1 Introduction

In this lecture, we are interested in studying the deviations of a random variable from it's mean and deviations of the average of random variables from their expectation. Concentration inequalities bound the probability of such deviations. Two notions of concentration are usually studied, asymptotic and non-asymptotic. Non asymptotic analysis deals with the concentration inequalities of averages of finite number of random variables. Asymptotic analysis studies the convergence of the averages of infinite sequences of random variables.

2 Asymptotic Analysis

Given a sequence of random variables $\{X_n\}$ and another random variable Y, the following two notions of convergence can be defined.

Definition 9.1 (Convergence in Probability)**.** $\{X_n\}$ converges in probability to Y if for every $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}(|X_n - Y| > \epsilon) = 0 \tag{9.1}$$

The above is also denoted by $X_n \xrightarrow{\text{p}} Y$.

Definition 9.2 (Convergence in Distribution)**.** Let $F_X(.)$ denote the CDF of a random variable X. $\{X_n\}$ converges in distribution to Y if

$$\lim_{n\to\infty} F_{X_n}(t) = F_Y(t) \tag{9.2}$$

for all points t where the F_Y is continuous. The above is also denoted by $X_n \xrightarrow{\text{d}} Y$.

2.1 Weak law of large numbers

The weak law of large numbers states that the average of i.i.d random variables converges in probability to their mean.

Theorem 9.3 (Weak law of large numbers)**.** *Let S_n denote the sum of n i.i.d random variables, each with mean μ and variance $\sigma^2 < \infty$, then*

$$\frac{S_n}{n} \xrightarrow{p} \mu \tag{9.3}$$

2.2 Central limit theorem

Let $N(0,1)$ denote the standard normal variable with mean 0 and variance 1, i.e. its probability density function is given by $\frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$. The central limit theorem gives us an idea about the distribution of the sum of a large collection of i.i.d random variables.

Theorem 9.4 (Central limit theorem)**.** *Let S_n denote the sum of n i.i.d random variables, each with mean μ and variance $\sigma^2 < \infty$, then*

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0,1) \tag{9.4}$$

3 Non Asymptotic Convergence Bounds

Markov's inequality gives a tail bound for a single random variable. This basic result is used for proving more involved tail bounds for average of random variables.

Theorem 9.5 (Markov's Inequality)**.** *Let X be a non negative random variable and $\lambda > 0$, then*

$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}(X)}{\lambda} \tag{9.5}$$

Proof. Let $f(x)$ be the probability density function of X.

$$\begin{aligned}
\mathbf{E}[X] &= \int_0^\infty x f(x) dx, \quad \text{since } X \geq 0 \\
&\geq \int_\lambda^\infty x f(x) dx \\
&\geq \lambda \int_\lambda^\infty f(x) dx = \lambda \, \mathbf{Pr}[X \geq \lambda]
\end{aligned}$$

□

4 Moment Inequalities

Markov's inequality bounds the tail of a random variable given it's mean. If we have more information about the higher order moments of the random variable, we can get stronger tail bounds.

Theorem 9.6 (Chebychev's inequality)**.** *For any random variable X with mean μ and variance σ^2, we have*

$$\mathbf{Pr}[|X - \mu| \geq \lambda] \leq \frac{\sigma^2}{\lambda^2}$$

Proof. Let $Y = (X - \mu)^2$ be a random variable. Now using Markov's inequality we get

$$\mathbf{Pr}[Y \geq \lambda^2] \leq \frac{\mathbf{E}[Y]}{\lambda^2}$$

However, note that $\mathbf{Pr}[Y \geq \lambda^2] = \mathbf{Pr}[|X - \mu| \geq \lambda]$. □

Theorem 9.7 (Moment inequalities)**.** *For any random variable X with mean μ and finite moments upto order $2k$ for any positive integer k, we have*

$$\mathbf{Pr}[|X - \mu| \geq \lambda] \leq \frac{\mathbb{E}((X - \mu)^{2k})}{\lambda^{2k}}$$

Proof. Let $Y = (X - \mu)^{2k}$ be a random variable. Now using Markov's inequality we get

$$\mathbf{Pr}[Y \geq \lambda^{2k}] \leq \frac{\mathbf{E}[Y]}{\lambda^{2k}}$$

However, note that $\mathbf{Pr}[Y \geq \lambda^{2k}] = \mathbf{Pr}[|X - \mu| \geq \lambda]$. □

We get stronger tail bounds for large values of k, however it becomes increasingly tedious to compute $E((X - \mu)^{2k})$.

