Lecture 07: Concentration Bounds (Basics)

- In the previous lectures, we learned to compute estimates of the expected value of a random variable
- But, is the expected value a good representation of the random variable?
- If the random variable concentrates most of its probability mass around the expected value, then we can consider the expected value to be a good representative of the random variable's behavior
- In the topic of concentration, we shall cover technique to argue the "typicality of a randomized experiment," i.e., say the mean or the median being a good representative of the random variable

Theorem (Markov Inequality)

Let $\mathbb{X}$ be a r.v. over the sample space $\Omega \subseteq \mathbb{R}_{\geqslant 0}$ (i.e., the set of non-negative real numbers), and $\mu = \mathbb{E}[\mathbb{X}]$. Then, the following is true

$$\mathbb{P}[\mathbb{X} \geqslant \lambda\mu] \leqslant \frac{1}{\lambda}$$

This is also equivalent to the expression

$$\mathbb{P}[\mathbb{X} \geqslant \lambda] \leqslant \frac{\mu}{\lambda}$$

Intuition: Suppose λ is large. Then, the probability that $\mathbb{X}$ deposits probability mass further than $\lambda\mu$ is unlikely.
I present the proof only for discrete Ω. The case of non-discrete Ω is similar.

Markov Inequality II

Proof.

If possible let, Markov's inequality is false. That is, there exists $\lambda > 1$ such that $\mathbb{P}\left[\mathbb{X} \geqslant \lambda\mu\right] > \frac{1}{\lambda}$. Then, let us lower-bound the expectation as follows.

$$\begin{aligned}
\mu &= \sum_{i \in \Omega} i\mathbb{P}\left[\mathbb{X} = i\right] \\
&= \sum_{\substack{i \in \Omega \\ i < \lambda\mu}} i\mathbb{P}\left[\mathbb{X} = i\right] + \sum_{\substack{i \in \Omega \\ i \geqslant \lambda\mu}} i\mathbb{P}\left[\mathbb{X} = i\right] \\
&\geqslant 0 + \sum_{\substack{i \in \Omega \\ i \geqslant \lambda\mu}} (\lambda\mu)\mathbb{P}\left[\mathbb{X} = i\right] \\
&= (\lambda\mu)\mathbb{P}\left[\mathbb{X} \geqslant \lambda\mu\right] > (\lambda\mu) \cdot \frac{1}{\lambda} = \mu
\end{aligned}$$

So, we have obtain $\mu > \mu$, a contradiction. □

- We emphasize that for every $\lambda \geqslant 1$, there is a distribution for which the Markov inequality is tight. Let $\mathbb{X}$ be a distribution such that $\mathbb{P}[\mathbb{X}=0] = 1 - \frac{1}{\lambda}$ and $\mathbb{P}[\mathbb{X}=1] = \frac{1}{\lambda}$
- If there exists B such that $\mathbb{P}[\mathbb{X} > B] = 0$, i.e., the sample space of $\mathbb{X}$ bounded above then we can also apply Markov inequality to the random variable $(B - \mathbb{X})$
- Think: How is Markov inequality equivalent to the pigeon-hole principle?
- Think: Consider the following problem. Suppose $(\mathbb{R}, \mathbb{C})$ be a joint-distribution over $\Omega = \{1, \dots, m\} \times \{1, \dots, n\}$.
  Intuitively, think of a matrix with m-rows and n-columns. The r.v. associated probability to the cells.
  Suppose there is a Fun event, and the following holds.

$$\mathbb{P}\left[(\mathbb{R}, \mathbb{C}) \in \mathsf{Fun}\right] \geqslant \varepsilon$$

That is, if you sample a cell according to the joint distribution $(\mathbb{R}, \mathbb{C})$ then the probability of Fun event occurring is at least ε. Consider the following expression.

$$\mathbb{P}\left[(\mathbb{R}, \mathbb{C}) \in \mathsf{Fun} | \mathbb{R} = r\right]$$

This expression represents the probability of the fun event happening if we restrict (condition) on the row $r \in \{1, \ldots, m\}$. Prove the following statement. The probability of sampling $r \sim \mathbb{R}$ such that it has

$$\mathbb{P}\left[(\mathbb{R}, \mathbb{C}) \in \mathsf{Fun} | \mathbb{R} = r\right] \geqslant \alpha$$

is $\geqslant \varepsilon/\alpha$.
Russel Impagliazzo refers to this result as the pigeon-hole principle. The proof of this result is similar to the proof of Markov inequality. It is an excellent exercise to think of ways how this result can be used for derandomization.

