Lecture 06: Chernoff Bound

Problem Introduction: Vanilla Form I

- Let $\mathbb{X}$ be a coin that outputs 1 (representing heads) with probability p, and outputs 0 (representing tails) with probability $1-p$. The exact probability p is not known. Our objective is to estimate the probability p.
- Informally, our strategy is to toss this coin (independently) n times and report the fraction of outcomes that were heads. We want to understand the probability that this estimate is far from the actual value of p.
- Let $\mathbb{X}^{(1)}, \mathbb{X}^{(2)}, \dots, \mathbb{X}^{(n)}$ represent n independent coin tosses that are identically distributed as the random variable $\mathbb{X}$
- We are interested in studying the random variable

$$\mathbb{S}_{n,p} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)} + \cdots + \mathbb{X}^{(n)}$$

This random variable $\mathbb{S}_{n,p}$ represents the total number of heads in the n coin tosses.

- Formally, given $\varepsilon > 0$, we are interested in computing the probability that

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

  That is, we are interested to prove that the probability of our estimate being "much larger" than p is small.

Approach using Stirling's Approximation I

- Suppose we have seen i heads. We can explicitly compute the probability that $\mathbb{S}_{n,p} = i$ as follows. There are $\binom{n}{i}$ ways to choose the coins that turn up heads. The probability that these coins turn up heads is p^i. Moreover, the probability that the remaining coins turn up tails is $(1-p)^{n-i}$. So, we can claim the following

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

- Threfore, from this result, our desired probability is

$$\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+\varepsilon)\right] = \sum_{i \geqslant n(p+\varepsilon)} \binom{n}{i} p^i (1-p)^{n-i}$$

- For simplicity, let us assume that $n(p+\varepsilon) = k$ is an integer

Approach using Stirling's Approximation II

- **Upper-bound.** We can *prove* that among the elements $\binom{n}{i}p^i(1-p)^{n-i}$, where $i \geqslant k$, the maximum element is one where $i = k$. We can use this observation to upper-bound the probability expression.

$$\begin{aligned}
\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+\varepsilon)\right] &= \sum_{i \geqslant k} \binom{}{} nip^i(1-p)^{n-i} \\
&\leqslant \sum_{i \geqslant k} \binom{n}{k} p^k(1-p)^{n-k} \\
&= (n-k)\binom{n}{k} p^k(1-p)^{n-k} \\
&\leqslant \frac{n-k}{\sqrt{2\pi n(p+\varepsilon)(1-p-\varepsilon)}} \exp\left(-n\mathrm{D}_{\mathrm{KL}}\left(p+\varepsilon, p\right)\right) \\
&= \sqrt{\frac{n-k}{2\pi(p+\varepsilon)}} \exp\left(-n\mathrm{D}_{\mathrm{KL}}\left(p+\varepsilon, p\right)\right)
\end{aligned}$$

Basically, this bound proves that

$$\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+\varepsilon)\right] = O(\sqrt{n}) \cdot \exp\left(-n\mathrm{D}_{\mathrm{KL}}\left(p+\varepsilon, p\right)\right)$$

- **Lower-bound.** We can prove a lower bound by using the fact that "the probability of observing $\geqslant k$ heads" is more than "the probability of observing exactly k heads."

$$\begin{aligned}
\mathbb{P}\left[\mathbb{S}_{n,p} = n(p+\varepsilon)\right] &> \mathbb{P}\left[\mathbb{S}_{n,p} = k\right] \\
&= \binom{n}{k} p^k (1-p)^{n-k} \\
&\geqslant \frac{1}{\sqrt{8n(p+\varepsilon)(1-p-\varepsilon)}} \exp\left(-n\mathrm{D}_{\mathrm{KL}}\left(p+\varepsilon, p\right)\right)
\end{aligned}$$

Basically, this bound proves that

$$\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+\varepsilon)\right] = \Omega(1/\sqrt{n}) \exp\left(-n\mathrm{D}_{\mathrm{KL}}\left(p+\varepsilon, p\right)\right)$$

- **Conclusion.** The upper and the lower-bounds can be combined to conclude that $\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+\varepsilon)\right]$ is $\mathrm{poly}(n) \cdot \exp(-n\mathrm{D}_{\mathrm{KL}}\left(p+\varepsilon, p\right))$.

