Probabilistic Method and Random Graphs

Lecture 3. Chernoff bounds: behind and beyond

Xingwu Liu

Institute of Computing Technology
Chinese Academy of Sciences, Beijing, China

Preface

Questions, comments, or suggestions?

A brief review

Moments

Expectation, k-moment, variance

Inequalities

Universal: Union bound
1-moment: Markov's inequality
2-moment: Chebychev's inequality

Chernoff bounds: independent sum

Let $X = \sum_{i=1}^{n} X_i$, where $X_i's$ are **independent** Poisson trials. Let $\mu = \mathbb{E}[X]$. Then

1. For $\delta > 0$, $\Pr(X \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \leq e^{-\frac{\delta^2}{2+\delta}\mu}$.
2. For $1 > \delta > 0$, $\Pr(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} \leq e^{-\frac{\delta^2}{2}\mu}$.

General bounds for independent sums

Each $X_i \in \{0, a_i\}$ where $a_i \leq 1$

Basic Chernoff bounds remain valid, by Homework 2 of Week 2.

Each $X_i \in [0, 1]$ but is not necessarily a Poisson trial

Basic Chernoff bounds remain valid (by $e^{\lambda x} \leq x e^{1\lambda} + (1-x)e^{0\lambda}$).

The domains (a_i, b_i) of X_i's differ

Hoeffding's Inequality: $\Pr(|X - \mathbb{E}[X]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_i (b_i - a_i)^2}}$.
Proposed in 1963.

Remarks of Hoeffding's Inequality

1. It considers the absolute, rather than relative, deviation.
Particularly useful if $\mu = 0$.
2. When each $X_i \in [0, s]$, it is tighter than the simplified basic
Chernoff bounds if δ is big, and looser otherwise.

Hoeffding's Inequality

Let $X = \sum_{i=1}^{n} X_i$, where $X_i \in [a_i, b_i]$ are independent r.v. Then $\Pr(|X - \mathbb{E}[X]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_i (b_i - a_i)^2}}$ for any $t > 0$

Idea of the proof

1. Given r.v. $Z \in [a, b]$ with $\mathbb{E}[Z] = 0$, $\mathbb{E}[e^{\lambda Z}] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}$
   -Hoeffding's Lemma
2. $$\begin{aligned}\Pr(X - \mathbb{E}[X] \geq t) &\leq \frac{\prod_i \mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}]}{e^{\lambda t}} \\ &\leq e^{\lambda^2 \sum_i \frac{(b_i - a_i)^2}{8} - \lambda t}\end{aligned}$$
3. Choose λ to minimize RHS. Likewise for $\Pr(X - \mathbb{E}[X] \leq -t)$.

Proof of Hoeffding's Lemma

Lemma: Given r.v. $Z \in [a, b]$ with $\mathbb{E}[Z] = 0$, $\mathbb{E}[e^{\lambda Z}] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}$

$e^{\lambda z} \leq \frac{z-a}{b-a} e^{\lambda b} + \frac{b-z}{b-a} e^{\lambda a}$, for $z \in [a, b]$

$$
\begin{aligned}
\mathbb{E}[e^{\lambda Z}] &\leq \frac{b e^{\lambda a}}{b-a} - \frac{a e^{\lambda b}}{b-a} \\
&= (1 - \theta + \theta e^{u}) e^{-\theta u} \quad \text{where } \theta = \frac{-a}{b-a}, u = \lambda(b-a) \\
&= e^{\phi(u)} \quad \text{where } \phi(u) \triangleq -\theta u + \ln(1 - \theta + \theta e^{u})
\end{aligned}
$$

Taylor expansion $\phi(u) = \phi(0) + \phi'(0) + \frac{\phi''(\xi)}{2} u^2$.
Then $\phi(u) \leq \frac{u^2}{8}$ since $\phi(0) = \phi'(0) = 0, \phi''(\xi) \leq \frac{1}{4}$

Example: Hoeffding's Inequality + Union bound

Set balancing

Given a matrix $A \in \{0,1\}^{n\times m}$, find $b \in \{-1,1\}^m$ s.t. $\| Ab \|_\infty$ is minimized.

