Probabilistic Method and Random Graphs

Lecture 4. 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 of Lecture 3

Chernoff bounds for 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 $1 > \delta > 0$, $\Pr(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} \leq e^{-\frac{\delta^2}{2}\mu}$.

Exponentially decreasing upper bound!
μ can be replaced by its upper/lower bound.

Trick in the proof: introduce λ and $e^{(\cdot)}$

$$\Pr(X \geq (1+\delta)\mu) = \Pr\left(e^{\lambda X} \geq e^{\lambda(1+\delta)\mu}\right) \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda(1+\delta)\mu}}$$

A brief review of Lecture 3

Specialized for i.i.d. case

$\Pr(|X - \mu| > t) \leq e^{-\frac{2t^2}{n}}$ for any $t > 0$.

Generalization

Other domains $[0, b_i]$, or non-binary over $[0, 1]$.

Hoeffding's Ineq. for $[a_i, b_i]$: $\Pr(|X - \mathbb{E}[X]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_i (b_i - a_i)^2}}$.
Bernstein's and McDiarmid's Ineq.: higher order and beyond sum.

McDiarmid's Ineq.

	Independent	Dependent (Qualitative)	Dependent (Quantitative)
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**

A brief review of Lecture 3

Paradigm: Union bound + Chernoff bounds.

Application

X: number of Heads in n tosses of a fair coin.

- Markov's inequality: $\Pr(X - \frac{n}{2} > \sqrt{n \ln n}) < 1$
- Chebyshev's inequality: $\Pr(X - \frac{n}{2} > \sqrt{n \ln n}) < \frac{1}{\ln n}$
- Chernoff bounds: $\Pr(X - \frac{n}{2} > \sqrt{n \ln n}) < \frac{1}{n^2}$

Reflections

Why are Chernoff bound so good?
Can it be improved by non-exponential functions?
Is there anything to do with moments?
How much information do moments capture?

The story begins with generating functions$\cdots$

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 distributions.

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 $M_X(t) = M_Y(t)$ and both converge around 0, then X and Y are identically distributed.

But

Moment generating functions 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 distributions:

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.
Due to invertibility of the Fourier transform.

Ready to get insights

Moments

How much information do moments capture?

Conditionally, moments=distribution.

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):
$F(x) = \sum_{n=0}^{\infty} F_n x^n = x + \sum_{n=2}^{\infty} (F_{n-1} + F_{n-2}) x^n =$
$x + xF(x) + x^2 F(x)$
$\Rightarrow F(x) = \frac{x}{1-x-x^2} = \frac{1}{\sqrt{5}} \left(\frac{\psi}{x+\psi} - \frac{\phi}{x+\phi} \right) = \sum_{n=0}^{\infty} \frac{1}{\sqrt{5}} \left(\phi^n - \psi^n \right) x^n$
$\Rightarrow F_n = \frac{1}{\sqrt{5}} \left(\phi^n - \psi^n \right)$, where $\phi = \frac{1+\sqrt{5}}{2}, \psi = \frac{1-\sqrt{5}}{2}$.

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.

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

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

$\Pr(X_n \geq \frac{n}{2} + \frac{n}{3}) \approx e^{-0.2426n+o(n)} \ll \text{Chernoff bound.}$
See *Large Deviations-Willperkins.pdf*

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 of $\omega(\sqrt{n})$).

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

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 fit for large deviation

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

Key assumption

Independence!

References

1. http://nowak.ece.wisc.edu/SLT07/lecture7.pdf
2. When Do the Moments Uniquely Identify a Distribution
3. http://willperkins.org/6221/slides/largedeviations.pdf