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

Applications to Coupon Collector's Problem

$\mathbb{E}[X] = nH(n) \approx n \ln n \approx 30 \ll 200$ when $n = 12$.
Markov's inequality: $\Pr(X > 200) < 1/6$.
Chebychev's inequality: $\Pr(X > 200) < 0.01$.
Union bound: $\Pr(X > 200) < 0.00001$.
The more information you have, the better bounds you get.

Chernoff bounds: inequalities of independent sum

Motiving Example

Flip a fair coin for n trials. Let X be the number of Heads, which is around the expectation $\frac{n}{2}$. How about its concentration?

- Markov's inequality: $\Pr(X - \frac{n}{2} > \sqrt{n \ln n}) < \frac{n}{n+2\sqrt{n \ln n}} \rightsquigarrow 1$
- Chebyshev's inequality: $\Pr(X - \frac{n}{2} > \sqrt{n \ln n}) < \frac{1}{\ln n}$
- Can we do better?
  - Due to independent sum: $X = \sum_{i=1}^{n} X_i$
  - YES!

Chernoff bounds: basic form

Chernoff bounds

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

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

Remarks

Note that $0 < \frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} < 1$ when $\delta > 0$.
The bound in 1 exponentially deceases w.r.t. μ!
And so is the bound in 2.

Proof of the upper tail bound

For any $\lambda > 0$,

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

Proof of the upper tail bound

For any $\lambda > 0$,
$$\Pr(X \geq (1+\delta)\mu) = \Pr\left(e^{\lambda X} \geq e^{\lambda(1+\delta)\mu}\right) \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda(1+\delta)\mu}}.$$

$$\mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n} X_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda X_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right].$$

Proof of the upper tail bound

For any $\lambda > 0$,
$\Pr(X \geq (1+\delta)\mu) = \Pr\left(e^{\lambda X} \geq e^{\lambda(1+\delta)\mu}\right) \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda(1+\delta)\mu}}$.

$\mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n} X_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda X_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right]$.

Let $p_i = \Pr(X_i = 1)$ for each i. Then,
$\mathbb{E}\left[e^{\lambda X_i}\right] = p_i e^{\lambda \cdot 1} + (1-p_i)e^{\lambda \cdot 0} = 1 + p_i(e^{\lambda} - 1) \leq e^{p_i(e^{\lambda}-1)}$.

Proof of the upper tail bound

For any $\lambda > 0$,
$\Pr(X \geq (1+\delta)\mu) = \Pr\left(e^{\lambda X} \geq e^{\lambda(1+\delta)\mu}\right) \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda(1+\delta)\mu}}$.

$$\mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n} X_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda X_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right].$$

Let $p_i = \Pr(X_i = 1)$ for each i. Then,
$\mathbb{E}\left[e^{\lambda X_i}\right] = p_i e^{\lambda \cdot 1} + (1-p_i)e^{\lambda \cdot 0} = 1 + p_i(e^{\lambda} - 1) \leq e^{p_i(e^{\lambda}-1)}$.

So, $\mathbb{E}\left[e^{\lambda X}\right] \leq \prod_{i=1}^{n} e^{p_i(e^{\lambda}-1)} = e^{\sum_{i=1}^{n} p_i(e^{\lambda}-1)} = e^{(e^{\lambda}-1)\mu}$.

Proof of the upper tail bound

For any $\lambda > 0$,
$\Pr(X \geq (1+\delta)\mu) = \Pr\left(e^{\lambda X} \geq e^{\lambda(1+\delta)\mu}\right) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda(1+\delta)\mu}}$.

$$\mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n} X_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda X_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right].$$

Let $p_i = \Pr(X_i = 1)$ for each i. Then,
$\mathbb{E}\left[e^{\lambda X_i}\right] = p_i e^{\lambda \cdot 1} + (1-p_i)e^{\lambda \cdot 0} = 1 + p_i(e^{\lambda} - 1) \leq e^{p_i(e^{\lambda}-1)}$.

So, $\mathbb{E}\left[e^{\lambda X}\right] \leq \prod_{i=1}^{n} e^{p_i(e^{\lambda}-1)} = e^{\sum_{i=1}^{n} p_i(e^{\lambda}-1)} = e^{(e^{\lambda}-1)\mu}$.

Thus, $\Pr(X \geq (1+\delta)\mu) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda(1+\delta)\mu}} \leq \frac{e^{(e^{\lambda}-1)\mu}}{e^{\lambda(1+\delta)\mu}} = \left(\frac{e^{(e^{\lambda}-1)}}{e^{\lambda(1+\delta)}}\right)^{\mu}$.
Let $\lambda = \ln(1+\delta) > 0$, and the proof ends.