4.1 Examples

Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli random variables with $\mathbf{Pr}[X_i = 0] = 1-p$ and $\mathbf{Pr}[X_i = 1] = p$. Let $S_n := \sum_i^n X_i$. Then S_n is distributed as a binomial random variable $Bin(n, p)$. Note that $\mathbf{E}[S_n] = np$ and $\mathbf{Var}[S_n] = np(1-p)$.

Example 1 $\left(Bin(n, \frac{1}{2})\right)$: Here Markov's inequality gives a bound on the probability that S_n is away from its mean $\frac{n}{2}$ as $\mathbf{Pr}[S_n - \frac{n}{2} \geq \beta n] \leq \frac{n/2}{n/2+\beta n} = \frac{1}{1+2\beta}$. However, Chebychev's inequality gives a much tighter bound as $\mathbf{Pr}[|S_n - \frac{n}{2}| \geq \beta n] \leq \frac{n/4}{\beta^2 n^2} = \frac{1}{4\beta^2 n}$.

Example 2 (Balls and Bins): Suppose we throw n balls uniformly at random into n bins. Then for a fix bin i the number of balls in it is distributed as a $Bin(n, \frac{1}{n})$ random variable. Markov's inequality gives a bound on the probability that S_n is away from its mean 1 (i.e. the number of balls in bin i deviates from its expected value) as $\mathbf{Pr}[S_n - 1 \geq \lambda] \leq \frac{1}{1+\lambda}$. However, Chebychev's inequality gives a much tighter bound as $\mathbf{Pr}[|S_n - 1| \geq \lambda] \leq \frac{(1-1/n)}{\lambda^2}$.

Example 3 (Random Walk): Suppose we start at the origin and at each step move a unit distance either left or right uniformly randomly and independently. We can then ask about the behaviour of the final position after n steps. Each step (X_i) can be modelled as a Rademacher random variable with the following distribution.

$$X_i = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2} \end{cases}$$

The position after n steps is $S_n = \sum_{i=1}^n X_i$. Note that the mean and standard deviation of S_n are 0 and $\sqrt{n}$ respectively. Applying Chebyshev's inequality on S_n, we get

$$\mathbb{P}(S_n > t\sqrt{n}) \leq \frac{1}{t^2}$$

. However we can do better by applying moment inequalities.

$$\mathbb{E}((S_n)^4) = \mathbb{E}(\sum_{i=1}^{n} X_i) \tag{9.6}$$

$$= \mathbb{E}(\sum_i X_i^4 + c_1 \sum_{i<j} X_i^3 X_j + c_2 \sum_{i<j} X_i^2 X_j^2 + c_3 \sum_{i<j<k} X_i^2 X_j X_k + c_4 \sum_{i<j<k<l} X_i X_j X_k X_l) \tag{9.7}$$

$$= n + c_2 \binom{n}{2} \tag{9.8}$$

Substituting this value of expectation in the fourth order moment inequality, we get a stronger tail bound.

$$\mathbb{P}[|S_n| \geq t\sqrt{n}] \leq \frac{\mathbb{E}((S_n)^4)}{t^4 n^2} \leq \frac{n + c_2\binom{n}{2}}{t^4 n^2} \leq \frac{\Theta(1)}{t^4}$$

We can analyze how good these bounds are by explicitly computing $\mathbb{P}(S_n = k)$. All possible sequence of steps can be thought of as different ways of choosing positions for $+1$ steps.

$$\frac{\mathbb{P}(S_n = 2\lambda)}{\mathbb{P}(S_n = 0)} = \frac{\binom{n}{\frac{n}{2}+\lambda}}{\binom{n}{\frac{n}{2}}}.$$

