3
Concentration of Measure: Chernoff-Hoeffding Bounds

3.1 Recap and Motivation

In our last lecture, we introduced two fundamental concentration inequalities:

- **Markov's Inequality**: For a non-negative RV X, $\Pr(X \geq \lambda) \leq \frac{\mathbb{E}[X]}{\lambda}$. This uses the first moment (the mean).
- **Chebyshev's Inequality**: For any RV X, $\Pr(|X - \mu| \geq \lambda) \leq \frac{\text{Var}(X)}{\lambda^2}$. This uses the second moment (the variance).

While useful, Chebyshev's inequality is often not strong enough. For certain "worst-case" distributions, the bound is tight. However, for many common scenarios, particularly those involving **sums of independent random variables**, the probability of deviating from the mean is much smaller than Chebyshev's inequality suggests.

Today, we will explore a powerful class of inequalities known as Chernoff-Hoeffding bounds. These bounds provide exponentially decreasing probabilities for large deviations, but they require stronger assumptions about the random variables involved.

3.1.1 A Motivating Example: Coin Flips

Consider 1000 coin flips. Let $Y_i = +1$ if the i-th flip is heads, and $Y_i = -1$ if tails. Let $Y = \sum_{i=1}^{1000}(Y_i)$, where Y_i are i.i.d. Rademacher. The number of heads is $X = \frac{1000+Y}{2}$. So getting ≥ 625 heads is equivalent to $Y \geq 250$. Note that $\mathbb{E}[Y] = 0$ and $\text{Var}(Y) = 1000$. What is $\Pr(Y \geq 250)$?

- **Chebyshev's inequality** gives:

$$\Pr(Y \geq 250) \leq \Pr(|Y - 0| \geq 250) \leq \frac{\text{Var}(Y)}{250^2} = \frac{1000}{62500} = 0.016$$

- **The actual probability**, as we will see from Chernoff bounds, is incredibly small:

$$\Pr(Y \geq 250) \leq e^{-250^2/(2\cdot 1000)} = e^{-31.25} \approx 10^{-14}$$

This enormous difference highlights the power of the bounds we will study today.

3.1.2 *Intuition from the Central Limit Theorem (CLT)*

The Central Limit Theorem tells us that the sum of many independent and identically distributed (i.i.d.) random variables, when properly normalized, converges in distribution to a standard normal (Gaussian) distribution.

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0,1)$$

Gaussian distributions have tails that decay exponentially fast (e.g., $\Pr(Z \geq \lambda) \approx e^{-\lambda^2/2}$ for $Z \sim N(0,1)$). This suggests that sums of i.i.d. variables should also have exponentially decaying tails. The CLT is a limiting statement, but Chernoff bounds make this quantitative for any finite n.

3.2 *Chernoff-Hoeffding Bounds*

These bounds apply to sums of independent random variables.

Theorem 3.1 (Concentration for Rademacher RVs)**.** *Let $Y_1, \ldots, Y_n$ be independent Rademacher random variables (i.e., $\Pr(Y_i = 1) = \Pr(Y_i = -1) = 1/2$). Let $Y = \sum_{i=1}^n Y_i$. Then for any $\lambda > 0$:*

$$\Pr(Y \geq \lambda) \leq \exp\left(-\frac{\lambda^2}{2n}\right)$$

Since Y is symmetric about 0, we also have $\Pr(Y \leq -\lambda) \leq \exp(-\lambda^2/2n)$, and by the union bound, $\Pr(|Y| \geq \lambda) \leq 2\exp(-\lambda^2/2n)$. This is a "Gaussian-type" tail, decaying much faster than the "polynomial tail" $(1/\lambda^2)$ from Chebyshev's inequality.

A more general and widely applicable version is Hoeffding's inequality.