Chernoff Bound: Proof I

- Let us now upper bound the probability $\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+\varepsilon)\right]$ using the Chernoff bound. The upper-bound will be slightly better than what we obtained using the naïve Stirling approximation presented above.
- Recall that $\mathbb{X}$ is a r.v. over the sample space $\{0, 1\}$. Moreover, we have $\mathbb{P}[\mathbb{X} = 1] = p$ and $\mathbb{P}[\mathbb{X} = 0] = 1 - p$. Note that we have $\mathbb{E}[\mathbb{X}] = p$.
- We are studying the r.v.

$$\mathbb{S}_{n,p} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)} + \cdots + \mathbb{X}^{(n)}$$

  Each random variable $\mathbb{X}^{(i)}$ is an independent copy of the random variable $\mathbb{X}$.
- Note that we have $\mathbb{E}\left[\mathbb{S}_{n,p}\right] = n\mathbb{E}[\mathbb{X}] = np$, by the linearity of expectation

Chernoff Bound: Proof II

Theorem (Chernoff Bound)

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

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 $n(1/2 + 1/4) = 3n/4$

Chernoff Bound: Proof III

Let us begin with the proof.

- We are interested in upper-bounding the probability

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

- Note that, for any positive h, we have

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

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

- Now, we apply Markov inequality to obtain

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

Chernoff Bound: Proof IV

- Now, we need an observation. Suppose $\mathbb{A}$ and $\mathbb{B}$ are two independent random variables. Then, we have $\mathbb{E}\left[\exp(\mathbb{A}+\mathbb{B})\right] = \mathbb{E}\left[\exp(\mathbb{A})\right] \cdot \mathbb{E}\left[\exp(\mathbb{B})\right]$. We emphasize that $\mathbb{A}$ and $\mathbb{B}$ 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 iteratively to obtain the following result.

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

- 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

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

Chernoff Bound: Proof V

- Substituting this value, we get

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

- 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 bound holds

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

Chernoff Bound: Proof VI

- To obtain the tightest upper-bound we should use the value of $h = h^*$ that minimizes the right-hand size expression. For simplicity let us make a variable substitution $H = \exp(h)$. Let us define
$$f(H) = \frac{1 - p + pH}{H^{p+\varepsilon}}$$
Our objective is to find $H = H^*$ that minimizes $f(H)$.
- Let us compute $f'(H)$ and solve for $f'(H^*) = 0$. Note that we have
$$f'(H) = \frac{p}{H^{p+\varepsilon}} - \frac{(p+\varepsilon)(1-p+pH)}{H^{p+\varepsilon+1}}$$
The solution $f'(H^*) = 0$ is given by
$$H^* = \frac{p+\varepsilon}{1-p-\varepsilon} \cdot \frac{1-p}{p}.$$

Chernoff Bound: Proof VII

We can check that, for $\varepsilon > 0$, we have $H^* > 1$, that is, $h > 0$.
We can consider the second derivative $f''(H)$ to prove that this extremum is a minima.
Instead of computing $f''(H)$, we can use a shortcut technique. We know that at H^*, the function $f(H)$ either has a maximum or a minimum. Moreover, there is only one extremum of the function $f(H)$. Note that $\lim_{H\to\infty} f(H) = \infty$, so $f(H^*)$ must be a minimum.

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

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

Overview of Generalization I

Our objective is to generalize the Chernoff Bound that we proved above. Let us first recall the Chernoff bound result that we proved.

- Let $\mathbb{X}$ be $\mathrm{Bern}(p)$
- Let $\mathbb{S}_{n,p} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)} + \cdots + \mathbb{X}^{(n)}$
- Chernoff bound states that

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

Overview of Generalization II

We shall generalize this result in two ways

1. For $1 \leqslant i \leqslant n$, let $\mathbb{X}_i$ be an independent $\mathrm{Bern}(p_i)$ random variable. That is, $\mathbb{X}_i$ be a r.v. over $\{0, 1\}$ such that $\mathbb{P}[\mathbb{X}_i = 0] = 1 - p_i$ and $\mathbb{P}[\mathbb{X}_i = 1] = p_i$. Each $\mathbb{X}_i$ is independent of the other $\mathbb{X}_j$s. Let $\mathbb{S}_{n,p} = \mathbb{X}_1 + \mathbb{X}_2 + \cdots + \mathbb{X}_n$, where $p = (p_1 + \cdots + p_n)/n$.
2. For $1 \leqslant i \leqslant n$, let $\mathbb{X}_i$ be a r.v. over $[0, 1]$ such that $\mathbb{E}[\mathbb{X}_i] = p_i$.

Despite these two generalizations, the following bound continues to hold true.

