Sample Complexity and Uniform Convergence for Learning and Data Analysis

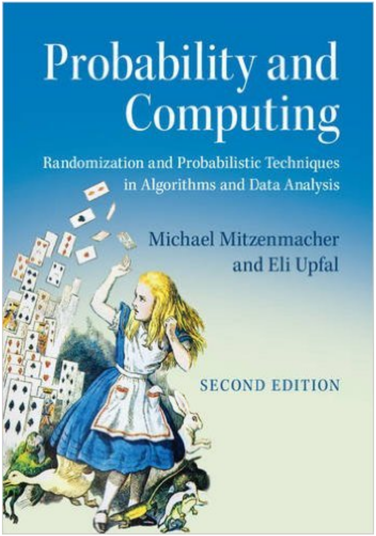
Eli Upfal
Brown University

BROWN

Outline - What I'll try to cover...

- Large Deviation
  - The basic scheme: How to create your own bound
  - iid bounds: Chernoff bound, Hoeffding bound
  - Martingale bounds: Azuma-Hoeffding bound, McDiarmid bound
- Uniform convergence
  - Sample complexity and machine learning
  - VC-dimension bounds
  - Rademacher complexity bounds
  - Applications beyond machine learning

It's (almost) all in the book:



Fine Sample Techniques

A typical probability theory statement:

Theorem (The Central Limit Theorem)

Let $X_1, \ldots, X_n$ be independent identically distributed random variables with common mean μ and variance σ^2. Then

$$\lim_{n\to\infty} \Pr(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt.$$

A typical CS probabilistic tool:

Theorem (Chernoff Bound)

Let $X_1, \ldots, X_n$ be independent Bernoulli random variables such that $Pr(X_i = 1) = p$, then

$$Pr(\sum_{i=1}^{n} X_i \geq (1+\delta)np) \leq e^{-np\delta^2/3}.$$

The Basic Idea of Large Deviation Bounds:

For any random variable X, by Markov inequality we have:
For any $t > 0$,

$$Pr(X \geq a) = Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

Similarly, for any $t < 0$

$$Pr(X \leq a) = Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

We use:

Theorem (Markov Inequality)

If a random variable X is non-negative ($X \geq 0$) then

$$Prob(X \geq a) \leq \frac{E[X]}{a}.$$

The General Scheme:

We obtain specific bounds for particular conditions/distributions by

1. Compute $E[e^{tX}]$
2. Optimize w.r.t t,

$$Pr(X \geq a) \leq \min_{t>0} \frac{\mathbf{E}[e^{tX}]}{e^{ta}}$$

$$Pr(X \leq a) \leq \min_{t<0} \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

3. Simplify

Moment Generating Function

Definition

The moment generating function of a random variable X is

$$M_X(t) = \mathbf{E}[e^{tX}].$$

Theorem

If $M_X(t)$ exists in some neighborhood of 0, then for all $n \geq 1$,

$$\mathbf{E}[X^n] = M_X^{(n)}(0) = \left. \frac{d^n M_X(t)}{dt} \right|_{t=0}.$$

Theorem

For independent random variables X and Y,

$$M_{X+Y}(t) = M_X(t) M_Y(t).$$

Chernoff Bound for Sum of Bernoulli Trials

Let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli trials with $Pr(X_i = 1) = p_i$. Let $X = \sum_{i=1}^{n} X_i$, and let

$$\mu = \mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{E}[X_i] = \sum_{i=1}^{n} p_i.$$

For each X_i:

$$\begin{aligned} M_{X_i}(t) &= \mathbf{E}[e^{tX_i}] \\ &= p_i e^t + (1 - p_i) \\ &= 1 + p_i(e^t - 1) \\ &\leq e^{p_i(e^t - 1)}. \end{aligned}$$

$$M_{X_i}(t) = \mathbf{E}[e^{tX_i}] \leq e^{p_i(e^t-1)}.$$

Taking the product of the n generating functions we get for $X = \sum_{i=1}^{n} X_i$

$$\begin{aligned} M_X(t) &= \prod_{i=1}^{n} M_{X_i}(t) \\ &\leq \prod_{i=1}^{n} e^{p_i(e^t-1)} \\ &= e^{\sum_{i=1}^{n} p_i(e^t-1)} \\ &= e^{(e^t-1)\mu} \end{aligned}$$

$$M_X(t) = \mathbf{E}[e^{tX}] = e^{(e^t-1)\mu}$$

Applying Markov's inequality we have for any $t > 0$

$$\begin{aligned} Pr(X \geq (1+\delta)\mu) &= Pr(e^{tX} \geq e^{t(1+\delta)\mu}) \\ &\leq \frac{\mathbf{E}[e^{tX}]}{e^{t(1+\delta)\mu}} \\ &\leq \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}} \end{aligned}$$