Lower tail bound

Can be proved likewise.

A tentative application

Recall the coin flipping example. By the Chernoff bound,

$$\Pr(X - \frac{n}{2} > \sqrt{n \ln n}) < \frac{e^{\sqrt{n \ln n}}}{\left(1 + 2\sqrt{\frac{\ln n}{n}}\right)^{\left(\frac{n}{2} + \sqrt{n \ln n}\right)}}$$

Even hard to figure out the order.

Is there a more *friendly* form?

Chernoff bounds: a simplified form

Simplified Chernoff bounds for i.i.d. case

Let $X = \sum_{i=1}^{n} X_i$, where $X_i's$ are i.i.d. Bernoulli r.v. Let $\mu = \mathbb{E}[X]$. Then $\Pr(|X - \mu| > t) \leq e^{-\frac{2t^2}{n}}$ for any $t > 0$.

Simplified Chernoff bounds

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

1. $\Pr(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2}{2+\delta}\mu}$ for any $\delta > 0$;
2. $\Pr(X \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2}{2}\mu}$ for any $1 > \delta > 0$.

Application to coin flipping

$\Pr(X - \frac{n}{2} > \sqrt{n \ln n}) \leq n^{-\frac{2}{3}}$. This is exponentially tighter than Chebychev's inequality $\frac{1}{\ln n}$.

Proof and Remarks

Idea of the proof

1. $\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} \leq e^{-\frac{\delta^2}{2+\delta}} \Leftrightarrow \delta - (1+\delta)\ln(1+\delta) < -\frac{\delta^2}{2+\delta} \Leftarrow$ $\ln(1+\delta) > \frac{2\delta}{2+\delta}$ for $\delta > 0$.
2. Use calculus to show that $\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}} \leq e^{-\frac{\delta^2}{2}}$.

Remark 1

When $1 > \delta > 0$, we have $-\frac{\delta^2}{2+\delta} < -\frac{\delta^2}{3}$, so
$\Pr(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2}{3}\mu}$, and $\Pr(|X-\mu| \geq \delta\mu) \leq 2e^{-\frac{\delta^2}{3}\mu}$.

Remark 2

The bound is simpler but looser. Generally, it is outperformed by the basic Chernoff bound. See example.

Example: random rounding

Minimum-congestion path planning

- $G = (V, E)$ is an undirected graph. $D = \{(s_i, t_i)\}_{i=1}^m \subseteq V^2$.
- Find a path P_i connecting (s_i, t_i) for every i.
- Objective: minimize the congestion $\max_{e \in E} cong(e)$, the number of the paths among $\{P_i\}_{i=1}^m$ that pass e.

This problem is NP-hard. We will give an approximation algorithm based on randomized rounding.

- Model as an integer program
- Relax it into a linear program
- Round the solution
- Analyze the approximation ratio

ILP and its relaxation

Notation

$\mathbb{P}_i$: the set of candidate paths connecting s_i and t_i;
f_P^i: the indicator of whether we pick path $P \in \mathbb{P}_i$ or not;
C: the congestion in the graph.

ILP

$$\begin{aligned} &\text{Min } C \\ &s.t. \sum_{P\in\mathbb{P}_i} f_P^i = 1, \forall i \\ &\quad \sum_i \sum_{e\in P\in\mathbb{P}_i} f_P^i \leq C, \forall e \\ &\quad \underline{f_P^i \in \{0,1\}}, \forall i, P \end{aligned}$$

$\Rightarrow$

LP

$$\begin{aligned} &\text{Min } C \\ &s.t. \sum_{P\in\mathbb{P}_i} f_P^i = 1, \forall i \\ &\quad \sum_i \sum_{e\in P\in\mathbb{P}_i} f_P^i \leq C, \forall e \\ &\quad \underline{f_P^i \in [0,1]}, \forall i, P \end{aligned}$$

Round a solution to the LP

For each i, $f_\cdot^i$ forms a probability function on $\mathbb{P}_i$, which enables to randomly choose **one** path (say P_i) from $\mathbb{P}_i$.
Use $P_1, \cdots, P_m$ as an approximate solution to the ILP.

Approximation ratio

Notation

C: optimum congestion of the ILP.
C^*: optimum congestion of the LP. $C^* \leq C$.
X_i^e: indicator of whether $e \in P_i$.
$X^e \triangleq \sum_i X_i^e$: congestion of the edge e.
$X \triangleq \max_e X^e$: the network congestion.