Chebyshev's Inequality I

Theorem

For any random variable $\mathbb{X}$ *and* $\mu = \mathbb{E}[\mathbb{X}]$. *Then, we have*

$$\mathbb{P}\left[|\mathbb{X} - \mu| \geqslant t\right] \leqslant \frac{\mathrm{Var}[\mathbb{X}]}{t^2}$$

Proof.

$$\begin{aligned}\mathbb{P}\left[|\mathbb{X} - \mu| \geqslant t\right] &= \mathbb{P}\left[(\mathbb{X} - \mu)^2 \geqslant t^2\right] \\ &\leqslant \frac{\mathbb{E}\left[(\mathbb{X} - \mu)^2\right]}{t^2} \\ &= \frac{\mathrm{Var}[\mathbb{X}]}{t^2}\end{aligned}$$

$\square$

Chebyshev's Inequality II

- In the previous proof, we used the following fact.

$$\mathbb{P}\left[|\mathbb{X}-\mu| \geqslant t\right] = \mathbb{P}\left[\,(\mathbb{X}-\mu)^2 \geqslant t^2\right]$$

In general, for which functions f does the following hold?

$$\mathbb{P}\left[|\mathbb{X}-\mu| \geqslant t\right] = \mathbb{P}\left[\,f\left(|\mathbb{X}-\mu|\right) \geqslant f(t)\right]$$

Answer: For functions f that is monotonically increasing in the sample space of $|\mathbb{X}-\mu|$. This is a very crucial trick that shall be used in various other problems.

- Think: So, we saw that "Markov studied the r.v. $\mathbb{X}$ and got a bound in $1/t$" and "Chebyshev studied the r.v. $\mathbb{X}^2$ and got a bound in $1/t^2$." Can we extrapolate this to use "high powers of $\mathbb{X}$" (technically referred to as <u>moments</u>) to obtain bounds that are "high polynomials in $1/t$?"

- Let $\mathbb{X}$ be r.v. over the sample space $\Omega = \{0, 1\}$ such that $\mathbb{P}[\mathbb{X} = 0] = 1 - p$ and $\mathbb{P}[\mathbb{X} = 1] = p$. Intuitively, think of $\mathbb{X}$ as a coin that says "heads" with probability p and "tails" with probability $1 - p$.
- Note that $\mathbb{E}[\mathbb{X}] = p$
- Suppose we consider n independent and identical samples of $\mathbb{X}$. That is, we consider $\left(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}, \dots, \mathbb{X}^{(n)}\right)$. Intuitively, we perform n independent coin tosses represented by $\mathbb{X}$.
- We are interested in studying the random variable

$$\mathbb{S}_{n,p} := \sum_{i=1}^{n} \mathbb{X}^{(i)}$$

Intuitively, the random variable $\mathbb{S}_{n,p}$ represents the number of heads when we perform n independent tosses of $\mathbb{X}$

Chernoff Bound II

- Note that the sample space of $\mathbb{S}_{n,p}$ is the set $\{0, 1, \ldots, n\}$. Moreover, we can exactly compute the probability that $\mathbb{S}_{n,p} = j$, for any $j \in \{0, 1, \ldots, n\}$. There are $\binom{n}{j}$ ways of choosing which coins output "heads." The probability of those coins outputting "heads" is p^j, and the probability that other coins outputting "tails" is $(1-p)^{n-j}$. So, overall, we have

$$\mathbb{P}\left[\mathbb{S}_{n,p} = j\right] = \binom{n}{j} p^j (1-p)^{n-j}$$

- Note that by linearity of expectation we have

$$\mathbb{E}\left[\mathbb{S}_{n,p}\right] = np$$

Chernoff bound states that $\mathbb{S}_{n,p}$ is <u>very concentrated</u> around its mean.

- Suppose we want to find the probability

$$\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+t)\right]$$

We can directly perform the following sum

$$\sum_{j \geqslant n(p+t)} \binom{n}{j} p^j (1-p)^{n-j}$$

However, this sum is extremely difficult to estimate. Chernoff bound provides an easy way to estimate this sequence.