We can approximate the above for large n with Stirling's formula $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$

$$\frac{\mathbb{P}(S_n = 2\lambda)}{\mathbb{P}(S_n = 0)} \sim \frac{(\frac{n}{2})^{(\frac{n}{2})}(\frac{n}{2})^{(\frac{n}{2})}}{(\frac{n}{2}+\lambda)^{(\frac{n}{2}+\lambda)}(\frac{n}{2}-\lambda)^{(\frac{n}{2}-\lambda)}} = \frac{1}{(1+\frac{2\lambda}{n})^{\frac{n}{2}+\lambda}(1-\frac{2\lambda}{n})^{\frac{n}{2}-\lambda}}$$

If $\lambda << n$, then we can approximate $1 + k\frac{\lambda}{n}$ by $e^{k\frac{\lambda}{n}}$

$$\frac{\mathbb{P}(S_n = 2\lambda)}{\mathbb{P}(S_n = 0)} \sim e^{\frac{-2\lambda}{n}(\frac{n}{2}+\lambda)}e^{\frac{2\lambda}{n}(\frac{n}{2}-\lambda)} = e^{\frac{-4\lambda^2}{n}}$$

Substituting $\lambda = t\sqrt{n}$, we get

$$\mathbb{P}(S_n = 2\lambda) \sim \mathbb{P}(S_n = 0)e^{-4t^2}$$

This shows that most of the probability mass lies in the region $|S_n| < t\sqrt{n}$, and drops off exponentially as we go far off.

5 Chernoff bounds - Hoeffding's inequality

Theorem 9.8 (Chernoff bounds - Hoeffding's inequality)**.** [1] *Let $X_1, X_2, \ldots, X_n$ be n independent random variables taking values in $[0,1]$. Let $S_n := X_1 + X_2 + \ldots + X_n$, $\mu_i := \mathbf{E}[X_i]$, and $\mu := \mathbf{E}[S_n] = \sum_i \mathbf{E}[X_i]$. Then for any $\beta \geq 0$ we have*

$$\textit{Upper tail}: \qquad \mathbf{Pr}[S_n \geq \mu(1+\beta)] \leq \exp\left(-\frac{\beta^2\mu}{2+\beta}\right) \tag{9.9}$$

$$\textit{Lower tail}: \qquad \mathbf{Pr}[S_n \leq \mu(1-\beta)] \leq \exp\left(-\frac{\beta^2\mu}{3}\right) \tag{9.10}$$

Before proving the above theorem, we consider its application for example 1 ($Bin(n, \frac{1}{2})$) mentioned in the previous section. The upper tail of the above theorem implies that $\mathbf{Pr}[S_n - \frac{n}{2} \geq \frac{\beta n}{2}] \leq \exp(-\frac{\beta^2 n/2}{2+\beta})$. Clearly this exponentially bound seems more prominent than the polynomial one achieved by Markov's or Chebychev's inequality.

Proof. We only prove Eq. (9.9). The proof for Eq. (9.10) is similar.

$$\begin{aligned}
\mathbf{Pr}[S_n \geq \mu(1+\beta)] &= \mathbf{Pr}[e^{tS_n} \geq e^{t\mu(1+\beta)}] \quad \forall t > 0 \\
&\leq \frac{\mathbf{E}[e^{tS_n}]}{e^{t\mu(1+\beta)}} \quad \text{(using Markov's inequality)} \\
&= \frac{\prod \mathbf{E}[e^{tX_i}]}{e^{t\mu(1+\beta)}} \quad \text{(using independence)}
\end{aligned}$$

Assumption: For now we assume that all $X_i \in \{0,1\}$, i.e. are Bernoulli random variables. We will later show how to remove this assumption.

Now using the above assumption we get $\mathbf{E}[e^{tX_i}] = 1 + \mu_i(e^t - 1) \leq \exp(\mu_i(e^t - 1))$. Hence, we get

$$\begin{aligned}
\mathbf{Pr}[S_n \geq \mu(1+\beta)] &\leq \frac{\prod \mathbf{E}[e^{tX_i}]}{e^{t\mu(1+\beta)}} \\
&\leq \frac{\prod \exp(\mu_i(e^t-1))}{e^{t\mu(1+\beta)}} \\
&= \exp(\mu(e^t - 1) - t\mu(1+\beta))
\end{aligned}$$

[1] In his paper Chernoff derive the corresponding inequality in the case that $X_1, ..., X_n$ are i.i.d Bernoulli random variables. Hoeffding gave the generalization where $X_1, ..., X_n$ are independent random variables all taking values in some bounded interval $[a, b]$.