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

First Generalization I

- Let $X_1, X_2, \ldots \mathbb{X}_n$ be independent random variables such that $\mathbb{X}_i = \mathrm{Bern}(p_i)$, for $1 \leqslant i \leqslant n$
- Let $p := (p_1 + p_2 + \cdots + p_n)/n$
- Define $\mathbb{S}_{n,p} = \mathbb{X}_1 + \mathbb{X}_2 + \cdots + \mathbb{X}_n$
- We bound the following probability. For any $H > 1$, we have

$$\mathbb{P}\left[\mathbb{S}_{n,p} \geqslant n(p+\varepsilon)\right] = \mathbb{P}\left[H^{\mathbb{S}_{n,p}} \geqslant H^{n(p+\varepsilon)}\right]$$

- Now, we apply the Markov inequality

$$\mathbb{P}\left[H^{\mathbb{S}_{n,p}} \geqslant H^{n(p+\varepsilon)}\right] \leqslant \frac{\mathbb{E}\left[H^{\mathbb{S}_{n,p}}\right]}{H^{n(p+\varepsilon)}} = \frac{\mathbb{E}\left[H^{\sum_{i=1}^{n} \mathbb{X}_i}\right]}{H^{n(p+\varepsilon)}} = \frac{\mathbb{E}\left[\prod_{i=1}^{n} H^{\mathbb{X}_i}\right]}{H^{n(p+\varepsilon)}}$$

First Generalization II

- Since, each $\mathbb{X}_i$ are independent of other $\mathbb{X}_j$s, we have

$$\frac{\mathbb{E}\left[\prod_{i=1}^{n} H^{\mathbb{X}_i}\right]}{H^{n(p+\varepsilon)}} = \frac{\prod_{i=1}^{n} \mathbb{E}\left[H^{\mathbb{X}_i}\right]}{H^{n(p+\varepsilon)}} = \frac{\prod_{i=1}^{n} 1 - p_i + p_i H}{H^{n(p+\varepsilon)}}$$

- We apply the AM-GM inequality to conclude that

$$\prod_{i=1}^{n} 1 - p_i + p_i H \leqslant \left(\frac{\sum_{i=1}^{n} 1 - p_i + p_i H}{n}\right)^n$$

Equality holds if and only if all $p_i = p$. This bound can now be substituted to conclude

$$\frac{\mathbb{E}\left[\prod_{i=1}^{n} H^{\mathbb{X}_i}\right]}{H^{n(p+\varepsilon)}} \leqslant \left(\frac{1 - p + pH}{H^{p+\varepsilon}}\right)^n$$

- This is identical to the bound that we had in the Chernoff bound proof. We can use the following choice of H in the bound above to obtain the tightest possible bound

$$H^* = \frac{(p+\varepsilon)(1-p)}{p(1-p-\varepsilon)}$$

So, we get the bound

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

Second Generalization I

- Let $1 \leqslant \mathbb{X}_i \leqslant 1$ be a r.v. such that $\mathbb{E}[\mathbb{X}_i] = p_i$ and each $\mathbb{X}_i$ is independent of other $\mathbb{X}_j$s
- Just like the previous setting, we have $\mathbb{S}_{n,p} = \mathbb{X}_1 + \mathbb{X}_2 + \cdots + \mathbb{X}_n$, where $p = (p_1 + p_2 + \cdots + p_n)/n$
- Note that if we prove the following bound, then we shall be done

$$\mathbb{E}\left[H^{\mathbb{X}_i}\right] \leqslant 1 - p_i + p_i H$$

We can use this bound in the previous proof and arrive at the identical upper-bound.

The proof follows from the following

$$\begin{aligned}
\mathbb{E}\left[H^{\mathbb{X}_i}\right] &= \sum_{x\in[0,1]} \mathbb{P}\left[\mathbb{X}_i = x\right] \cdot H^x \\
&= \sum_{x\in[0,1]} \mathbb{P}\left[\mathbb{X}_i = x\right] \cdot H^{(1-x)\cdot 0 + x\cdot 1} \\
&\leqslant \sum_{x\in[0,1]} \mathbb{P}\left[\mathbb{X}_i = x\right] \cdot \left((1-x)\cdot H^0 + x\cdot H^1\right), && \text{(By Jensen's)} \\
&= \sum_{x\in[0,1]} \mathbb{P}\left[\mathbb{X}_i = x\right] \cdot (1 - x + xH) \\
&= \sum_{x\in[0,1]} \mathbb{P}\left[\mathbb{X}_i = x\right] - \sum_{x\in[0,1]} \mathbb{P}\left[\mathbb{X}_i = x\right]\cdot x + H \sum_{x\in[0,1]} \mathbb{P}\left[\mathbb{X}_i = x\right] \cdot x \\
&= 1 - p_i + p_i H, && \text{(Because } \mathbb{E}\left[\mathbb{X}_i\right] = p_i\text{)}
\end{aligned}$$

The appendix provides additional intuition for this analysis.

Conclusion

- Let $1 \leqslant \mathbb{X}_i \leqslant 1$ are independent random variables, for $1 \leqslant i \leqslant n$. Let $p_i = \mathbb{E}[\mathbb{X}_i]$, for $1 \leqslant i \leqslant n$. Define $\mathbb{S}_{n,p} := \mathbb{X}_1 + \mathbb{X}_2 + \cdots + \mathbb{X}_n$, where $p := (p_1 + \cdots + p_n)/n$.