For any $\delta > 0$, we can set $t = \ln(1+\delta) > 0$ to get:

$$Pr(X \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}.$$

Theorem

Let $X_1, \ldots, X_n$ *be independent Bernoulli random variables such that* $Pr(X_i = 1) = p_i$. *Let* $\mu = E[X] = \sum_{i=1}^{n} p_i$, *then*

- *For any* $\delta > 0$,

$$Pr(X \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \right)^{\mu}. \tag{1}$$

- *For* $0 < \delta \leq 1$,

$$Pr(X \geq (1+\delta)\mu) \leq e^{-\mu\delta^2/3}. \tag{2}$$

- *For* $R \geq 6\mu$,

$$Pr(X \geq R) \leq 2^{-R}. \tag{3}$$

Theorem

Let $X_1, \ldots, X_n$ *be independent Bernoulli random variables such that* $Pr(X_i = 1) = p_i$. *Let* $X = \sum_{i=1}^{n} X_i$ *and* $\mu = \mathbf{E}[X]$.
For $0 < \delta < 1$:

- $$Pr(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}. \tag{4}$$

- $$Pr(X \leq (1-\delta)\mu) \leq e^{-\mu\delta^2/2}. \tag{5}$$

Using Markov's inequality, for any $t < 0$,

$$\begin{aligned} Pr(X \leq (1-\delta)\mu) &= Pr(e^{tX} \geq e^{(1-\delta)t\mu}) \\ &\leq \frac{\mathbf{E}[e^{tX}]}{e^{t(1-\delta)\mu}} \\ &\leq \frac{e^{(e^t-1)\mu}}{e^{t(1-\delta)\mu}} \end{aligned}$$

For $0 < \delta < 1$, we set $t = \ln(1-\delta) < 0$ to get:

$$Pr(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}} \right)^{\mu}$$

This proves (4).
We need to show:

$$f(\delta) = -\delta - (1-\delta)\ln(1-\delta) + \frac{1}{2}\delta^2 \leq 0.$$

We need to show:

$$f(\delta) = -\delta - (1-\delta)\ln(1-\delta) + \frac{1}{2}\delta^2 \le 0.$$

Differentiating $f(\delta)$ we get

$$\begin{aligned} f'(\delta) &= \ln(1-\delta) + \delta, \\ f''(\delta) &= -\frac{1}{1-\delta} + 1. \end{aligned}$$

Since $f''(\delta) < 0$ for $\delta \in (0,1)$, $f'(\delta)$ decreasing in that interval. Since $f'(0) = 0$, $f'(\delta) \le 0$ for $\delta \in (0,1)$. Therefore $f(\delta)$ is non increasing in that interval.
$f(0) = 0$. Since $f(\delta)$ is non increasing for $\delta \in [0,1)$, $f(\delta) \le 0$ in that interval, and (5) follows.

Example: Coin flips

Let X be the number of heads in a sequence of n independent fair coin flips.

$$Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{1}{2}\sqrt{4n\ln n}\right)$$

$$= Pr\left(X \geq \frac{n}{2}\left(1 + \sqrt{\frac{4\ln n}{n}}\right)\right) + Pr\left(X \leq \frac{n}{2}\left(1 - \sqrt{\frac{4\ln n}{n}}\right)\right)$$

$$\leq e^{-\frac{1}{3}\frac{n}{2}\frac{4\ln n}{n}} + e^{-\frac{1}{2}\frac{n}{2}\frac{4\ln n}{n}} \leq \frac{2}{n}.$$

Note that the standard deviation is $\sqrt{n}/2$

The probability of $\geq 3n/4$ heads

Markov Inequality gives

$$Pr\left(X \geq \frac{3n}{4}\right) \leq \frac{n/2}{3n/4} \leq \frac{2}{3}.$$

Using the Chebyshev's bound we have:

$$Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq \frac{n/4}{(n/4)^2} = \frac{4}{n}.$$

