Lecture 06: Chernoff Bound
Let $X$ 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 $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ represent $n$ independent coin tosses that are identically distributed as the random variable $X$.

We are interested in studying the random variable

$$S_{n,p} = X^{(1)} + X^{(2)} + \ldots + X^{(n)}$$

This random variable $S_{n,p}$ represents the total number of heads in the $n$ coin tosses.
Formally, given \( \varepsilon > 0 \), we are interested in computing the probability that

\[ \mathbb{P} \left[ S_{n,p} \geq n(p + \varepsilon) \right] \leq ??? \]

That is, we are interested to prove that the probability of our estimate being “much larger” than \( p \) is small.
Suppose we have seen $i$ heads. We can explicitly compute the probability that $S_{n,p} = i$ 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

$$P[S_{n,p} = i] = \binom{n}{i} p^i (1-p)^{n-i}$$

Therefore, from this result, our desired probability is

$$P[S_{n,p} \geq n(p+\varepsilon)] = \sum_{i \geq n(p+\varepsilon)} \binom{n}{i} p^i (1-p)^{n-i}$$

For simplicity, let us assume that $n(p+\varepsilon) = k$ is an integer
Upper-bound. We can prove that among the elements \(( \binom{n}{i} p^i (1 - p)^{n-i} )\), where \(i \geq k\), the maximum element is one where \(i = k\). We can use this observation to upper-bound the probability expression.

\[
P[\mathbb{S}_{n,p} \geq n(p + \varepsilon)] = \sum_{i \geq k} \binom{n}{i} n i p^i (1 - p)^{n-i}
\leq \sum_{i \geq k} \binom{n}{k} p^k (1 - p)^{n-k}
= (n - k) \binom{n}{k} p^k (1 - p)^{n-k}
\leq \frac{n - k}{\sqrt{2\pi n(p + \varepsilon)(1 - p - \varepsilon)}} \exp\left(-nD_{KL}(p + \varepsilon, p)\right)
= \sqrt{\frac{n - k}{2\pi(p + \varepsilon)}} \exp\left(-nD_{KL}(p + \varepsilon, p)\right)
\]
Approach using Stirling’s Approximation III

Basically, this bound proves that

\[ \mathbb{P} [S_{n, p} \geq n(p + \varepsilon)] = O(\sqrt{n}) \cdot \exp \left( -nD_{\text{KL}}(p + \varepsilon, p) \right) \]

- **Lower-bound.** We can prove a lower bound by using the fact that “the probability of observing \( \geq k \) heads” is more than “the probability of observing exactly \( k \) heads.”

\[
\mathbb{P} [S_{n, p} = n(p + \varepsilon)] > \mathbb{P} [S_{n, p} = k] \\
= \binom{n}{k} p^k (1 - p)^{n-k} \\
\geq \frac{1}{\sqrt{8n(p + \varepsilon)(1 - p - \varepsilon)}} \exp \left( -nD_{\text{KL}}(p + \varepsilon, p) \right)
\]

Basically, this bound proves that

\[ \mathbb{P} [S_{n, p} \geq n(p + \varepsilon)] = \Omega(1/\sqrt{n}) \exp \left( -nD_{\text{KL}}(p + \varepsilon, p) \right) \]
**Conclusion.** The upper and the lower-bounds can be combined to conclude that $\Pr[S_{n,p} \geq n(p + \varepsilon)]$ is $\text{poly}(n) \cdot \exp(-nD_{KL}(p + \varepsilon, p))$. 
Let us now upper bound the probability $\mathbb{P}[S_{n,p} \geq n(p + \varepsilon)]$ using the Chernoff bound. The upper-bound will be slightly better than what we obtained using the naïve Stirling approximation presented above.

Recall that $X$ is a r.v. over the sample space $\{0, 1\}$. Moreover, we have $\mathbb{P}[X = 1] = p$ and $\mathbb{P}[X = 0] = 1 - p$. Note that we have $\mathbb{E}[X] = p$.

We are studying the r.v.

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

Each random variable $X^{(i)}$ is an independent copy of the random variable $X$.

Note that we have $\mathbb{E}[S_{n,p}] = n\mathbb{E}[X] = np$, by the linearity of expectation.
Theorem (Chernoff Bound)

\[ P \left[ S_{n,p} \geq n(p + \varepsilon) \right] \leq \exp\left( -nD_{\text{KL}}(p + \varepsilon, p) \right) \]

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 \( S_{n,p} \) surpasses \( n(1/2 + 1/4) = 3n/4 \).
Let us begin with the proof.