- How tight is Chernoff bound? One can show that the Chernoff bound is very tight. We shall not explicitly cover the proof for this tightness result. However, one can use the following Stirling's approximation to easily demonstrate the tightness of the Chernoff bound.

$$\sqrt{2\pi n}\left(\frac{n}{\mathrm{e}}\right)^n \exp({}^{1}/_{12n+1}) \leqslant n! \leqslant \sqrt{2\pi n}\left(\frac{n}{\mathrm{e}}\right)^n \exp({}^{1}/_{12n})$$

  There is a more combinatorial technique using "the method of types."

- The random variable $\mathbb{S}_{n,p}$ is also referred to as the "Binomial distribution" $B(n, p)$

Theorem (Chernoff Bound)

$$\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+t)\right] \leqslant \exp(-n\mathrm{D}_{\mathrm{KL}}\left(p+t,p\right)) \leqslant \exp(-2nt^2),$$

where the function $\mathrm{D}_{\mathrm{KL}}\left(\cdot,\cdot\right)$ *is the Kullback–Leibler divergence defined as follows*

$$\mathrm{D}_{\mathrm{KL}}\left(p+t,p\right) := (p+t)\log\frac{p+t}{p} + (1-p-t)\log\frac{1-p-t}{1-p}$$

Chernoff Bound VI

Before we proceed to proving this result, let us interpret this theorem statement.

- Suppose $p = 1/2$ and $t = 1/4$. Then, it is exponentially unlikely that $\mathbb{S}_{n,p}$ surpasses $3n/4$
- Suppose $p = 1/2$ and $t = c/\sqrt{n}$. Then, the probability that $\mathbb{S}_{n,p}$ surpasses $n(p+t)$ is only a constant.
- Note that the last bound $\exp(-2nt^2)$ is not a function of p at all. So, in many contexts, this is not a good estimate to use. However, the bound $\exp(-n\mathrm{D}_{\mathrm{KL}}(p+t, p))$ is a very good estimate for all values of p.

Proof of Chernoff Bound I

- We are interested in upper-bounding the probability

$$\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+t)\right]$$

- Note that, for any positive h, we have

$$\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+t)\right] = \mathbb{P}\left[\exp\left(h\mathbb{S}_{n,p}\right) \geqslant \exp(hn(p+t))\right]$$

  The exact value of h will be chosen later. The intuition of using the $\exp(\cdot)$ function is to consider all moments of $\mathbb{S}_{n,p}$.

- Now, we apply Markov to obtain

$$\mathbb{P}\left[\exp\left(h\mathbb{S}_{n,p}\right) \geqslant \exp(hn(p+t))\right] \leqslant \frac{\mathbb{E}\left[\exp\left(h\mathbb{S}_{n,p}\right)\right]}{\exp(hn(p+t))}$$

Proof of Chernoff Bound II

- Now, we need an observation. Suppose $\mathbb{X}$ and $\mathbb{Y}$ are two independent random variables. Then, we have $\mathbb{E}\left[\exp(\mathbb{X}+\mathbb{Y})\right] = \mathbb{E}\left[\mathbb{X}\right] \cdot \mathbb{E}\left[\mathbb{Y}\right]$. We emphasize that $\mathbb{X}$ and $\mathbb{Y}$ have to be independent to apply this result.
- Note that we have $\mathbb{S}_{n,p} = \sum_{i=1}^{n} \mathbb{X}^{(i)}$. So, we can apply the previous observation to obtain the following result.

$$\frac{\mathbb{E}\left[\exp\left(h\mathbb{S}_{n,p}\right)\right]}{\exp(hn(p+t))} = \frac{\prod_{i=1}^{n} \mathbb{E}\left[\exp(h\mathbb{X}^{(i)})\right]}{\exp(hn(p+t))} = \left(\frac{\mathbb{E}\left[\exp(h\mathbb{X})\right]}{\exp(h(p+t))}\right)^{n}$$