Theorem 3.2 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ for all i. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X]$. Then for any $t > 0$:*

$$\Pr(|X - \mu| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

A simpler, common form is for variables bounded in $[0, 1]$*:*

$$\Pr(|X - \mu| \geq \epsilon) \leq 2 \exp\left(-\frac{2\epsilon^2}{n}\right)$$

There is also a version for relative error, which is often very useful:

$$\Pr(|X - \mu| \geq \epsilon\mu) \leq 2 \exp\left(-\frac{\epsilon^2 \mu}{3}\right) \quad \textit{for } 0 < \epsilon < 1$$

3.3 The Chernoff Bound Technique: Proof Sketch

The proof of all these bounds follows a beautiful, unified method. We will sketch it for the Rademacher case.

The goal is to bound $\Pr(Y \geq \lambda)$.

1. **Exponentiate**: For any $t > 0$, the function $z \mapsto e^{tz}$ is monotonically increasing. Therefore:

$$\Pr(Y \geq \lambda) = \Pr(e^{tY} \geq e^{t\lambda})$$

2. **Apply Markov's Inequality**:

$$\Pr(e^{tY} \geq e^{t\lambda}) \leq \frac{\mathbb{E}[e^{tY}]}{e^{t\lambda}}$$

   This term, $\mathbb{E}[e^{tY}]$, is the Moment Generating Function (MGF) of Y.

3. **Use Independence**: The MGF of a sum of independent variables is the product of their individual MGFs.

$$\mathbb{E}[e^{tY}] = \mathbb{E}\left[e^{t\sum Y_i}\right] = \mathbb{E}\left[\prod_i e^{tY_i}\right] = \prod_i \mathbb{E}[e^{tY_i}]$$

4. **Bound the Individual MGFs**: For a Rademacher Y_i:

$$\mathbb{E}[e^{tY_i}] = \frac{1}{2}e^t + \frac{1}{2}e^{-t} = \cosh(t)$$

   Using the inequality $\cosh(t) \leq e^{t^2/2}$, we get $\mathbb{E}[e^{tY_i}] \leq e^{t^2/2}$. This inequality is a standard result, often proven using Taylor series expansions. The Taylor series for $\cosh(t)$ is:

$$\cosh(t) = 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \frac{t^6}{6!} + \cdots = \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!}$$

   The Taylor series for e^x is $e^x = 1 + x + \frac{x^2}{2!} + \ldots$. If we substitute $x = t^2/2$, we get:

$$e^{t^2/2} = 1 + \frac{t^2}{2} + \frac{(t^2/2)^2}{2!} + \frac{(t^2/2)^3}{3!} + \ldots$$

$$= 1 + \frac{t^2}{2} + \frac{t^4}{8} + \frac{t^6}{48} + \cdots = \sum_{k=0}^{\infty} \frac{t^{2k}}{2^k k!}$$

By comparing the denominators of the corresponding terms (t^{2k}) in both series, you can see that $(2k)! \geq 2^k k!$ for all $k \geq 0$. This means that each term in the $\cosh(t)$ series is less than or equal to the corresponding term in the $e^{t^2/2}$ series. Therefore, the inequality $\cosh(t) \leq e^{t^2/2}$ holds for all t.

5. **Combine and Optimize**: Substitute back into the main inequality:

$$\Pr(Y \geq \lambda) \leq \frac{\prod_i e^{t^2/2}}{e^{t\lambda}} = \frac{e^{nt^2/2}}{e^{t\lambda}} = e^{nt^2/2 - t\lambda}$$

This bound holds for any $t > 0$. To get the tightest possible bound, we minimize the exponent with respect to t. The minimum occurs at $t = \lambda/n$. Plugging this back in gives the final result:

$$\Pr(Y \geq \lambda) \leq \exp\left(n\frac{\lambda^2}{2n^2} - \frac{\lambda^2}{n}\right) = \exp\left(-\frac{\lambda^2}{2n}\right)$$

This general strategy of exponentiating, applying Markov, and optimizing the free parameter t is the core of the Chernoff bound method.

3.4 *Applications of Chernoff Bounds*

3.4.1 *Mean Estimation (Polling)*

Suppose we want to estimate the fraction p of a large population that supports a certain candidate. We can poll n people, chosen independently and uniformly at random. Let $X_i \sim \text{Bernoulli}(p)$ be 1 if the i-th person is a supporter, and 0 otherwise. Our estimate for p is the sample mean $\bar{X} = \frac{1}{n}\sum X_i$.

