Lecturer: Anupam Gupta Scribe: Michael Anastos, Sahil Singla

1 Introduction

Consider n independent identically distributed (i.i.d.) random variables $X_1, X_2, \ldots, X_n$, each with mean μ. We are interested in the sum of these random variables $S_n := \sum_i^n X_i$. Note that $\mathbf{E}[S_n] = n\mu$ by linearity of expectation. From the week law of large numbers we know that as n tends to infinity, the random variable $\frac{S_n}{n}$ converges in probability to the mean μ, i.e. $\lim_{n\to\infty} \mathbf{Pr}[|S_n/n - \mu| > \epsilon] = 0$ for any positive constant ϵ. In this lecture we are interested quantify the concentration of S_n around its mean $n\mu$ for some finite n. Equivalently, we are interested in upper bounding the probability $\mathbf{Pr}[|S_n - n\mu| \geq \lambda]$ for some positive λ.

1.1 Central limit theorem

We say a sequence of random variables $\{X_n\}$ converges in distribution to a random variable Y (written as $X_n \xrightarrow{d} Y$) if $\forall u \in \mathbb{R}$

$$\mathbf{Pr}(X_n \geq u) \xrightarrow{n\to\infty} \mathbf{Pr}(Y \geq u)$$

Let $N(0,1)$ denote the standard normal variable ("Gaussian variable") with mean 0 and variance 1, i.e. its probability density function is given by $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$. The central limit theorem gives us an idea on how far S_n is from $n\mu$ as n tends to infinity.

Theorem 18.1 (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*

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

1.2 Markov's inequality

Markov's inequality is the most basic concentration bound.

Theorem 18.2 (Markov's inequality)**.** *Let X be a non-negative random variable, then*

$$\mathbf{Pr}[X \geq \lambda] \leq \frac{\mathbf{E}[X]}{\lambda}$$

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

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

□

1.3 Chebychev inequality

Theorem 18.3 (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 $\mathbf{Pr}[Y \geq \lambda^2] = \mathbf{Pr}[|X - \mu| \geq \lambda]$. □

Remark: One can obtain stronger inequalities than the Chebychev's inequality by taking higher moments and applying the Markov's inequality. In particular, we may define a random variable $Y = (X - \mu)$. Then for every positive integer t we have $\mathbf{Pr}[|X - \mu| \geq \lambda] = \mathbf{Pr}[Y^{2t} \geq \lambda^{2t}] \leq \frac{\mathbf{E}[Y^{2t}]}{\lambda^{2t}}$. Such inequalities are commonly called generalized Chebychev or moment inequality. The problem with this approach is that calculating $\mathbf{E}[Y^{2t}]$ becomes tedious for large values of t.

1.4 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 $\mathbf{Pr}[|S_n - 1| \geq \lambda] \leq \frac{(1-1/n)}{\lambda^2}$.

2 Chernoff bounds - Hoeffding's inequality

Theorem 18.4 (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*

$$\textit{Upper tail}: \qquad \mathbf{Pr}[S_n \geq \mu(1 + \beta)] \leq \exp\left(-\frac{\beta^2 \mu}{2 + \beta}\right) \tag{18.1}$$

$$\textit{Lower tail}: \qquad \mathbf{Pr}[S_n \leq \mu(1 - \beta)] \leq \exp\left(-\frac{\beta^2 \mu}{3}\right) \tag{18.2}$$

[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 $[a, b]$.

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. (18.1). The proof for Eq. (18.2) 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

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

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

$$\mathbf{Pr}[S_n \geq \mu(1+\beta)] \leq \left(\frac{e^\beta}{(1+\beta)^{1+\beta}}\right)^\mu \tag{18.3}$$

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. (18.3). Furthermore this bound is tight in the sense that w.h.p. there is a bin with load $(1+o(1))\frac{\log n}{\log\log n}$ [4]. This fact has two immediate implications. First Eq. (18.3) 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

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

3 Other concentration bounds

Theorem 18.5 (Bernstein's inequality [5]). *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 18.6 (McDiarmid's inequality [5]). *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 18.7 (Philips and Nelson [6] 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 18.8 (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 18.9 (Hoeffding-Azuma inequality [5])**.** *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 $t \geq 0$*

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

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

References

[1] S. N. Bernstein. Gastehizdat Publishing House, 1946.

[2] H. Chernoff. A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observations,. *Ann. Math. Stat.*, 23:493–509, 1952.

[3] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–60, 1963.

[4] S. Kotz and N. L. Johnson. *Urn models and their applications.* John Wiley & Sons, 1977. 2

[5] C. McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998. 18.5, 18.6, 18.9

[6] T. K. Philips and R. Nelson. The moment bound is tighter than Chernoff's bound for positive tail probabilities. *The American Statistician*, 49(2):175–178, 1995. 18.7