11. Concentration Bounds & Error Reduction

Chernoff Bound for Bernoulli Variables

Bernoulli Random Variables are given by

$$X = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$$

Now, we take n independent Bernoulli trials, $X_1, X_2, \ldots, X_n$, and define $X = \sum\limits_{i=1}^{n} X_i$

The probability that $X = i$ is

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

We want to find a tighter bound on the probability that X deviates from $\mathbb{E}[X]$.

Theorem -

Given that we have $X_1, X_2, \ldots, X_n$ trials with probability $p_1, p_2, \ldots, p_n$, and $X = \sum\limits_i X_i$, for $\delta > 0$, where $\mu = \mathbb{E}[X]$ -

$$\mathbb{P}\left[X > (1+\delta)\mu\right] < \left[\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right]^\mu$$

Proof -

Since $e^{tX} \geq 0$, we know that $X \geq k \iff e^{tX} \geq e^{tk}$ when $t > 0$.

Therefore using Markov's Inequality, we get

$$\mathbb{P}[X \geq k] = \mathbb{P}[e^{tX} \geq e^{tk}] \leq \frac{\mathbb{E}[e^{tX}]}{e^{tk}}$$

This inequality is meant to be held irrespective of the value of t (as long as it's positive). Therefore, we can find the tightest bound by finding for which t is $\frac{\mathbb{E}[e^{tX}]}{e^{tk}}$ minimized.

First, we compute what is $\mathbb{E}[e^{tX_i}]$

$$\mathbb{E}[e^{tX_i}] = p_i e^t + (1-p_i) = 1 + p_i(e^t - 1) \leq e^{p_i(e^t-1)}$$

Noting that

$$\mathbb{E}\left[e^{t\left(\sum_i X_i\right)}\right] = \prod_i \mathbb{E}[e^{tX_i}]$$

We can write

$$\mathbb{P}[X \geq k] \leq \min_{t>0} \frac{\prod_{i=1}^n \mathbb{E}[e^{tX_i}]}{e^{tk}} \leq \min_{t>0} \frac{e^{\sum p_i(e^t-1)}}{e^{tk}} = \min_{t>0} \frac{e^{\mu(e^t-1)}}{e^{tk}}$$

If we take $k = (1+\delta)\mu$, then we can find that at $t = \ln(1+\delta)$, the value is minimized. Plugging in this value, we get the bound

$$\mathbb{P}[X \geq (1+\delta)\mu] \leq \frac{e^{\mu(e^{\ln(1+\delta)}-1)}}{e^{(1+\delta)\mu\ln(1+\delta)}} = \frac{e^{\mu\delta}}{(1+\delta)^{(1+\delta)\mu}}$$

Similarly, the bound for $\mathbb{P}[X \leq (1-\delta)\mu]$ can be found using a similar procedure, except optimizing for $t < 0$.

$$\mathbb{P}[X < (1-\delta)\mu] \leq \frac{e^{-\mu\delta}}{(1-\delta)^{(1-\delta)\mu}} \leq \exp\left(\frac{-\mu\delta^2}{2}\right)$$

Theorem -
For any $0 < \delta \leq 1$,

$$\mathbb{P}[X \geq (1+\delta)\mu] \leq e^{\frac{-\mu\delta^2}{3}}$$

Proof -
We essentially just want to show that for $0 < \delta \leq 1$

$$\frac{e^\delta}{(1+\delta)^{1+\delta}} \leq e^{\frac{-\delta^2}{3}}$$

Taking the logarithm on both sides, we can instead prove

$$f(\delta) = \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{3} \leq 0$$

Computing the derivatives of $f(\delta)$, we get

$$f'(\delta) = 1 - \ln(1+\delta) - \frac{1+\delta}{1+\delta} + \frac{2\delta}{3} = -\ln(1+\delta) + \frac{2\delta}{3}$$

$$f''(\delta) = -\frac{1}{1+\delta} + \frac{2}{3}$$

$f''(\delta) < 0$ for $0 \leq \delta < \frac{1}{2}$, and $f''(\delta) > 0$ for $\delta > \frac{1}{2}$. Therefore, $f'(\delta)$ first decreases, then increases.
Since we are working with the interval $[0, 1]$, we try to see how $f'(\delta)$ values evolve in this interval. We see that $f'(0) = 0$ and $f'(1) < 0$. Since $f'(\delta)$ decreases, then increases, the max value will be either at 0 or 1, hence the values of $f'(\delta)$ in this interval is ≤ 0. Therefore, the value of $f(\delta)$ will always be non-increasing when going from 0 to 1. We see that $f(0) = 0$, hence $f(\delta) \leq 0$, hence proving the inequality.
Theorem -
For $R \geq 6\mu$,

$$\mathbb{P}(X \geq R) \leq 2^{-R}$$

Proof -
Taking $R = (1+\delta)\mu$, if $R \geq 6\mu$, then $\delta = \frac{R}{\mu} - 1 \geq 5$.
Then, using the earlier Chernoff bound, we get

$$\begin{aligned}
\mathbb{P}(X \geq R) &\leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu \\
&\leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu} \\
&\leq \left(\frac{e}{6}\right)^R \\
&\leq 2^{-R} && \left(\frac{e}{6} \leq \frac{1}{2}\right)
\end{aligned}$$

Error Reduction

Theorem - Let A be a randomized algorithm for task T such that A runs in polynomial time and succeeds with probability $\geq \frac{2}{3}$. Then, there exists another randomized algorithm that runs in polynomial time and succeeds with probability $\geq 1 - 2^{-n^{O(1)}}$.

