15-859E: Advanced Algorithms **CMU, Spring 2015**
Lecture #18: Concentration Bounds: "Chernoff-Hoeffding" February 23, 2015
Lecturer: Anupam Gupta Scribe: 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 $S_n := \sum_i X_i$. Note that $\mathbf{E}[S_n] = n\mu$ by linearity of expectation. From the 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 in understanding how far do we expect S_n to be from its mean $n\mu$ for some finite n. To be more precise, 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 probability to a random variable Y (written as $\lim_{n\to\infty} X_n \to Y$) if for any constant $\epsilon > 0$

$$\lim_{n\to\infty} \mathbf{Pr}[|X_n - Y| \geq \epsilon] = 0$$

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. Then*

$$\lim_{n\to\infty} \frac{S_n - n\mu}{\sqrt{n}\sigma} \to N(0,1)$$

1.2 Markov's inequality

Markov's inequality is the most basic concentration bound.

Theorem 18.2 (Markov's inequality)**.** *For any non-negative random variable X, we have*

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

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 define a random variable $Y = (X-\mu)^{2t}$ for some positive integer t and use $\mathbf{Pr}[|X-\mu| \geq \lambda] = \mathbf{Pr}[Y \geq \lambda^{2t}] \leq \frac{\mathbf{E}[Y^{2t}]}{\lambda^{2t}}$. 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

Consider n i.i.d. Bernoulli random variables $X_1, X_2, \ldots, X_n$, i.e. $X_i \in \{0,1\}$ for each i, with $\mathbf{Pr}[X_i = 0] = 1-p$ and $\mathbf{Pr}[X_i = 1] = p$. Let $S_n := Bin(n,p) := \sum_i X_i$ be the sum of these random variables. 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 $\left(Bin(n, \frac{1}{n})\right)$: Here Markov's inequality gives a bound on the probability that S_n is away from its mean 1 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-Hoeffding" bounds

Theorem 18.4 ("Chernoff-Hoeffding" bounds)**.** *Consider n independent $[0,1]$ random variables $X_1, X_2, \ldots, X_n$. Let $S_n := X_1 + X_2 + \ldots + X_n$, let $\mu_i := \mathbf{E}[X_i]$, and let $\mu := \mathbf{E}[S_n] = \sum_i \mathbf{E}[X_i]$. Then for any non-negative β we have*

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

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

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 theorem gives us $\mathbf{Pr}[S_n - \frac{n}{2} \geq \frac{\beta n}{2}] \leq exp(-\frac{\beta^2 n/2}{2+\beta})$. Clearly this bound is exponentially stronger than Markov's or Chebychev's inequality for any constant β.

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 that this is actually the worst possible case.

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 is true for all positive t and we wish to minimize it, we put its derivative w.r.t. t to zero and obtain $t = ln(1+\beta)$. This gives

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

We make another observation that for all positive x the following is true $\frac{x}{1+\frac{x}{2}} \leq ln(1+x)$. 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 is 0 with probability $1-\mu_i$ and is 1 with probability μ_i. Now note that the function e^{tX_i} is convex for any positive value of t. 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 $x \in [0,1]$.

□

Example 3 (Balls and Bins): Suppose we throw n balls uniformly at random into n bins. The problem is to bound the maximum number of balls falling into a bin. Here we observe that the probability that a given ball falls into a given bin is $\frac{1}{n}$. Hence, the expected number of balls into any bin is 1. Now we use Chernoff-Hoeffding inequality to bound the probability that bin i receives at least $1+\beta$ balls:

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

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. The correct answer for this problem is actually $O\left(\frac{log\ n}{log\ log\ n}\right)$, which can be obtained by using the stronger bound given in Eq. (18.3).

Remark: Chernoff-Hoeffding 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, for all disjoint sets A, B and monotone increasing functions f, g, we want

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

3 Other concentration bounds

Theorem 18.5 (Bernstein's inequality [1]). *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 non-negative β 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 [1]). *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 [2] 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 [1]). *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 $t \geq 0$*

$$\mathbf{Pr}\left[|\sum_i Y_i| \geq t\right] \leq 2\ exp\left(-\frac{t^2}{2\sum_i c_i^2}\right)$$

References

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

[2] 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