Motivation

$$\begin{array}{r} \text{feature } 1: \\ \text{feature } 2: \\ \vdots \\ \text{feature } n: \end{array} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}, \text{ each column is an object.}$$

Want to partition the objects so that every feature is balanced.

Example: Hoeffding's Inequality + Union bound

Set balancing

Given a matrix $A \in \{0,1\}^{n\times m}$, find $b \in \{-1,1\}^m$ s.t. $\| Ab \|_\infty$ is minimized.

Motivation

$$\begin{array}{l} \text{feature 1:} \\ \text{feature 2:} \\ \vdots \\ \text{feature } n\text{:} \end{array} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}, \text{ each column is an object.}$$

Want to partition the objects so that every feature is balanced.

Algorithm

Uniformly randomly sample b.

Performance analysis

Performance

$\Pr(\| Ab \|_\infty \geq \sqrt{4m \ln n}) \leq \frac{2}{n}$

Proof

For any $1 \leq i \leq n$, $Z_i = \sum_j a_{ij} b_j$ is the ith entry of Ab. By union bound, it suffices to prove $\Pr(|Z_i| \geq \sqrt{4m \ln n}) \leq \frac{2}{n^2}$ for each i.

Fix i. W.l.o.g, assume $a_{ij} = 1$ iff $1 \leq j \leq k$ for some $k \leq m$. Then $Z_i = b_1 + ... + b_k$.

Note that b_j's are independent over $\{-1, 1\}$ with $\mathbb{E}[b_j] = 0$.

By Hoeffding's Inequality, $\Pr(|Z_i| \geq \sqrt{4m \ln n}) \leq 2e^{-\frac{8m \ln n}{4k}} \leq \frac{2}{n^2}$

Reflection on moments and Chernoff bounds

Chernoff Bounds

Why is it so good?
Can it be improved by non-exponential functions?
Anything to do with moments?

Moments

Do moments uniquely determine the distribution?

The story begins with generating functions.

Generating functions

Informal definition

A power series whose coefficients encode information about a sequence of numbers.

Example: Probability generating function

Given a discrete random variable X whose values are non-negative integers, $G_X(t) \triangleq \sum_{n \geq 0} \Pr(X = n) t^n = \mathbb{E}[t^X]$.
Example: Bernoulli and binomial random variables.

Properties

Convergence: It converges if $|t| < 1$.
Uniqueness: $G_X(\cdot) \equiv G_Y(\cdot)$ implies the same distribution.

Application

Toy: Use uniqueness to show that the summation of independent identical binomial distribution is binomial.
Deriving Moments: $G_X^{(k)}(1) = \mathbb{E}[X(X-1) \cdots (X-k+1)]$.

Moment generating functions

Shortcoming of probability generating functions

Only valid for non-nagetive integer random variables.

Moment generating functions

$M_X(t) \triangleq \sum_x \Pr(X = x)e^{tx} = \mathbb{E}[e^{tX}]$.
Example of Bernoulli and binomial distributions.

Properties

- If $M_X(t)$ converges around 0, $M_X^{(k)}(0) = \mathbb{E}[X^k]$, meaning the moments are exactly the coefficients of the Taylor's expansion.
- **Convergence**: $M_X(t)$ converges when X is bounded.
- If independent, $M_{X+Y} = M_X M_Y$.
- **Uniqueness**: If $M_X(t)$ converges around 0, the distribution is uniquely determined by the moments. (Why? See later)

But

Moments generating function may not converge

Cauchy distribution: density function $f(x) = \frac{1}{\pi(1+x^2)}$ does not have moments for any order.

An example of non-uniqueness of moments

Log-Normal-like distribution:

density function $f_{X_n}(x) = \frac{e^{-\frac{1}{2}(\ln x)^2}}{\sqrt{2\pi x}}(1 + \sin(2n\pi \ln x))$.

k-Moments $\mathbb{E}[X_n^k] = e^{k^2/2}$ for non-negative integers k.

Characteristic functions

Definition

$\varphi_X(t) \triangleq \int_{\mathbb{R}} e^{itx} dF_X(x)$ where $i = \sqrt{-1}$ and t is real.

Properties

Convergence: It always exists.
Uniqueness: It uniquely determines the distribution.
Rationale of the uniqueness.