Using the Chernoff bound in this case, we obtain

$$\begin{aligned} Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) &= Pr\left(X \geq \frac{n}{2}\left(1 + \frac{1}{2}\right)\right) \\ &+ Pr\left(X \leq \frac{n}{2}\left(1 - \frac{1}{2}\right)\right) \\ &\leq e^{-\frac{1}{3}\frac{n}{2}\frac{1}{4}} + e^{-\frac{1}{2}\frac{n}{2}\frac{1}{4}} \\ &\leq 2e^{-\frac{n}{24}}. \end{aligned}$$

Chernoff's vs. Chebyshev's Inequality

Assume for all i we have $p_i = p; 1 - p_i = q$.

$$\mu = \mathbf{E}[X] = np$$

$$Var[X] = npq$$

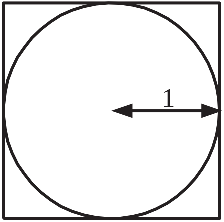
If we use Chebyshev's Inequality we get

$$Pr(|X - \mu| > \delta\mu) \leq \frac{npq}{\delta^2\mu^2} = \frac{npq}{\delta^2 n^2 p^2} = \frac{q}{\delta^2\mu}$$

Chernoff bound gives

$$Pr(|X - \mu| > \delta\mu) \leq 2e^{-\mu\delta^2/3}.$$

Example: Estimate the value of π



- Choose X and Y independently and uniformly at random in $[0, 1]$.
- Let

$$Z = \begin{cases} 1 & \text{if } \sqrt{X^2 + Y^2} \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

- $\Pr(Z = 1) = \frac{\pi}{4}$.
- $4\mathbf{E}[Z] = \pi$.

- Let $Z_1, \ldots, Z_m$ be the values of m independent experiments. $W = \sum_{i=1}^{m} Z_i$.
- $$\mathbf{E}[W] = \mathbf{E}\left[\sum_{i=1}^{m} Z_i\right] = \sum_{i=1}^{m} \mathbf{E}[Z_i] = \frac{m\pi}{4},$$
- $W' = \frac{4}{m}W$ is an unbiased estimate for π.
- $$\begin{aligned} \Pr(|W' - \pi| \geq \epsilon\pi) &= \Pr\left(|W - \frac{m\pi}{4}| \geq \frac{\epsilon m\pi}{4}\right) \\ &= \Pr\left(|W - \mathbf{E}[W]| \geq \epsilon\mathbf{E}[W]\right) \\ &\leq 2\mathrm{e}^{-\frac{1}{12}m\pi\epsilon^2} = \delta. \end{aligned}$$

For fixed ϵ and δ we need $m \geq O(\frac{\ln\frac{2}{\delta}}{\pi\epsilon^2})$ samples.

Set Balancing

Given an $n \times n$ matrix $\mathcal{A}$ with entries in $\{0, 1\}$, let

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ \dots \\ b_n \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \dots \\ \dots \\ c_n \end{pmatrix}.$$

Find a vector $\bar{b}$ with entries in $\{-1, 1\}$ that minimizes

$$||\mathcal{A}\bar{b}||_\infty = \max_{i=1,\dots,n} |c_i|.$$

Theorem

For a random vector $\bar{b}$, with entries chosen independently and with equal probability from the set $\{-1, 1\}$,

$$Pr(||\mathcal{A}\bar{b}||_\infty \geq \sqrt{4n \ln n}) \leq \frac{2}{n}.$$

The $\sum_{i=1}^{n} a_{j,i} b_i$ (excluding the zero terms) is a sum of independent $-1, 1$ random variable. We need a bound on such sum.

Chernoff Bound for Sum of $\{-1, +1\}$ Random Variables

Theorem

Let $X_1, ..., X_n$ be independent random variables with

$$Pr(X_i = 1) = Pr(X_i = -1) = \frac{1}{2}.$$

Let $X = \sum_1^n X_i$. For any $a > 0$,

$$Pr(X \geq a) \leq e^{-\frac{a^2}{2n}}.$$

de Moivre – Laplace approximation: For any k, such that $|k - np| \leq a$

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{a^2}{2np(1-p)}}$$

For any $t > 0$,

$$\mathbf{E}[e^{tX_i}] = \frac{1}{2}e^t + \frac{1}{2}e^{-t}.$$

$$e^t = 1 + t + \frac{t^2}{2!} + \cdots + \frac{t^i}{i!} + \ldots$$

and

$$e^{-t} = 1 - t + \frac{t^2}{2!} + \cdots + (-1)^i\frac{t^i}{i!} + \ldots$$

Thus,

$$\begin{aligned}\mathbf{E}[e^{tX_i}] &= \frac{1}{2}e^t + \frac{1}{2}e^{-t} = \sum_{i\geq 0} \frac{t^{2i}}{(2i)!} \\ &\leq \sum_{i\geq 0} \frac{(\frac{t^2}{2})^i}{i!} = e^{t^2/2}\end{aligned}$$

$$\mathbf{E}[e^{tX}] = \prod_{i=1}^{n} \mathbf{E}[e^{tX_i}] \leq e^{nt^2/2},$$

$$Pr(X \geq a) = Pr(e^{tX} > e^{ta}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}} \leq e^{t^2 n/2 - ta}.$$