Since the above expression holds for all positive t and we wish to minimize it. By setting its derivative w.r.t. t to zero we obtain $t = ln(1+\beta)$. This gives

$$\mathbf{Pr}[S_n \geq \mu(1+\beta)] \leq \left(\frac{e^{\beta}}{(1+\beta)^{1+\beta}}\right)^{\mu} \tag{9.11}$$

Now observe that for $x \geq 0$ we have that $\frac{x}{1+\frac{x}{2}} \leq ln(1+x)$. Hence, we can simplify the above expression for $x = \beta$ to obtain

$$\mathbf{Pr}[S_n \geq \mu(1+\beta)] \leq \exp\left(-\frac{\beta^2\mu}{2+\beta}\right)$$

Removing the assumption $X_i \in \{0,1\}$: For each i in $[n]$, we define a new Bernoulli random variable Y_i which take value 0 with probability $1-\mu_i$ and value 1 with probability μ_i. You can think of Y_i as being formed by starting with probability density function of X_i and then moving the mass from every point in $(0,1)$ to the endpoints $0,1$ in a way that preserve the mean. Now note that the function e^{tX_i} is convex for every value of $t \geq 0$. Thus we have $\mathbf{E}[e^{tX_i}] \leq \mathbf{E}[e^{tY_i}] = 1+\mu_i(e^t-1) \leq \exp(\mu_i(e^t-1))$, and the above proof goes through even for the general case where $x \in [0,1]$. In the case that $X_1, ..., X_n$ are n independent variables that take values in $[a,b]$ we can define $Y_i = \frac{X_i - a}{b-a}$. Now $Y_1, ..., Y_n$ are independent random variables that take values in $[0,1]$. Furthermore with $S_n = \sum_i = 1^n X_i$ and $S'_n = \sum_{i=1}^n Y_i$ we have that $(b-a)S_n + na = S'_n$. Hence $\mathbf{Pr}(S_n \geq (1+\beta)\mu) = \mathbf{Pr}[S'_n \geq ((1+\beta)\mu - na)/(b-a)]$. The latest probability can be calculated using Hoeffding's inequality. □

Example 3 (Balls and Bins): With the same setting as in Example 2 we now want to bound the maximum number of balls falling into a bin. The expected number of balls into any bin is 1. Thus Chernoff bounds imply that the probability that:

$$\mathbf{Pr}[\text{Balls in bin } i \geq 1+\beta] \leq \exp\left(-\frac{\beta^2}{2+\beta}\right)$$

If we ensure that the above probability is less than $\frac{1}{n^2}$ (i.e. $\beta = O(\log n)$) then even if we take union bound over all the bins, we get that the probability that a bin receives at least $1+\beta$ balls is at most $\frac{1}{n}$. Hence, we have with high probability that no bin receives more than $O(\log\ n)$ balls. A better bound for this problem is $(1+o(1))\left(\frac{\log\ n}{\log\log n}\right)$, which can be obtained by using the stronger bound given in Eq. (9.11). Furthermore this bound is tight in the sense that w.h.p. there is a bin with load $(1+o(1))\frac{\log n}{\log\log n}$ [KJ77]. This fact has two immediate implications. First Eq. (9.11) found in the proof of Hoeffding's inequality can is stronger than Hoeffding's inequality. Second any inequality of the form $\mathbf{Pr}[S_n \geq \mu(1+\beta)] \leq \exp(-C\beta^2\mu)$ for some constant $C > 0$ does not hold. That is because any such inequality would imply that the maximum load would be $O(\log^{0.5} n)$.