Theorem (Chernoff Bound)

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

- **Objective of the next lecture.** We shall obtain easier to compute, albeit weaker, upper bounds on this probability. These bounds shall rely on the following inequalities
  1. $\mathrm{D}_{\mathrm{KL}}(p+\varepsilon, p) \geqslant 2\varepsilon^2$,
  2. $\mathrm{D}_{\mathrm{KL}}\left(p(1+\varepsilon), p\right) \geqslant \frac{p\varepsilon^2}{2\left(1+\varepsilon/3\right)}$, and
  3. $\mathrm{D}_{\mathrm{KL}}\left(1-p(1-\varepsilon), 1-p\right) \geqslant p\varepsilon^2/2$.

  Check them out at:

  `https://www.desmos.com/calculator/pyessio3v2`

Appendix: Intuition for the Analysis I

- Let $\mathbb{X}$ be an r.v. over $[a, b]$ such that $\mathbb{E}[\mathbb{X}] = \mu$
- Let $f : \mathbb{R} \to \mathbb{R}$ be a concave upwards function (that is, it looks like $f(x) = x^2$)
- Jensen's inequality states that $f(\mathbb{E}[\mathbb{X}]) \leqslant \mathbb{E}\left[f(\mathbb{X})\right]$, and equality holds if and only if $\mathbb{X}$ has its entire probability mass at μ. Therefore, we can conclude that $f(\mu) \leqslant \mathbb{E}\left[f(\mathbb{X})\right]$
- So, we have a lower-bound on $\mathbb{E}\left[f(\mathbb{X})\right]$. Now, we are interested in obtaining an upper-bound on $\mathbb{E}\left[f(\mathbb{X})\right]$
- For the upper-bound note that is $\mathbb{X}$ deposits more probability mass away from μ, then $\mathbb{E}\left[f(\mathbb{X})\right]$ increases. In fact, increasing the mass further away increases $\mathbb{E}\left[f(\mathbb{X})\right]$ more. So, the maximum value of $\mathbb{E}\left[f(\mathbb{X})\right]$ is achieved when $\mathbb{X}$ deposits the entire probability mass either at a or b only. Let us find such a probability distribution under the constraint that $\mathbb{E}[\mathbb{X}] = \mu$

Appendix: Intuition for the Analysis II

- Suppose $\mathbb{P}[\mathbb{X}^* = a] = p$. Then, we have $\mathbb{P}[\mathbb{X}^* = b] = 1 - p$. Further, the constraint $\mathbb{E}[\mathbb{X}^*] = \mu$ becomes $pa + (1-p)b = \mu$. Solving, we get

$$p = \frac{b-\mu}{b-a}$$

Therefore, we get $1 - p = \frac{\mu-a}{b-a}$. For this probability, we get

$$\mathbb{E}\left[f(\mathbb{X}^*)\right] = \frac{b-\mu}{b-a}f(a) + \frac{\mu-a}{b-a}f(b)$$

So, we expect the following bound to hold for a general r.v. $\mathbb{X}$

$$\mathbb{E}\left[f(\mathbb{X})\right] \leqslant \mathbb{E}\left[f(\mathbb{X}^*)\right] = \frac{b-\mu}{b-a}f(a) + \frac{\mu-a}{b-a}f(b)$$

This is not a formal proof. Let us prove this intuition formally.

Appendix: Intuition for the Analysis III

- Let $\mathbb{X}$ be an r.v. over $[a, b]$ with $\mathbb{E}[\mathbb{X}] = \mu$. Note that by Jensen's inequality, we have

$$f(x) = f\left(\frac{b-x}{b-a}a + \frac{x-a}{b-a}b\right) \leqslant \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b)$$

Now, we take expectation on both sides to conclude that

$$\begin{aligned}\mathbb{E}\left[f(\mathbb{X})\right] &\leqslant \mathbb{E}\left[\frac{b-\mathbb{X}}{b-a}f(a) + \frac{\mathbb{X}-a}{b-a}f(b)\right] \\ &= \frac{b-\mathbb{E}\left[\mathbb{X}\right]}{b-a}f(a) + \frac{\mathbb{E}\left[\mathbb{X}\right]-a}{b-a}f(b) \\ &= \frac{b-\mu}{b-a}f(a) + \frac{\mu-a}{b-a}f(b)\end{aligned}$$

- To conclude, we have the following bound.

$$f(\mu) \leqslant \mathbb{E}\left[f(\mathbb{X})\right] \leqslant \frac{b-\mu}{b-a}f(a) + \frac{\mu-a}{b-a}f(b)$$