Uniqueness of convergent moments generating functions

Suppose $M_X(t) = M_Y(t)$ converges around 0.

- $\phi_X(t)$ and $\phi_Y(t)$ can be extended to the belt with small imaginary part (since formally, $M_X(t) = \phi_X(it)$)
- $\phi_X(t) = \phi_Y(t)$ when t is purely imaginary in this belt
- By the unique continuation theorem of analytic complex functions, the characteristic functions are equal

Ready to get insights

Moments

Do moments uniquely determine the distribution?
Yes, but conditionally.

Chernoff Bounds

- Why is it so good?
- Can it be improved by non-exponential functions?
- Anything to do with moments?

What's your answer?

A story of generating function

Introduced in 1730 by Abraham de Moivre, to solve the general linear recurrence problem

Wisdom: A generating function is a clothesline on which we hang up a sequence of numbers for display. -Herbert Wilf

Application to Fibonacci numbers (by courtesy of de Moivre):
$F(x) = \sum_{n=0}^{\infty} F_n x^n = x + \sum_{n=2}^{\infty} (F_{n-1} + F_{n-2}) x^n =$
$x + xF(x) + x^2 F(x)$
$\Rightarrow F(x) = \frac{x}{1-x-x^2} = \frac{1}{\sqrt{5}} \left(\frac{\phi}{x+\phi} - \frac{\psi}{x+\psi} \right) = \sum_{n=0}^{\infty} \frac{1}{\sqrt{5}} \left(\phi^n - \psi^n \right) x^n$
$\Rightarrow F_n = \frac{1}{\sqrt{5}} \left(\phi^n - \psi^n \right).$

Brief introduction to Abraham de Moivre

- May 26, 1667-Nov. 27, 1754
- A French mathematician

- de Moivre's formula
- Binet's formula
- Central limit theorem
- Stirling's formula

Legend

- Friends: Isaac Newton, Edmond Halley, and James Stirling
- Struggled for a living and lived for mathematics
- The Doctrine of Chances was prized by gamblers
  - 2nd probability textbook in history
- Predicted the exact date of his death

Chernoff bound in a big picture

Fundamental laws of probability theory

Law of large numbers (Cardano, Jacob Bernoulli 1713, Poisson 1837): The sample average converges to the expected value.

Central limit theorem (Abraham de Moivre 1733, Laplace 1812, Lyapunov 1901, Pólya 1920): The arithmetic mean of independent random variables is approximately normally distributed.

$$\lim_{n\to\infty} \Pr\left(\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right) \leq x\right) = \Phi\left(\frac{x}{\sigma}\right)$$

Marvelous but ...

Say nothing about the rate of convergence

Large deviation theory

How fast does it converge? Beyond central limit theorem

A glance at large deviation theory

Motivation

X_n: the number of heads in n flips of a fair coin.
By the central limit theorem, $\Pr(X_n \geq \frac{n}{2} + \sqrt{n}) \to 1 - \Phi(1)$.
What about $\Pr(X_n \geq \frac{n}{2} + \frac{n}{3})$? Nothing but converging to 0.

Chernoff bounds say...

$$\Pr(X_n \geq \tfrac{n}{2} + \tfrac{n}{3}) \leq \left(\frac{e^{\frac{2}{3}}}{\left(\frac{5}{3}\right)^{\frac{5}{3}}} \right)^{\frac{n}{2}} \approx e^{-0.092n}.$$

Actually

Direct calculation shows that
$\Pr(X_n \geq \frac{n}{2} + \frac{n}{3}) \approx e^{-0.2426n+o(n)} \ll \text{Chernoff bound.}$

Oh, no!

Mission of Large Deviation Theory

Find the asymptotic probabilities of *rare* events - how do they decay to 0 as $n \to \infty$?

Rare events mean large deviation.
So large that CLT is almost useless (deviation up to $\sqrt{n}$).

Intuition

Inspired by Chernoff bounds, conjecture that probabilities of rare events will be exponentially small in n : e^{-cn} for some c.
Q: Does $\lim_{n\to\infty} \frac{1}{n} \ln \Pr(\mathcal{E}_n^{\text{rare}})$ exist? If so, what's it?