- Recall that $\mathbb{X}$ is a random variable such that $\mathbb{P}\left[\mathbb{X}=0\right] = 1-p$ and $\mathbb{P}\left[\mathbb{X}=1\right] = p$. So, the random variable $\exp(h\mathbb{X})$ is such that $\mathbb{P}\left[\exp(h\mathbb{X})=1\right] = 1-p$ and $\mathbb{P}\left[\exp(h\mathbb{X})=\exp(h)\right] = p$. Therefore, we can conclude that

$$\mathbb{E}\left[\exp(h\mathbb{X})\right] = (1-p) + p\exp(h)$$

Proof of Chernoff Bound III

- Substituting this value, we get

$$\left(\frac{\mathbb{E}\left[\exp(h\mathbb{X})\right]}{\exp(h(p+t))}\right)^n = \left(\frac{(1-p)+p\exp(h)}{\exp(h(p+t))}\right)^n$$

- So, let us take a pause at this point and recall what we have proven thus far. We have shown that, for all positive h, the following holds

$$\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+t)\right] \leqslant \left(\frac{(1-p)+p\exp(h)}{\exp(h(p+t))}\right)^n$$

Proof of Chernoff Bound IV

- To obtain the tightest upper-bound we should use the value of $h = h^*$ that minimizes the right-hand side expression.

$$f(h) = \frac{(1-p) + p\exp(h)}{\exp(h(p+t))} = (1-p)\exp(-h(p+t)) + p\exp(h(1-p-t))$$

- Let us compute $f'(h)$ and solve for $f'(h^*) = 0$. Note that we have

$$f'(h) = -h(1-p)(p+t)\exp(-h(p+t)) + hp(1-p-t)\exp(h(1-p-t))$$

Equating $f'(h) = 0$, we get

$$h^*(1-p)(p+t)\exp(-h^*(p+t)) = h^* p(1-p-t)\exp(h^*(1-p-t))$$

This is equivalent to

$$\exp(h^*) = \frac{(1-p)(p+t)}{p(1-p-t)}$$

Proof of Chernoff Bound V

So, can check that h^* is positive because the $\frac{(1-p)(p+t)}{p(1-p-t)} > 1$. Further, taking $h \to \infty$ we can verify that $f(h) \geqslant f(h^*)$. So, we can conclude that h^* is a minimum. (Otherwise, you can also show that $f''(h^*) > 0$ to conclude that h^* is minimum).

Proof of Chernoff Bound VI

- Now, let us substitute the value of h^* to obtain

$$\begin{aligned}
\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+t)\right] &\leqslant \left(\frac{(1-p) + \frac{(1-p)(p+t)}{(1-p-t)}}{\left(\frac{(1-p)(p+t)}{p(1-p-t)}\right)^{p+t}} \right)^n \\
&= \left(\frac{\frac{1-p}{1-p-t}}{\left(\frac{1-p}{1-p-t}\right)^{p+t}\left(\frac{p+t}{p}\right)^{p+t}} \right)^n \\
&= \left(\frac{\left(\frac{1-p}{1-p-t}\right)^{1-p-t}}{\left(\frac{p+t}{p}\right)^{p+t}} \right)^n \\
&= \left(\left(\frac{1-p}{1-p-t}\right)^{1-p-t} \left(\frac{p}{p+t}\right)^{p+t} \right)^n \\
&= \left(\left(\frac{1-p-t}{1-p}\right)^{1-p-t} \left(\frac{p+t}{p}\right)^{p+t} \right)^{-n}
\end{aligned}$$

Proof of Chernoff Bound VII

- Now, we can use the definition of $\mathrm{D}_{\mathrm{KL}}(p+t, p)$ to obtain

$$\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+t)\right] \leqslant \left(\left(\frac{1-p-t}{1-p} \right)^{1-p-t} \left(\frac{p+t}{p} \right)^{p+t} \right)^{-n}$$
$$= \exp(-n\mathrm{D}_{\mathrm{KL}}(p+t, p))$$

Proof of Chernoff Bound VIII

- We have proven one party of the Chernoff bound. All that remains is to prove that

$$\exp(-n\mathrm{D}_{\mathrm{KL}}(p+t, p)) \leqslant \exp(-2nt^2)$$

Or, equivalently, we need to prove that

$$\mathrm{D}_{\mathrm{KL}}(p+t, p) \geqslant 2t^2$$

That is,

$$(p+t)\log\frac{p+t}{p} + (1-p-t)\log\frac{1-p-t}{1-p} \geqslant 2t^2$$