}$. What does this gain us? Think of $e^{\gamma X}$ as a random variable Z. Then Z is *always* positive, and we can apply Markov on it. This will hold *even when* X could have been negative (which doesn't happen if all X_i's are $\{0,1\}$, but this allows us to prove even more general Chernoff bounds.)

b. (Idea 2). What does Markov give? If we let $\theta := e^{\gamma(1+t)\,\mathbf{Exp}[X]}$ (which note is a fixed quantity with nothing random in it), then

$$\mathbf{Pr}[Z \geq \theta] \leq \frac{\mathbf{Exp}[Z]}{\theta} \tag{4}$$

What is $\mathbf{Exp}[Z]$? Here we use

$$\mathbf{Exp}[Z] = \mathbf{Exp}[e^{\gamma X}] = \mathbf{Exp}[e^{\gamma \sum_{i=1}^{n} X_i}] = \mathbf{Exp}[\prod_{i=1}^{n} e^{\gamma X_i}]$$

and now the kicker, by independence (which implies $\mathbf{Exp}[f(X_1)f(X_2)] = \mathbf{Exp}[f(X_1)]\,\mathbf{Exp}[f(X_2)]$),

$$\mathbf{Exp}[Z] = \prod_{i=1}^{n} \mathbf{Exp}[e^{\gamma X_i}] \tag{5}$$

c. (Idea 3). Now, we use that X_i is a Bernoulli random variable with some probability p_i of being 1 and $(1-p_i)$ probability of being 0. And therefore,

$$\mathbf{Exp}[e^{\gamma X_i}] = p_i e^{\gamma} - (1-p_i) = 1 + p_i(e^{\gamma} - 1)$$

And finally we use one bit of analytic trick which is again a pervasive one: we use the inequality $1 + z \leq e^{z}$ for any z, to get

$$\mathbf{Exp}[e^{\gamma X_i}] \leq e^{p_i(e^{\gamma}-1)} \tag{6}$$

Now one plugs (6) into (5) to get an upper bound on $\mathbf{Exp}[Z]$, which is plugged into (4) giving an upper bound on the probability as a function of γ. Then, one uses calculus to find the γ which minimizes this quantity. And after some sweat, one proves Theorem 2. A similar strategy works for Theorem 3. I once again recommend every one doing this exercise from this point on; it is something one needs to do at least once in their lives!

As mentioned above, the similar strategy can actually prove Chernoff style bounds even when the X_i's are not $\{0,1\}$ variables. Indeed, the *same* bounds hold even when each X_i is a random variable whose domain is $[0,1]$. And in the proof strategy above, a little more care (analysis) is needed in (Idea 3) to figure out $\mathbf{Exp}[e^{\gamma X_i}]$; in fact one shows the Bernoulli case is the "worst" (which, if you think about it, makes sense as that has the maximum variance).

Learning Tidbits:

- Algorithm Design: *Taking "median" of independent averages/means* boosts *confidence. This is a general principle.*
- Analysis: *Seeing how Chernoff bounds apply : express random variable of interest as sum of independent random variables, and then it lingers "very close" to the expectation.*