Large Deviation Principle

Simple form (By courtesy of Cramer, 1938)

Let $X_1, ... X_n, ... \in \mathbb{R}$ be i.i.d. r.v. which satisfy $\mathbb{E}[e^{tX_1}] < \infty$ for $t \in \mathbb{R}$. Then for any $t > \mathbb{E}[X_1]$, we have

$$\lim_{n\to\infty} \frac{1}{n} \ln \Pr\left(\sum_{i=1}^{n} X_i \geq tn\right) = -I(t),$$

where

$$I(t) \triangleq \sup_{\lambda>0} \lambda t - \ln \mathbb{E}[e^{\lambda X_1}].$$

Remark

$I(\cdot)$: rate function.
Many variants: the factor $\frac{1}{n}$, random variables

Large Deviation Principle: Proof

Large Deviation Principle

$\lim_{n\to\infty} \frac{1}{n} \ln \Pr(\sum_{i=1}^{n} X_i \geq tn) = -\left(\sup_{\lambda>0} \lambda t - \ln \mathbb{E}[e^{\lambda X_1}]\right)$.

Proof: Upper bound

Let $Y_n = \frac{\sum_{i=1}^{n} X_i}{n}$, $M(\lambda) = \mathbb{E}[e^{\lambda X_1}]$, and $\psi(\lambda) = \ln M(\lambda)$.

$\Pr(Y_n \geq t) \leq e^{-\lambda n t}(M(\lambda))^n$ for any $\lambda \geq 0$.

$\frac{1}{n} \ln \Pr(Y_n \geq t) \leq -\lambda t + \psi(\lambda)$.

$\frac{1}{n} \ln \Pr(Y_n \geq t) \leq -\sup_{\lambda \geq 0}(\lambda t - \psi(\lambda))$.

Large Deviation Principle: Proof

Lower bound

The maximizer λ_0 of $\lambda t - \psi(\lambda)$ satisfies $t = \int \frac{xe^{\lambda_0 x}}{M(\lambda_0)} d\mu(x)$.

Let $d\mu_0(x) = \frac{e^{\lambda_0 x}}{M(\lambda_0)} d\mu(x)$. Its expectation $\int x d\mu_0(x) = t$.

Let $A = \{Y_n \geq t\} \subseteq \mathbb{R}^n, A_\delta = \{Y_n \in [t, t+\delta]\} \subseteq \mathbb{R}^n$.

$$\begin{aligned}
\Pr_\mu(A) \geq \Pr_\mu(A_\delta) &= \int_{A_\delta} \Pi_{i=1}^n d\mu(x_i) \\
&= \int_{A_\delta} (M(\lambda_0))^n e^{-\lambda_0 \sum_{i=1}^n x_i} \Pi_{i=1}^n d\mu_0(x_i) \\
&\geq \left(M(\lambda_0) e^{-\lambda_0 (t+\delta)} \right)^n \Pr_{\mu_0}(A_\delta).
\end{aligned}$$

Applying CLT to μ_0, we have $\lim_{n\to\infty} \Pr_{\mu_0}(A_\delta) = \frac{1}{2}$.

$\lim_{n\to\infty} \frac{1}{n} \ln \Pr(Y_n \geq t) \geq \psi(\lambda_0) - (t+\delta)\lambda_0$, and let $\delta \to 0$.

Remarks

Large deviation theory vs CLT

Seemingly easy to get exponential decay in many cases, but hard to calculate.

Chernoff bounds concern large deviation

- Con: Generally weaker
- Pro: Always holds, not just asymptotically

Key assumption

Independence!

One-order generalization of Chernoff Bound

General $f(X_1, \ldots, X_n)$	**McDiarmid 1989**	**Zhang, Liu et al. 2019**	**Kontorovich et al. 2008**
Linear $X_1 + \cdots + X_n$	**Chernoff 1948**	**Janson 2004**	**Bosq 2012**
	Independent	**Dependent (Qualitative)**	**Dependent (Quantitative)**

References

1. http://nowak.ece.wisc.edu/SLT07/lecture7.pdf
2. https://www.math.illinois.edu/~psdey/414CourseNotes.pdf
3. When Do the Moments Uniquely Identify a Distribution
4. http://willperkins.org/6221/slides/largedeviations.pdf