- We are interested in upper-bounding the probability

\[ P \left[ S_{n,p} \geq n(p + \varepsilon) \right] \]

- Note that, for any positive \( h \), we have

\[ P \left[ S_{n,p} \geq n(p + \varepsilon) \right] = P \left[ \exp(hS_{n,p}) \geq \exp(hn(p + \varepsilon)) \right] \]

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

- Now, we apply Markov inequality to obtain

\[ P \left[ \exp(hS_{n,p}) \geq \exp(hn(p + \varepsilon)) \right] \leq \frac{\mathbb{E} \left[ \exp(hS_{n,p}) \right]}{\exp(hn(p + \varepsilon))} \]
Now, we need an observation. Suppose $A$ and $B$ are two independent random variables. Then, we have 

$$\mathbb{E} \left[ \exp(A + B) \right] = \mathbb{E} \left[ \exp(A) \right] \cdot \mathbb{E} \left[ \exp(B) \right].$$

We emphasize that $A$ and $B$ have to be independent to apply this result.

Note that we have $S_{n,p} = \sum_{i=1}^{n} X^{(i)}$. So, we can apply the previous observation iteratively to obtain the following result.

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

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

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

Concentration Bounds
Substituting this value, we get

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

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

\[
P \left[ S_{n,p} \geq n(p + \varepsilon) \right] \leq \left( \frac{1 - p + p \exp(h)}{\exp(h(p + \varepsilon))} \right)^n
\]
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

$$f(H) = \frac{1 - p + pH}{Hp+\varepsilon}$$

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

$$f'(H) = \frac{p}{Hp+\varepsilon} - \frac{(p + \varepsilon)(1 - p + pH)}{Hp+\varepsilon+1}$$

The solution $f'(H^*) = 0$ is given by

$$H^* = \frac{p + \varepsilon}{1 - p - \varepsilon} \cdot \frac{1 - p}{p}.$$
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

$$\mathbb{P}[S_{n,p} \geq n(p + \varepsilon)] \leq \left( \frac{1 - p + \frac{(1-p)(p+\varepsilon)}{1-p-\varepsilon}}{\frac{(1-p)(p+\varepsilon)}{p(1-p-\varepsilon)}} \right)^n$$

$$= \left( \frac{1-p}{1-p-\varepsilon} \right)^n$$

$$= \left( \left( \frac{p}{p+\varepsilon} \right)^{p+\varepsilon} \left( \frac{1-p}{1-p-\varepsilon} \right)^{1-p-\varepsilon} \right)^n$$

$$= \exp(-nD_{KL}(p + \varepsilon, p))$$
Our objective is to generalize the Chernoff Bound that we proved above. Let us first recall the Chernoff bound result that we proved.

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

$$
\Pr \left[ S_{n,p} \geq n(p + \varepsilon) \right] \leq \exp \left( -nD_{KL}(p + \varepsilon, p) \right)
$$
We shall generalize this result in two ways

1. For $1 \leq i \leq n$, let $X_i$ be an independent $\text{Bern}(p_i)$ random variable. That is, $X_i$ be a r.v. over $\{0, 1\}$ such that $P[X_i = 0] = 1 - p_i$ and $P[X_i = 1] = p_i$. Each $X_i$ is independent of the other $X_j$s. Let $S_{n,p} = X_1 + X_2 + \cdots + X_n$, where $p = (p_1 + \cdots + p_n)/n$.

2. For $1 \leq i \leq n$, let $X_i$ be a r.v. over $[0, 1]$ such that $\mathbb{E}[X_i] = p_i$. Despite these two generalizations, the following bound continues to hold true.

$$P[S_{n,p} \geq n(p + \varepsilon)] \leq \exp(-nD_{KL}(p + \varepsilon, p))$$
First Generalization I

- Let $X_1, X_2, \ldots X_n$ be independent random variables such that $X_i = \text{Bern}(p_i)$, for $1 \leq i \leq n$
- Let $p := (p_1 + p_2 + \cdots + p_n)/n$
- Define $S_{n,p} = X_1 + X_2 + \cdots + X_n$
- We bound the following probability. For any $H > 1$, we have

$$P[S_{n,p} \geq n(p + \varepsilon)] = P[H^{S_{n,p}} \geq H^n(p + \varepsilon)]$$

- Now, we apply the Markov inequality

$$P[H^{S_{n,p}} \geq H^n(p + \varepsilon)] \leq \frac{\mathbb{E}[H^{S_{n,p}}]}{H^n(p + \varepsilon)} = \frac{\mathbb{E}[H^{\sum_{i=1}^n X_i}]}{H^n(p + \varepsilon)} = \frac{\mathbb{E}[\prod_{i=1}^n H^{X_i}]}{H^n(p + \varepsilon)}$$
First Generalization II

Since, each $X_i$ are independent of other $X_j$'s, we have

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

We apply the AM-GM inequality to conclude that

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

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

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

$$\mathbb{P} \left[ S_{n,p} \geq n(p + \varepsilon) \right] \leq \exp(-nD_{KL}(p + \varepsilon, p))$$
Let $1 \leq X_i \leq 1$ be a r.v. such that $\mathbb{E}[X_i] = p_i$ and each $X_i$ is independent of other $X_j$s.