Objective: $\Pr(X > (1+\delta)C)$ is small for a small δ

Note that $X > (1+\delta)C$ equals $\bigcup_{e \in E}(X^e > (1+\delta)C)$.
By union bound, we just show $\Pr(X^e > (1+\delta)C) < \frac{1}{n^3}$ for any e.

Apply Chernoff bound to $X^e = \sum_i X_i^e$

Prove $\Pr(X^e > (1+\delta)C) < \frac{1}{n^3}$

Easy facts

$\mathbb{E}[X_i^e] = \sum_{e \in P \in \mathbb{P}_i} f_P^i$.
$\mu = \mathbb{E}[X^e] = \sum_i \mathbb{E}[X_i^e] = \sum_i \sum_{e \in P \in \mathbb{P}_i} f_P^i \leq C^* \leq C$.

If $C = \omega(\ln n)$, δ can be arbitrarily small

Proof: For any $0 < \delta < 1$, $\Pr(X^e > (1+\delta)C) \leq e^{-\frac{\delta^2 C}{2+\delta}} \leq \frac{1}{n^3}$.

If $C = O(\ln n)$, $\delta = \Theta(\ln n)$

Proof: $\Pr(X^e > (1+\delta)C) \leq e^{-\frac{\delta^2 C}{2+\delta}} \leq e^{-\frac{\delta}{2}}$ for $\delta \geq 2$.
So, $\Pr(X^e > (1+\delta)C) \leq \frac{1}{n^3}$ when $\delta = 6 \ln n$.

Prove $\Pr(X^e > (1+\delta)C) < \frac{1}{n^3}$

If $C = O(\ln n)$, δ can be improved to be $\delta = \Theta\left(\frac{\ln n}{\ln \ln n}\right)$

Proof: By the basic Chernoff bounds,

$$\Pr(X^e > (1+\delta)C) \le \left[\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right]^C \le \frac{e^\delta}{(1+\delta)^{(1+\delta)}}.$$

When $\delta = \Theta\left(\frac{\ln n}{\ln \ln n}\right)$, $(1+\delta)\ln(1+\delta) = \Theta(\ln n)$ and
$\delta - (1+\delta)\ln(1+\delta) = \Theta(\ln n)$.

Remarks of the application

Remark 1

It illustrates the practical difference of various Chernoff bounds.

Remark 2

Is it a mistake to use the inaccurate expectation?
No! It's a powerful trick.
If $\mu_L \leq \mu \leq \mu_H$, the following bounds hold:

- Upper tail: $\Pr(X \geq (1+\delta)\mu_H) \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu_H}$.
- Lower tail: $\Pr(X \leq (1-\delta)\mu_L) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu_L}$.

Chernoff bounds + Union bound: a paradigm

A high-level picture: Want to upper-bound $\Pr(\text{something bad})$.
1. By Union bound, $\Pr(\text{something bad}) \leq \sum_{i=1}^{\text{Large}} \Pr(\text{bad}_i)$;
2. By Chernoff bounds, $\Pr(\text{bad}_i) \leq \text{minuscule}$ for each i;
3. $\Pr(\text{something bad}) \leq \text{Large} \times \text{minuscule} = \text{small}$.

Why the Chernoff bound is better? Note that it's rooted at Markov's Inequality.

Can it be improved by using functions other than exponential?

General bounds for independent sums

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

Basic Chernoff bounds remain valid.

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.

$$\Pr(X - \mathbb{E}[X] \geq t) \leq \frac{\prod_i \mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}]}{e^{\lambda t}}$$
$$\leq e^{\lambda^2 \sum_i \frac{(b_i - a_i)^2}{8} - \lambda t}$$

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

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

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

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

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 $Z_i = b_1 + \ldots + b_k$.

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

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

Concentration Inequalities: higher order and beyond sum

Bernstein Inequality (Bernstein, 1911)

Let $X = \sum_{i=1}^{n} X_i$, where X_i's are independent r.v satisfying $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] \le b_i$, $\mathbb{E}[|X_i|^k] \le \frac{b_i}{2} H^{k-2} k!$ for $k > 2$. Then

$\Pr(|X| \ge t) \le 2e^{-\frac{t^2}{2\left(Ht + \sum_{i=1}^{n} b_i\right)}}$ for any $t > 0$

- Extended to dependent or multi-dimensional variables

McDiarmid Inequality (McDiarmid, 1989)

Let $X_1, \cdots, X_n$ be independent r.v. and $f(\cdot)$ is an n-ary function with bounded differences $c_1, \cdots, c_n$. Let $Y = f(X_1, \cdots, X_n)$.

Then $\Pr(|Y - \mathbb{E}[Y]| \ge t) \le 2e^{-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}}$ for any $t > 0$

- Extended to (quantitative and qualitative) dependence

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?