Setting $t = a/n$ yields

$$Pr(X \geq a) \leq e^{-\frac{a^2}{2n}}.$$

By symmetry we also have

Corollary

Let $X_1, \ldots, X_n$ be independent random variables with

$$Pr(X_i = 1) = Pr(X_i = -1) = \frac{1}{2}.$$

Let $X = \sum_{i=1}^{n} X_i$. Then for any $a > 0$,

$$Pr(|X| > a) \leq 2e^{-\frac{a^2}{2n}}.$$

Application: Set Balancing

Theorem

For a random vector $\bar{b}$, with entries chosen independently and with equal probability from the set $\{-1, 1\}$,

$$Pr(||\mathcal{A}\bar{b}||_\infty \geq \sqrt{4n \ln n}) \leq \frac{2}{n} \tag{6}$$

- Consider the i-th row $\bar{a}_i = a_{i,1}, ..., a_{i,n}$.
- Let k be the number of 1's in that row.
- $Z_i = \sum_{j=1}^{k} a_{i,i_j} b_{i_j}$.
- If $k \leq \sqrt{4n \ln n}$ then clearly $Z_i \leq \sqrt{4n \ln n}$.

If $k > \sqrt{4n \log n}$, the k non-zero terms in the sum Z_i are independent random variables, each with probability $1/2$ of being either $+1$ or -1.
Using the Chernoff bound:

$$Pr\left\{|Z_i| > \sqrt{4n \log n}\right\} \leq 2e^{-4n \log n/(2k)} \leq \frac{2}{n^2},$$

where we use the fact that $n \geq k$.
The result follows by union bound (n rows).

Hoeffding's Inequality

Large deviation bound for more general random variables:

Theorem (Hoeffding's Inequality)

Let $X_1, \ldots, X_n$ be independent random variables such that for all $1 \leq i \leq n$, $E[X_i] = \mu$ and $Pr(a \leq X_i \leq b) = 1$. Then

$$Pr(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

Lemma

(Hoeffding's Lemma) Let X be a random variable such that $Pr(X \in [a, b]) = 1$ and $E[X] = 0$. Then for every $\lambda > 0$,

$$\mathbf{E}[e^{\lambda X}] \leq e^{\lambda^2 (a-b)^2}/8.$$

Proof of the Lemma

Since $f(x) = e^{\lambda x}$ is a convex function, for any $\alpha \in (0,1)$ and $x \in [a,b]$,

$$f(X) \leq \alpha f(a) + (1-\alpha) f(b).$$

Thus, for $\alpha = \frac{b-x}{b-a} \in (0,1)$,

$$e^{\lambda x} \leq \frac{b-x}{b-a} e^{\lambda a} + \frac{x-a}{b-a} e^{\lambda b}.$$

Taking expectation, and using $\mathbf{E}[X] = 0$, we have

$$E[e^{\lambda X}] \leq \frac{b}{b-a} e^{\lambda a} + \frac{a}{b-a} e^{\lambda b} \leq e^{\lambda^2 (b-a)^2/8}.$$

Proof of the Bound

Let $Z_i = X_i - \mathbf{E}[X_i]$ and $Z = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$Pr(Z \geq \epsilon) \leq e^{-\lambda\epsilon}\mathbf{E}[e^{\lambda Z}] \leq e^{-\lambda\epsilon}\prod_{i=1}^{n}\mathbf{E}[e^{\lambda X_i/n}] \leq e^{-\lambda\epsilon+\frac{\lambda^2(b-a)^2}{8n}}$$

Set $\lambda = \frac{4n\epsilon}{(b-a)^2}$ gives

$$Pr(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \epsilon) = Pr(Z \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

A More General Version

Theorem

Let $X_1, \ldots, X_n$ *be independent random variables with* $\mathbf{E}[X_i] = \mu_i$ *and* $Pr(B_i \leq X_i \leq B_i + c_i) = 1$, *then*

$$Pr(|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i| \geq \epsilon) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n} c_i^2}}$$