Just like the previous setting, we have

$$S_{n,p} = X_1 + X_2 + \cdots + X_n,$$

where $p = (p_1 + p_2 + \cdots + p_n)/n$.

Note that if we prove the following bound, then we shall be done

$$\mathbb{E}[H^{X_i}] \leq 1 - p_i + p_i H$$

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

\[ E \left[ H^{X_i} \right] = \sum_{x \in [0,1]} \mathbb{P}[X_i = x] \cdot H^x \]

\[ = \sum_{x \in [0,1]} \mathbb{P}[X_i = x] \cdot H^{(1-x) \cdot 0 + x \cdot 1} \]

\[ \leq \sum_{x \in [0,1]} \mathbb{P}[X_i = x] \cdot \left( (1 - x) \cdot H^0 + x \cdot H^1 \right), \quad \text{(By Jensen’s)} \]

\[ = \sum_{x \in [0,1]} \mathbb{P}[X_i = x] \cdot (1 - x + xH) \]

\[ = \sum_{x \in [0,1]} \mathbb{P}[X_i = x] - \sum_{x \in [0,1]} \mathbb{P}[X_i = x] \cdot x + H \sum_{x \in [0,1]} \mathbb{P}[X_i = x] \cdot x \]

\[ = 1 - p_i + p_i H, \quad \text{(Because } E[X_i] = p_i \text{)} \]

The appendix provides additional intuition for this analysis.
Conclusion

Let $1 \leq X_i \leq 1$ are independent random variables, for $1 \leq i \leq n$. Let $p_i = \mathbb{E}[X_i]$, for $1 \leq i \leq n$. Define $S_{n,p} := X_1 + X_2 + \cdots + X_n$, where $p := (p_1 + \cdots + p_n)/n$.

**Theorem (Chernoff Bound)**

$$
\mathbb{P} \left[ S_{n,p} \geq n(p + \varepsilon) \right] \leq \exp\left(-nD_{KL}(p + \varepsilon, p)\right)
$$

**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. $D_{KL}(p + \varepsilon, p) \geq 2\varepsilon^2$,
2. $D_{KL}(p(1 + \varepsilon), p) \geq \frac{p\varepsilon^2}{2(1+\varepsilon/3)}$, and
3. $D_{KL}(1 - p(1 - \varepsilon), 1 - p) \geq p\varepsilon^2/2$.

Check them out at:
https://www.desmos.com/calculator/pyessio3v2
Appendix: Intuition for the Analysis I

- Let $X$ be an r.v. over $[a, b]$ such that $E[X] = \mu$
- Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be a concave upwards function (that is, it looks like $f(x) = x^2$)
- Jensen’s inequality states that $f(E[X]) \leq E[f(X)]$, and equality holds if and only if $X$ has its entire probability mass at $\mu$. Therefore, we can conclude that $f(\mu) \leq E[f(X)]$
- So, we have a lower-bound on $E[f(X)]$. Now, we are interested in obtaining an upper-bound on $E[f(X)]$
- For the upper-bound note that is $X$ deposits more probability mass away from $\mu$, then $E[f(X)]$ increases. In fact, increasing the mass further away increases $E[f(X)]$ more. So, the maximum value of $E[f(X)]$ is achieved when $X$ deposits the entire probability mass either at $a$ or $b$ only. Let us find such a probability distribution under the constraint that $E[X] = \mu$
Suppose $\mathbb{P}[X^* = a] = p$. Then, we have $\mathbb{P}[X^* = b] = 1 - p$.
Further, the constraint $\mathbb{E}[X^*] = \mu$ 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

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

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

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

This is not a formal proof. Let us prove this intuition formally.
Let $X$ be an r.v. over $[a, b]$ with $E[X] = \mu$. Note that by Jensen’s inequality, we have

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

Now, we take expectation on both sides to conclude that

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

$$= \frac{b - E[X]}{b - a}f(a) + \frac{E[X] - a}{b - a}f(b)$$

$$= \frac{b - \mu}{b - a}f(a) + \frac{\mu - a}{b - a}f(b)$$

To conclude, we have the following bound.

$$f(\mu) \leq E[f(X)] \leq \frac{b - \mu}{b - a}f(a) + \frac{\mu - a}{b - a}f(b)$$