Algorithm A' runs A s times (we find s later), and we output the answer that occurs a majority of the time as the answer for A'.

X_i- The indicator random variable that the i^{th} iteration of A succeeded.

$$\mathbb{E}[X_i] = \mathbb{P}[X_i = 1] \geq \frac{2}{3}$$

We then define $X = \sum_i X_i$ as the number of successful trials, therefore $\mathbb{E}[X] \geq \frac{2}{3}s$ by linearity of expectation.

The output that A' is incorrect if A fails more than half the times out of the s trials, i.e $X < \frac{s}{2}$

$$\mathbb{P}\left[X < \frac{s}{2}\right] = \mathbb{P}[X < (1-\delta)\mathbb{E}[X]]$$

$$\frac{s}{2} = (1-\delta)\frac{2s}{3} \Rightarrow \frac{3}{4} = 1 - \delta \Rightarrow \delta = \frac{1}{4}$$

Hence, we can bound the probability of A' failure by

$$\mathbb{P}\left[X < \frac{s}{2}\right] < \exp\left(-\frac{s}{3 \cdot 16}\right) = \exp(-\Omega(s))$$

Setting s to n^2 would cause the error probability to be bounded by

$$\frac{1}{2^{\Omega(n^2)}}$$

Benefit of Chernoff over Markov & Chebyshev

If we try to solve the same problem of finding a bound on error probability using Markov's Inequality, we'd get

$$\mathbb{P}\left[X > \frac{n}{2}\right] \leq \frac{\frac{2}{3}s}{\frac{s}{2}} \leq \frac{4}{3}$$

Giving us no information essentially.

Finding the bound using Chebyshev, we want to see what's the probability of it deviating greater than $\frac{2s}{3} - \frac{s}{2} = \frac{s}{6}$

Assuming $p_i = \frac{2}{3}$, the variance of one Bernoulli trial is $p(1-p) = \frac{2}{9}$.
Therefore, the variance of X is given by $\mathrm{Var}(X) = \sum_i \mathrm{Var}(X_i) = \frac{2}{9}s$

Using these values, we get the inequality

$$\mathbb{P}\left[|X - \mu| > \frac{s}{6}\right] \leq \frac{\frac{2}{9}s}{\frac{s^2}{36}} = \frac{8}{s}$$

Which is a better inequality than what we got with Markov's Inequality, but not as good as Chernoff's Bound.

Set Balancing

Corollary - Let $X_1, X_2, \ldots, X_n$ be independent random variables with

$$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = \frac{1}{2}$$

Let $X = \sum_i^n X_i$, then for any $a > 0$,

$$\mathbb{P}(|X| \geq a) \leq 2e^{\frac{-a^2}{2n}}$$

Using this Corollary, we look at the application of this in Set Balancing. You are given an $n \times m$ matrix A with entries $\in \{0, 1\}$

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

We are trying to find a vector $\bar{b}$ with entries $\in \{-1, 1\}$ such that it minimizes

$$||A\bar{b}||_\infty = \max_i |c_i|$$

This problem arises in designing statistical experiments. Each column of the matrix A represents a subject in the experiment and each row represents a feature. The vector $\bar{b}$ partitions the subjects into two disjoint groups, so that each feature is roughly as balanced as possible between the two groups. One of the groups serves as a control group for an experiment that is run on the other group.

The randomized algorithm essentially picks each entry of $\bar{b}$ with $\mathbb{P}[b_i = 1] = \mathbb{P}[b_i = -1] = \frac{1}{2}$.

Theorem - For a random vector $\bar{b}$, with entries chosen randomly,

$$\mathbb{P}\left[||A\bar{b}||_\infty \geq \sqrt{4m \ln n}\right] \leq \frac{2}{n}$$

(Try to understand this as the dot product of the i^{th} row and $\bar{b}$).
If we consider the i^{th} row, $\bar{a}_i = a_{i1}, a_{i2}, \ldots, a_{im}$ and let k be the number of 1s, in that row. If $k \leq \sqrt{4m \ln n}$, then we trivially know that $\langle \bar{a}_i, \bar{b} \rangle = |c_i| \leq \sqrt{4m \ln n}$.

If $k > \sqrt{4m \ln n}$, then we define a random variable in terms of the k ones in the row,

$$Z_i = \sum_{j=1}^{m} a_{ij} b_j$$

Using the earlier Corallary, and knowing that $m \geq k$, we can say

$$\mathbb{P}\left[|Z_i| \geq \sqrt{4m \ln n}\right] \leq 2e^{-\frac{4m \ln n}{2k}} = 2n^{-\frac{2m}{k}} = \frac{2}{n^{\frac{2m}{k}}} \leq \frac{2}{n^2}$$

Probability that any of of them is greater than $\sqrt{4m \ln n}$ is given by

$$\mathbb{P}\left[\bigcup_{i}^{n} |Z_i| \geq \sqrt{4m \ln n}\right] \leq \sum_{i}^{n} \mathbb{P}\left[|Z_i| \geq \sqrt{4m \ln n}\right] \leq \frac{2}{n^2} \cdot n = \frac{2}{n}$$