Remark: Hoeffding's inequality also holds if the random variables are not independent but negatively correlated, i.e. if some variables are 'high' then it makes more likely for the other variables to be 'low'. Formally X_i and X_j are negatively correlated if for all disjoint sets A, B and for all monotone increasing functions f, g, we have

$$\mathbf{E}[f(X_i : i \in A)g(X_j : j \in B)] \leq \mathbf{E}[f(X_i : i \in A)]\,\mathbf{E}[g(X_j : j \in B)].$$

6 Other concentration bounds

Theorem 9.9 (Bernstein's inequality [McD98]). *Consider n independent random variables $X_1, X_2, \ldots, X_n$ with $|X_i - \mathbf{E}[X_i]| \leq b$ for each i. Let $S_n := X_1 + X_2 + \ldots + X_n$, and let S_n have mean μ variance σ^2. Then for any $\beta \geq 0$ we have*

$$\textit{Upper tail}: \qquad \mathbf{Pr}[S_n \geq \mu(1+\beta)] \leq exp\left(-\frac{\beta^2\mu}{2\sigma^2/\mu + 2\beta b/3}\right)$$

Theorem 9.10 (McDiarmid's inequality [McD98]). *Consider n independent random variables $X_1, X_2, \ldots, X_n$ with X_i taking values in a set A_i for each i. Suppose a real valued function f is defined on $\prod A_i$ satisfying $|f(x) - f(x')| \leq c_i$ whenever x and x' differ only in the ith coordinate. Let μ be the expected value of the random variable $f(X)$. Then for any non-negative β we have*

$$\textit{Upper tail}: \qquad \mathbf{Pr}[f(X) \geq \mu(1+\beta)] \leq exp\left(-\frac{2\mu^2\beta^2}{\sum_i c_i^2}\right)$$

$$\textit{Lower tail}: \qquad \mathbf{Pr}[f(X) \leq \mu(1-\beta)] \leq exp\left(-\frac{2\mu^2\beta^2}{\sum_i c_i^2}\right)$$

Theorem 9.11 (Philips and Nelson [PN95] show moment bounds are tighter than Chernoff-Hoeffding bounds). *Consider n independent random variables $X_1, X_2, \ldots, X_n$, each with mean 0. Let $S_n = \sum X_i$. Then*

$$\mathbf{Pr}[S_n \geq \lambda] \leq \min_{k \geq 0} \frac{\mathbf{E}[X^k]}{\lambda^k} \leq \inf_{t \geq 0} \frac{\mathbf{E}[e^{tX}]}{e^{t\lambda}}$$

Theorem 9.12 (Matrix Chernoff bounds). *Consider n independent symmetric matrices $X_1, X_2, \ldots, X_n$ of dimension d. Moreover, $X_i \succeq 0$ and $I \succeq X_i$ for each i, i.e. eigenvalues are between 0 and 1. Let $\mu_{min} = \lambda_{min}(\sum \mathbf{E}[X_i])$ and $\mu_{max} = \lambda_{max}(\sum \mathbf{E}[X_i])$, then*

$$\mathbf{Pr}\left[\lambda_{max}\left(\sum X_i\right) \geq \mu_{max} + \gamma\right] \leq d\ exp\left(-\frac{\gamma^2}{2\mu_{max} + \gamma}\right)$$

In some applications the random variables are not independent, but have limited influence on the overall function. We can still give concentration bounds if the random variables form a martingale.

Theorem 9.13 (Hoeffding-Azuma inequality [McD98]). *Let $c_1, c_2, \ldots, c_n$ be n constants, and let $Y_1, Y_2, \ldots, Y_n$ be a martingale difference sequence with $|Y_i| \leq c_i$ for each i. Then for any $t \geq 0$*

$$\mathbf{Pr}\left[|\sum_{i=1}^{n} Y_i| \geq t\right] \leq 2\ exp\left(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}\right)$$

Remark:McDiarmid's inequality and Azuma-Hoeffding inequality can be used to bound functions of $X_1, ..., X_n$ other than their sum.

Acknowledgments

These lecture notes were scribed by Sarthak Garg, based on previous scribe notes of Michael Anastos and Sahil Singla.

Acknowledgments

These lecture notes were scribed by Sarthak Garg, based on previous scribe notes of Michael Anastos and Sahil Singla.

References

[KJ77] Samuel Kotz and Norman Lloyd Johnson. *Urn models and their applications*. John Wiley & Sons, 1977. 5

[McD98] Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998. 9.9, 9.10, 9.13

[PN95] Thomas K Philips and Randolph Nelson. The moment bound is tighter than Chernoff's bound for positive tail probabilities. *The American Statistician*, 49(2):175–178, 1995. 9.11