How many people do we need to poll to be confident our estimate is accurate? We want to bound $\Pr(|\bar{X} - p| \geq \epsilon)$. This is equivalent to $\Pr(|X - \mu| \geq n\epsilon)$, where $X = \sum X_i$ and $\mu = np$.

Using a Chernoff bound:

$$\Pr(|X - \mu| \geq n\epsilon) \leq 2\exp\left(-\frac{(n\epsilon)^2 \cdot \text{const}}{n}\right) = 2e^{-C\epsilon^2 n}$$

If we want this probability to be at most δ (our error tolerance, e.g., $\delta = 0.05$ for 95% confidence), we can solve for n:

$$2e^{-C\epsilon^2 n} \leq \delta \implies n \geq \frac{1}{C\epsilon^2}\log\frac{2}{\delta}$$

This tells us the number of samples needed grows as $1/\epsilon^2$ (for better accuracy) and $\log(1/\delta)$ (for higher confidence), which is a cornerstone of statistical sampling and machine learning.

3.4.2 *Low Congestion Routing via Randomized Rounding*

- **Problem**: Given a graph $G = (V, E)$ and k pairs of terminals $(s_1, t_1), (s_2, t_2), \ldots, (s_k, t_k)$, find a single path for each pair to minimize the *congestion*, defined as the maximum number of paths using any single edge.
- **Hardness**: This integer routing problem is NP-hard.
- **Algorithm Idea**: We can use a two-step "relax and round" approach.
  1. **LP Relaxation**: Formulate the problem as a linear program. Instead of choosing one whole path, allow for "fractional" paths. For each pair (s_i, t_i), we can send flow along multiple paths, as long as the total flow for that pair sums to 1. Specifically, the linear program has a variable x_p for every path between two pairs, and is formulated as follows:

$$\begin{aligned} \min \quad & z \\ \text{s.t.} \quad & \sum_{p \in P_i} x_p = 1, && \forall (s_i, t_i) \\ & \sum_{p : e \in p} x_p \leq z, && \forall e \\ & x \geq 0 \end{aligned}$$

     Here, we let P_i denote the paths from s_i to t_i. We can solve this LP efficiently to find the minimum possible fractional congestion z. Let's call it $C_{opt} = \max(z, 1)$, which is clearly a lower bound on the integral congestion.
  2. **Randomized Rounding**: For each pair (s_i, t_i), the LP solution gives us a set of paths and fractional values $x_p \in [0, 1]$ that sum to 1. We treat these as probabilities. For each pair i, we randomly choose exactly one path in P_i to route its traffic, where path p is chosen with probability x_p.
- **Analysis**: What is the congestion of this randomly chosen integer solution? Consider a single edge e. Let $X_{e,i}$ be an indicator RV that is 1 if the path chosen for pair i uses edge e.

$$\mathbb{E}[X_{e,i}] = \sum_{p : e \in p} x_p$$

  The total congestion on edge e is $X_e = \sum_{i=1}^{k} X_{e,i}$. By linearity of expectation:

$$\mathbb{E}[X_e] = \sum_{i=1}^{k} \mathbb{E}[X_{e,i}] \leq C_{opt}$$

The random variables $X_{e,1}, \ldots, X_{e,k}$ are independent. This is exactly the setup for a Chernoff bound! We have a sum of independent (Bernoulli) random variables. We can apply the bound to show that the probability of X_e being much larger than its mean (C_{opt}) is exponentially small.

$$\Pr(X_e \geq (1+\delta)C_{opt}) \leq \exp\left(-\frac{\delta^2 C_{opt}}{2+\delta}\right)$$

By choosing δ appropriately (e.g., $\delta = \frac{c \log m}{C_{opt}}$), we can make this probability smaller than $1/m^2$. Then, by a union bound over all m edges, the probability that *any* edge has high congestion is less than $m \cdot (1/m^2) = 1/m$, which goes to 0 for large m.

This proves that this simple randomized algorithm produces a routing whose congestion is within a $O(\log m)$ factor of the optimal solution, with high probability.